IN SILICO METHODS FOR DRUG DESIGN AND DISCOVERY

EDITED BY: Simone Brogi, Teodorico Castro Ramalho, José L. Medina-Franco, Kamil Kuca and Marian Valko

PUBLISHED IN: Frontiers in Chemistry

frontiers Research Topics

frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-88966-057-5
DOI 10.3389/978-2-88966-057-5

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

IN SILICO METHODS FOR DRUG DESIGN AND DISCOVERY

Topic Editors:
Simone Brogi, University of Pisa, Italy
Teodorico Castro Ramalho, Universidade Federal de Lavras, Brazil
José L. Medina-Franco, National Autonomous University of Mexico, Mexico
Kamil Kuca, University of Hradec Králové, Czechia
Marian Valko, Slovak University of Technology in Bratislava, Slovakia

Citation: Brogi, S., Ramalho, T. C., Medina-Franco, J. L., Kuca, K., Valko, M., eds. (2020). *In Silico* Methods for Drug Design and Discovery. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-057-5

Table of Contents

06 ***Editorial:* In silico *Methods for Drug Design and Discovery***
Simone Brogi, Teodorico Castro Ramalho, Kamil Kuca, José L. Medina-Franco and Marian Valko

11 ***GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism***
Christina de Bruyn Kops, Conrad Stork, Martin Šícho, Nikolay Kochev, Daniel Svozil, Nina Jeliazkova and Johannes Kirchmair

26 ***Docking, Interaction Fingerprint, and Three-Dimensional Quantitative Structure–Activity Relationship (3D-QSAR) of Sigma1 Receptor Ligands, Analogs of the Neuroprotective Agent RC-33***
José Luis Velázquez-Libera, Giacomo Rossino, Carlos Navarro-Retamal, Simona Collina and Julio Caballero

46 ***Refinement and Rescoring of Virtual Screening Results***
Giulio Rastelli and Luca Pinzi

52 ***Predicting Synergism of Cancer Drug Combinations Using NCI-ALMANAC Data***
Pavel Sidorov, Stefan Naulaerts, Jérémy Ariey-Bonnet, Eddy Pasquier and Pedro J. Ballester

65 ***Finding Constellations in Chemical Space Through Core Analysis***
J. Jesús Naveja and José L. Medina-Franco

75 **In silico *Guided Drug Repurposing: Discovery of New Competitive and Non-competitive Inhibitors of Falcipain-2***
Lucas N. Alberca, Sara R. Chuguransky, Cora L. Álvarez, Alan Talevi and Emir Salas-Sarduy

95 ***Identification of Novel 3-Hydroxy-pyran-4-One Derivatives as Potent HIV-1 Integrase Inhibitors Using* in silico *Structure-Based Combinatorial Library Design Approach***
Hajar Sirous, Giulia Chemi, Sandra Gemma, Stefania Butini, Zeger Debyser, Frauke Christ, Lotfollah Saghaie, Simone Brogi, Afshin Fassihi, Giuseppe Campiani and Margherita Brindisi

115 ***Pushing the Ligand Efficiency Metrics: Relative Group Contribution (RGC) Model as a Helpful Strategy to Promote a Fragment "Rescue" Effect***
Andrés Felipe Vásquez and Andrés González Barrios

122 ***Structural, Thermodynamic, and Kinetic Traits of Antiestrogen-Compounds Selectively Targeting the Y537S Mutant Estrogen Receptor α Transcriptional Activity in Breast Cancer Cell Lines***
Matic Pavlin, Luca Gelsomino, Ines Barone, Angelo Spinello, Stefania Catalano, Sebastiano Andò and Alessandra Magistrato

138 ***Teaching an Old Dog New Tricks: Strategies That Improve Early Recognition in Similarity-Based Virtual Screening***
Ruifeng Liu, Mohamed Diwan M. AbdulHameed and Anders Wallqvist

149 ***Fast Screening of Inhibitor Binding/Unbinding Using Novel Software Tool CaverDock***
Gaspar P. Pinto, Ondrej Vavra, Jiri Filipovic, Jan Stourac, David Bednar and Jiri Damborsky

163 ***Computational Insight Into the Small Molecule Intervening PD-L1 Dimerization and the Potential Structure-Activity Relationship***
Danfeng Shi, Xiaoli An, Qifeng Bai, Zhitong Bing, Shuangyan Zhou, Huanxiang Liu and Xiaojun Yao

178 ***Comparison Study of Computational Prediction Tools for Drug-Target Binding Affinities***
Maha Thafar, Arwa Bin Raies, Somayah Albaradei, Magbubah Essack and Vladimir B. Bajic

197 ***Integrative Multi-Kinase Approach for the Identification of Potent Antiplasmodial Hits***
Marilia N. N. Lima, Gustavo C. Cassiano, Kaira C. P. Tomaz, Arthur C. Silva, Bruna K. P. Sousa, Leticia T. Ferreira, Tatyana A. Tavella, Juliana Calit, Daniel Y. Bargieri, Bruno J. Neves, Fabio T. M. Costa and Carolina Horta Andrade

211 ***Structure-Based and Molecular Modeling Studies for the Discovery of Cyclic Imides as Reversible Cruzain Inhibitors With Potent Anti-*Trypanosoma cruzi *Activity***
Rafael A. A. Ferreira, Ivani Pauli, Thiago S. Sampaio, Mariana L. de Souza, Leonardo L. G. Ferreira, Luma G. Magalhães, Celso de O. Rezende Jr., Rafaela S. Ferreira, Renata Krogh, Luiz C. Dias and Adriano D. Andricopulo

232 ***Could Quantum Mechanical Properties Be Reflected on Classical Molecular Dynamics? The Case of Halogenated Organic Compounds of Biological Interest***
Lucas de Azevedo Santos, Ingrid G. Prandi and Teodorico C. Ramalho

242 ***Novel Quinoline Compound Derivatives of NSC23925 as Potent Reversal Agents Against P-Glycoprotein-Mediated Multidrug Resistance***
Xingping Quan, Hongzhi Du, Jingjing Xu, Xiaoying Hou, Xiaofeng Gong, Yao Wu, Yuqi Zhou, Jingwei Jiang, Ligong Lu, Shengtao Yuan, Xiangyu Yang, Lei Shi and Li Sun

257 **In silico** ***Methods for Design of Kinase Inhibitors as Anticancer Drugs***
Zarko Gagic, Dusan Ruzic, Nemanja Djokovic, Teodora Djikic and Katarina Nikolic

282 ***Vienna LiverTox Workspace—A Set of Machine Learning Models for Prediction of Interactions Profiles of Small Molecules With Transporters Relevant for Regulatory Agencies***
Floriane Montanari, Bernhard Knasmüller, Stefan Kohlbacher, Christoph Hillisch, Christine Baierová, Melanie Grandits and Gerhard F. Ecker

290 ***ChEMBL-Likeness Score and Database GDBChEMBL***
Sven Bühlmann and Jean-Louis Reymond

297 ***A Pilot Study of All-Computational Drug Design Protocol–From Structure Prediction to Interaction Analysis***
Yifei Wu, Lei Lou and Zhong-Ru Xie

306 **In silico** ***Strategies to Support Fragment-to-Lead Optimization in Drug Discovery***
Lauro Ribeiro de Souza Neto, José Teófilo Moreira-Filho, Bruno Junior Neves, Rocío Lucía Beatriz Riveros Maidana, Ana Carolina Ramos Guimarães, Nicholas Furnham, Carolina Horta Andrade and Floriano Paes Silva Jr.

324 **In silico** ***Investigations of the Mode of Action of Novel Colchicine Derivatives Targeting β-Tubulin Isotypes: A Search for a Selective and Specific β-III Tubulin Ligand***
Lorenzo Pallante, Antonio Rocca, Greta Klejborowska, Adam Huczynski, Gianvito Grasso, Jack A. Tuszynski and Marco A. Deriu

332 ***Exploring the RNA-Recognition Mechanism Using Supervised Molecular Dynamics (SuMD) Simulations: Toward a Rational Design for Ribonucleic-Targeting Molecules?***
Maicol Bissaro, Mattia Sturlese and Stefano Moro

345 ***LigBuilder V3: A Multi-Target*** **de novo** ***Drug Design Approach***
Yaxia Yuan, Jianfeng Pei and Luhua Lai

363 ***A Definition of "Multitargeticity": Identifying Potential Multitarget and Selective Ligands Through a Vector Analysis***
Juan Francisco Sánchez-Tejeda, Juan F. Sánchez-Ruiz, Juan Rodrigo Salazar and Marco A. Loza-Mejía

376 **In silico** ***Design of Novel HIV-1 NNRTIs Based on Combined Modeling Studies of Dihydrofuro[3,4-d]pyrimidines***
Yanming Chen, Yafeng Tian, Ya Gao, Fengshou Wu, Xiaogang Luo, Xiulian Ju and Genyan Liu

393 ***Competition Between Phenothiazines and BH3 Peptide for the Binding Site of the Antiapoptotic BCL-2 Protein***
Aline Lagoeiro do Carmo, Fernanda Bettanin, Michell Oliveira Almeida, Simone Queiroz Pantaleão, Tiago Rodrigues, Paula Homem-de-Mello and Kathia Maria Honorio

406 ***Integrating Ligand and Target-Driven Based Virtual Screening Approaches With*** **in vitro** ***Human Cell Line Models and Time-Resolved Fluorescence Resonance Energy Transfer Assay to Identify Novel Hit Compounds Against BCL-2***
Gurbet Tutumlu, Berna Dogan, Timucin Avsar, Muge Didem Orhan, Seyma Calis and Serdar Durdagi

424 ***High-Throughput Docking Using Quantum Mechanical Scoring***
Claudio N. Cavasotto and M. Gabriela Aucar

434 ***Exploring the Use of Compound-Induced Transcriptomic Data Generated From Cell Lines to Predict Compound Activity Toward Molecular Targets***
Benoît Baillif, Joerg Wichard, Oscar Méndez-Lucio and David Rouquié

453 ***Structure-Based Virtual Screening: From Classical to Artificial Intelligence***
Eduardo Habib Bechelane Maia, Letícia Cristina Assis, Tiago Alves de Oliveira, Alisson Marques da Silva and Alex Gutterres Taranto

471 ***An Innovative Sequence-to-Structure-Based Approach to Drug Resistance Interpretation and Prediction: The Use of Molecular Interaction Fields to Detect HIV-1 Protease Binding-Site Dissimilarities***
Nuno G. Alves, Ana I. Mata, João P. Luís, Rui M. M. Brito and Carlos J. V. Simões

488 **In silico** ***Druggability Assessment of the NUDIX Hydrolase Protein Family as a Workflow for Target Prioritization***
Maurice Michel, Evert J. Homan, Elisée Wiita, Kia Pedersen, Ingrid Almlöf, Anna-Lena Gustavsson, Thomas Lundbäck, Thomas Helleday and Ulrika Warpman Berglund

frontiers in Chemistry

EDITORIAL
published: 07 August 2020
doi: 10.3389/fchem.2020.00612

Editorial: *In silico* Methods for Drug Design and Discovery

Simone Brogi[1]*, *Teodorico Castro Ramalho*[2]*, *Kamil Kuca*[3]*, *José L. Medina-Franco*[4]* and *Marian Valko*[5]*

[1] Department of Pharmacy, University of Pisa, Pisa, Italy, [2] Laboratory of Molecular Modeling, Chemistry Department, Federal University of Lavras, Lavras, Brazil, [3] Department of Chemistry, Faculty of Science, University of Hradec Kralove, Kralove, Czechia, [4] DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, Mexico, [5] Faculty of Chemical and Food Technology, Slovak Technical University, Bratislava, Slovakia

Keywords: chemoinformatics, computational chemistry, computational methods in medicinal chemistry, computer-aided-drug design, drug discovery, molecular modeling, web-servers

Editorial on the Research Topic

In silico Methods for Drug Design and Discovery

Computer-aided drug design (CADD) methodologies are playing an ever-increasing role in drug discovery that are critical in the cost-effective identification of promising drug candidates. These computational methods are relevant in limiting the use of animal models in pharmacological research, for aiding the rational design of novel and safe drug candidates, and for repositioning marketed drugs, supporting medicinal chemists and pharmacologists during the drug discovery trajectory. Within this field of research, we launched a Research Topic in *Frontiers in Chemistry* in March 2019 entitled "*In silico Methods for Drug Design and Discovery*," which involved two sections of the journal: Medicinal and Pharmaceutical Chemistry and Theoretical and Computational Chemistry. For the reasons mentioned, this Research Topic attracted the attention of scientists and received a large number of submitted manuscripts. Among them 27 Original Research articles, five Review articles, and two Perspective articles have been published within the Research Topic. The Original Research articles cover most of the topics in CADD, reporting advanced *in silico* methods in drug discovery, while the Review articles offer a point of view of some computer-driven techniques applied to drug research. Finally, the Perspective articles provide a vision of specific computational approaches with an outlook in the modern era of CADD.

Regarding the Original Research articles, two of them are related to innovative approaches concerning ADMET properties of the molecules. In particular, de Bruyn Kops et al. reported the development and validation of GLORY, an innovative tool for predicting the metabolism of molecules, identifying chemical structures of metabolites formed by cytochrome P450 enzyme family (CYPs). The mentioned software combines two main ideas: a literature-based pool of CYP-mediated reaction rules and the site of metabolism (SoM) prediction. This approach is relevant since a tool for the *in silico* prediction of the metabolism of xenobiotic compounds can offer key information for developing novel chemical entities with improved metabolic stability (i.e., cosmetics, drugs, agrochemicals). The GLORY web-server version is accessible at https://acm.zbh.uni-hamburg.de/glory/ (de Bruyn Kops et al.). Montanari et al. described a computational approach for predicting potential toxicity of molecules taking into account transporter proteins. These latter proteins, expressed in the liver, are crucial in drug pharmacokinetics and are important constituents of the physiological bile flow and their inhibition could be relevant to the drug-induced liver toxicity. Using a comprehensive analysis of the publicly available data, a set of classification models was developed for predicting the inhibition of the transport for a set of liver transporters deemed relevant by different regulatory agencies. The models were computationally

OPEN ACCESS

Edited and reviewed by:
Michael Kassiou,
The University of Sydney, Australia

***Correspondence:**
Simone Brogi
simone.brogi@unipi.it
Teodorico Castro Ramalho
teo@ufla.br
Kamil Kuca
kamil.kuca@fnhk.cz
José L. Medina-Franco
medinajl@unam.mx
Marian Valko
marian.valko@stuba.sk

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *08 May 2020*
Accepted: *11 June 2020*
Published: *07 August 2020*

Citation:
Brogi S, Ramalho TC, Kuca K, Medina-Franco JL and Valko M (2020) Editorial: In silico Methods for Drug Design and Discovery. Front. Chem. 8:612. doi: 10.3389/fchem.2020.00612

validated demonstrating an ability to predict the interaction profile of small molecules with liver transporters. These computational tools can assist medicinal chemists and toxicologists in prioritizing compounds at the initial steps of the development of drug candidates. The models are freely available as a web-service at https://livertox.univie.ac.at (Montanari et al.). Another work, regarding the pharmacological profiles of molecules, was presented by Sidorov et al.. They investigated the possibility to predict synergism of cancer drug combinations using NCI-ALMANAC data. This topic is of extreme interest since drug combinations could represent a promising strategy for treating cancer. The authors described an *in silico* approach to investigate drug combination synergy by exploiting the largest available dataset reporting synergism of anticancer drugs (NCI-ALMANAC, with over 290,000 synergy determinations). Two machine learning (ML) procedures, Random Forest (RF), and Extreme Gradient Boosting (XGBoost), were employed on the selected dataset. The assessment of these computational tools indicated that the prediction of the synergy of undisclosed drug combinations is a feasible task. Accordingly, by using these kind of models it will be possible to significantly reduce the number of *in vitro* tests, by evaluating *in silico* which of the selected combinations are expected to be synergistic (Sidorov et al.).

A relevant number of papers are focused on different ligand- and structure-based approaches or on a combination thereof to identify promising molecules for a given target. Accordingly, Velázquez-Libera et al. described a combined structure- and ligand-based approach for investigating the structural requirements governing the affinity of a series of molecules for the human Sigma1 receptor (S1R). This receptor represents a valuable drug target for treating neuropsychological disorders. The authors discovered an effective S1R agonist namely RC-33 as a promising neuroprotective agent. In the paper presented in this Research Topic, the authors computationally investigated the interactions of RC-33 and its novel derivatives within the S1R active site. To this end, different *in silico* techniques [docking, interaction fingerprints, and receptor-guided alignment three-dimensional quantitative structure-activity relationship (3D-QSAR)] were applied for investigating a potential mechanism of action of the developed compounds. The presented data could be useful for designing novel S1R modulators (Velázquez-Libera et al.). Wu et al. also described a combination of different computational procedures (*de novo* protein structure prediction and ligand-protein interaction simulation) to investigate the structural requirements of compounds governing the affinity for the *h*SK2/calmodulin complex. The authors developed a homology model of SK2/calmodulin in order to predict potential binding sites. The ligand-protein interaction, using a series of computational procedures, was then investigated. The obtained results confirmed that the combination of different *in silico* techniques could facilitate the drug discovery process (Wu et al.). Furthermore, some computational approaches, including 3D-QSAR, molecular docking, virtual screening (VS), ADME prediction, and molecular dynamics (MD), were used by Chen et al. to identify some HIV-1 non-nucleoside reverse transcriptase (RT) inhibitors (NNRTIs). Starting from a novel series of dihydrofuro[3,4-d]pyrimidine (DHPY) related compounds, endowed with antiviral activity, a computational investigation was performed employing 52 DHPYs. By applying sequential *in silico* methods, nine promising compounds were identified. These hit compounds could represent novel potential HIV-1 NNRTIs. Chen et al.. For identifying novel BCL-2 inhibitors from the Specs -SC- database, Tutumlu et al. employed multistep screening and filtering methods combining structure- and ligand-based techniques. The mentioned database was screened using a computational tool called "*cancer-QSAR*" and 26 toxicity QSAR models. The resulting non-toxic compounds were selected for two different target-driven approaches: (a) a molecular docking approach was applied to rank compounds considering their docking scores. Top-ranked compounds were employed in extensive MD simulations (100 ns) and biological assays; (b) the retrieved top-docking poses of each compound, derived from the subset selected by QSAR studies, were submitted to short MD simulations (1 ns), calculating their binding energies using the molecular mechanics generalized Born surface area (MM/GBSA) technique. By following this scheme, seven molecules were tested against different cancer cell lines. Four molecules were found to be able to reduce the proliferation of cancer cells, behaving as pro-apoptotic agents (Tutumlu et al.). The study performed by do Carmo et al. was also focused on BCL-2 and potential ligands based on a phenothiazine scaffold. The authors investigated some phenothiazines derivatives for their pro-apoptotic profile, performing an *in silico* study to relate their structures with their biological activities. By employing molecular docking simulation coupled to MD, the main interactions between compounds and the active site of the selected protein were highlighted. Notably, through these computational studies, the inhibition of BCL-2 by phenothiazines allowed for rationalizing the apoptosis-inducing effect on tumor cells (do Carmo et al.).

Naveja and Medina-Franco described a computational approach for selecting lead compounds from large datasets of chemical entities, acquired by high-throughput screening (HTS). They introduced the Constellation Plots as a general method for merging diverse and complementary molecular representations, to enhance the info contained in a visual representation and analysis of chemical space. This approach combines a sub-structure-based representation and classification of molecules with a "classical" coordinate-based representation of chemical space. A characteristic result of the mentioned technique is that organizing the molecules in analog series leads to the formation of groups of compounds, also known as "constellations," in chemical space. Notably, this proposed method is useful in identifying, for example, insightful and "bright" Structure-Activity Relationships (StARs) in chemical space that are simple to interpret. The authors applied the developed method on two datasets of DNA methyltransferases (DNMTs) and AKT1 inhibitors (Naveja and Medina-Franco). Alberca et al. reported a computational approach that allowed for the repurposing of old drugs as antimalarial agents. The authors developed and experimentally validated a collection of ligand-based models that are able to identify falcipain-2 inhibitors. These models were used in a VS campaign, using

two different databases (DrugBank and Sweetlead). The authors identified four potential hits to submit for biological evaluation. Among them, two drugs (odanacatib and methacycline) were confirmed as falcipain-2 inhibitors. Methacycline was found to be a non-competitive inhibitor of falcipain-2. Furthermore, the effects of both drugs on falcipain-2 hemoglobinase activity and on the growth of *P. falciparum* have been investigated (Alberca et al.). Baillif et al. reported a computational study, using a public dataset of compound-induced transcriptomic, for predicting the potential activity of compounds against 69 drug targets. The authors investigated the performances of the ML models constructed with transcriptomics data, with the computational tools generated by Morgan fingerprints. Active molecules against a given target could display comparable signatures in one or multiple cell lines, independent of the similarity in chemical structure, among the selected active chemical entities. For 25% of the tasks, RF computational tools employing transcriptomics signatures showed similar or better performances than those created by Morgan fingerprints. Compound-induced transcriptomic data offers a good chance for predicting targets based on cell response similarity, allowing to overcome the chemical space limitation of QSAR models (Baillif et al.). Shi et al. computationally investigated the SAR of some inhibitors of the dimerization process of PD-L1 by elucidating their potential binding and unbinding mechanism, using classical MD and metadynamics simulations. The contact analysis, R-group based QSAR analysis, and molecular docking provided additional insights about the SAR of these compounds. Accordingly, the outcomes of this research can be useful for optimizing compounds targeting PD-L1 (Shi et al.). Liu et al. introduced two methods for improving the selection of active molecules by using similarity information of all compounds. One technique ranks a molecule considering its highest z-score as an alternative of its highest Tanimoto index, while the other method ranks compounds by calculating an aggregated score taking into account their Tanimoto similarity related to all identified active and inactive molecules. These evaluations, performed using datasets available from PubChem, belonging to over 20 HTS studies, suggested that both approaches accomplished a ∼10% higher Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) score, compared to the classical approaches. Interestingly, the presented methods could offer an enhancement in early recognition of lead compounds during VS campaigns (Liu et al.). Lima et al. presented a computational approach for finding multi-kinase inhibitors against *Plasmodium falciparum* calcium-dependent protein kinases 1/ 4 (CDPK1 and CDPK4, respectively) and protein kinase 6 (PK6), in order to select novel multi-target compounds as antimalarials. By using shape-based and ML models, employing chemical databases of drug-like compounds, the authors identified 10 hit compounds to submit for biological evaluation. Among them, LabMol-171, LabMol-172, and LabMol-181 behaved as nanomolar antiplasmodial agents. Furthermore, LabMol-171 and LabMol-181 also inhibited *P. berghei* ookinete development, representing novel transmission-blocking agents (Lima et al.). Bühlmann and Reymond reported an approach to address the limitation of the GDB17 database (166.4 billion molecules), which contains numerous molecules that are too complex to synthesize. To this end, the authors developed the GDBChEMBL database, a small set of GDB17, which consists of 10 million molecules identified by means of the calculation of their ChEMBL-likeness score (CLscore). This subset contains compounds with higher synthetic accessibility, maintaining a comprehensive coverage of chemical space distinctive of the GDB17 database. GDBChEMBL is downloadable from http://gdb.unibe.ch; interactive chemical space map: http://faerun.gdb.tools (Bühlmann and Reymond).

Sirous et al. developed and experimentally validated an *in silico* procedure useful for hit-to-lead optimization. In particular, from micromolar HIV integrase (HIV IN) inhibitors, the authors described a computational workflow based on an *in silico* structure-based combinatorial library designing technique. The mentioned methodology is useful for combining the design of a combinatorial library and side-chain hopping with Quantum Polarized Ligand Docking (QPLD) and MD simulations. This method indicated the most valuable decorations for a promising scaffold. From this final set of optimized molecules, three representative compounds were synthesized and evaluated by *in vitro* tests. Among them, one compound was found to be an effective inhibitor of HIV IN in the low nanomolar range. Moreover, the biological characterization of the molecule showed that this compound is able to inhibit HIV-1 replication and HIV-1 IN strand transfer activity, with potency comparable to that found for Raltegravir (Sirous et al.). Ferreira et al. presented an article describing the development of cyclic imides as inhibitors of cruzain, a validated drug target of *Trypanosoma cruzi*. By using a micromolar-range cruzain inhibitor, the *in silico* optimization scheme led to the development of a non-toxic inhibitor of *T. cruzi* intracellular amastigotes in the nanomolar-range. By following the mentioned procedure, the authors identified a protocol useful for the rational design of novel trypanocidal agents targeting the cruzain enzyme (Ferreira et al.). Pallante et al. proposed a computational approach based on different *in silico* techniques such as homology modeling, molecular docking, and MD for investigating the interactions between several novel colchicine derivatives and tubulin isotype βIII. These derivatives were screened and ranked considering their binding affinity and conformational stability in the colchicine binding site. This study could be extremely relevant for rationally designing novel colchicine-based compounds as effective anticancer agents (Pallante et al.). Pavlin et al. focused their research article on the application and experimental validation of a VS protocol to identify small molecules that are able to target a particular variant of estrogen receptor alpha (ERα Y357S) that confers endocrine resistance, disease relapse, and increased mortality rates in patients affected by ER-positive breast cancer. By applying a VS procedure for screening different commercial databases, the authors identified five compounds active on recurrent Y537S ERα polymorphism in MCF7, and MDA-MB-231 breast cancer cell lines. Among the identified compounds, one of them showed selectivity for Y537S ERα, exhibiting no toxicity against breast cells. Remarkably, 4.5 μs of biased and unbiased MD was used for investigating the structural, thermodynamics, and the kinetics of these active ligands against wild type and diverse ERα variants (Y537S, Y537N, D538G). The information provided by

the mentioned study could be relevant for discovering mutant specific drug-candidates for improving breast cancer therapies (Pavlin et al.). Quan et al. investigated novel quinoline derivatives as P-glycoprotein (P-gp) inhibitors useful for counteracting the multidrug resistance, which represents a significant cause of cancer treatment failure. Among the mentioned derivatives, YS-7a was proposed as the most promising P-gp inhibitor. YS-7a blocked the P-gp transport without influencing the P-gp expression. Furthermore, YS-7a promoted the ATPase activity of P-gp in a dose-dependent manner. This compound could represent a valid starting point for developing novel derivatives that are able to treat multidrug resistant cancers (Quan et al.).

Michel et al., using different web-servers (DoGSite, FTMap, and CryptoSite) and a commercial tool (Schrödinger's SiteMap), comprehensively predicted ligand binding cavities, druggability scores, and conformationally active regions of the nucleoside diphosphates attached to the sequence-x (NUDIX) hydrolase protein family. Subsequently, a molecular docking study, employing Glide software, was carried out to assess the affinity of a subset of the ZINC FragNow database for the identified potential binding sites. This preliminary dual ranking, of druggable sites within the NUDIX protein family, was then compared with experimental hit rates acquired from biological studies. The detected correlation indicated that the described workflow could represent a valuable protocol for prioritizing targets and for excluding them in VS approaches (Michel et al.). Michel et al. presented a sequence-to-structure-based methodology for predicting drug resistance. The developed workflow produced and compared Molecular Interaction Fields (MIF), mapping the areas of energetically favorable interactions, between numerous chemical probes and the target active site. The technique appears to be appropriate for understanding changes of the three-dimensional structures and the physicochemical environment caused by mutations affecting the target active site. This approach was applied to four datasets of known HIV-1 protease sequences, displaying that it is able to correctly classify resistant and susceptible sequences given as the input. The described study is a novel step for interpreting the influence of genetic variability on the response to HIV 1 treatments (Alves et al.). Sánchez-Tejeda et al. proposed a vector analysis for measuring and defining "multitargeticity." The research on multi-target drugs could be relevant for identifying therapeutic agents to treat multifaceted diseases. The authors, considering the order and force of a ligand, described two "multi-target" indexes namely, 1 and 2. By combining the mentioned indexes, it is possible to discriminate multi-target drugs. These indexes were used for screening a chemical library of potential ligands that possess an affinity for diverse targets involved in multiple sclerosis. The application of the protocol allowed the identification of 10 molecules that could represent potential lead compounds for developing multi-target drugs (Sanchez-Tejeda et al.). Bissaro et al. applied the SuMD technique to ribonucleotide targets of pharmacological interest. SuMD is a modified MD protocol for accelerating the sampling of molecular recognition steps on a nanosecond timescale. Interestingly, they demonstrated the methodological ability of SuMD to reproduce the binding mode of viral or prokaryotic ribonucleic complexes and artificially engineered aptamers with a remarkable accuracy (Bissaro et al.).

Cavasotto and Aucar proposed a new approach for scoring results obtained from high-throughput docking (HTD) approaches. For better characterizing protein-ligand interactions, the authors proposed a quantum mechanical (QM)-based docking scoring function in order to obtain more accurate HTD results. This novel technique was investigated using 10 different drug targets belonging to various families with diverse binding site features. The output clearly demonstrated that the application of the QM scoring function could improve the performance of HTD methods (Cavasotto and Aucar). Pinto et al. presented a novel screening software, namely, CaverDock. In particular, the authors focused their studies on protein tunnels and channels that could represent promising drug targets. In fact, compounds able to hinder the entrance of substrates or release of products could be considered effective modulators of the biological activity. To this end, the influence of rigid and flexible side-chains on various substrates and inhibitors of seven unrelated drug targets was assessed. The accuracy of the software was evaluated by comparing the data found by CaverDock with experimental results obtained for the heat shock protein 90α. As a final point, CaverDock was used in a VS campaign employing anti-inflammatory and anticancer FDA-approved drugs against two drug targets [CYP450-17A1; leukotriene-A4 hydrolase (LTA4H)/aminopeptidase (AP)]. The analysis of the potential energies of binding and unbinding trajectories allowed for identifying functional tunnels. Accordingly, the presented software is a valuable computational resource useful in VS campaigns. CaverDock is accessible from https://loschmidt.chemi.muni.cz/caverdock/; web https://loschmidt.chemi.muni.cz/caverweb/ (Pinto et al.). Yuan et al. presented LigBuilder-V3, a software for *de novo* multi-target drug design. This computational tool can be useful for rationally designing and optimizing molecules with multi-target profiles. For validating the computational approach, LigBuilder-V3 was employed to design inhibitors that are able to target HIV protease and HIV RT, employing three different approaches. The resulting molecules, assessed by MM/GBSA, behaved as potential inhibitors for the selected drug targets. The software can be found at http://www.pkumdl.cn/ligbuilder3/ (Yuan et al.).

In this Research Topic two Perspective articles have been published. In the first article, Rastelli and Pinzi focused the attention on the need to obtain a valid post-docking analysis in VS campaigns. In fact, nowadays, HTD is a valuable *in silico* methodology extremely useful for rapidly identifying hit compounds for a given target. Unfortunately, HTD has some weaknesses (i.e., approximated scoring functions, limited sampling of ligand-target complexes), making docking outputs inevitably approximate. So, post-docking analyses are required to overcome these mentioned issues. The authors proposed a comprehensive method for the post-docking analysis in VS approaches, developing BEAR (Binding Estimation After Refinement), a post-docking resource this is able to refine docking poses employing MD, and re-scores ligands using (MM/PB(GB)SA). The article provides a rational perspective about the introduction of more accurate refinement and

rescoring approaches in VS to provide more reliable results (Rastelli and Pinzi). In the second Perspective article, Vásquez and González Barrios investigated the ligand efficiency (LE) metrics and how to improve it for an efficient and reliable hit identification in CADD. LE is still employed as a decision-making criteria in CADD. Furthermore, in fragment-based drug discovery (FBDD), LE metrics are extremely efficient in selecting promising "core" fragments for optimization. The authors presented a relative group contribution (RGC) model for improving FBDD. Accordingly, this approach could be useful in virtual fragment-based screening studies (Vásquez and González Barrios).

Finally, five Review articles reported a critical discussion about different computational approaches in various CADD fields. Santos et al. discussed the use of Molecular Mechanics (MM) for describing protein-ligand interactions. The authors reported a case study on the halogen bonds (XBs). Although less efficient than hydrogen bonds (HB), XBs were found to be relevant in CADD. XBs are able to improve the affinity and selectivity of a molecule against a potential drug target. Due to the limited ability of MM techniques to describe XBs, the authors indicated that a parametrization of the force-fields equations could be considered for improving the definition of XBs. This review highlighted some suggestions for parametrizing force-fields to accomplish reliable outcomes of complex non-covalent interactions (Santos et al.). Gagic et al. reviewed the *in silico* approaches employed for designing compounds behaving as kinase inhibitors. Kinases are pivotal in designing anticancer agents. They discussed and compared, considering some representative case studies, different methods such as pharmacophore modeling, MD, VS, and molecular docking for the rational design of kinase inhibitors (Gagic et al.). de Souza Neto et al. proposed a Review article for outlining the FBDD strategies exploring numerous computational strategies to apply for fragment-to-lead optimization. They considered potential fragment expansion strategies such as hot spot analysis, druggability prediction, SAR, application of ML/deep learning (DL) models for VS and some *de novo* approaches for suggesting synthesizable novel molecules. Moreover, the authors highlighted some recent case studies in FBDD, and how computational approaches were successfully used for developing lead compounds (de Souza Neto et al.). Maia et al. presented a Review article which reported a comprehensive outlook on the tasks in CADD for performing structure-based VS (SBVS). The authors compared methods and tools for SBVS employed in the modern drug development trajectory (Maia et al.). Thafar et al. discussed some established techniques using artificial intelligence (AI), ML, and DL approaches for identifying drug-target interactions (DTIs) and for predicting drug-target binding affinities (DTBA). This Review article reported an inclusive summary about the computational approaches for predicting DTBA. Notably, this review performed the first inclusive comparison analysis of *in silico* tools focused on DTBA related to AI/ML/DL (Thafar et al.).

In summary, as Guest Editors, we would like to thank all the authors and co-authors for their important contributions to this Research Topic, all the reviewers for their valuable work in evaluating the submitted manuscripts, and the editorial staff of Frontiers for their kind continued assistance. Taken together all these combined efforts allowed for the great success of this Research Topic. We expect this topic to contribute to the advancement of drug design and discovery and believe it serves as a valuable source of information and inspiration to scientists and students. The Research Topic is freely accessible through the following link https://www.frontiersin.org/research-topics/10032/in-silico-methods-for-drug-design-and-discovery.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

The authors wish to thank all the contributors of the Research Topic, reviewers and the Editorial Office of *Frontiers in Chemistry* for the helpful advice during the management of the submitted manuscripts.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Brogi, Ramalho, Kuca, Medina-Franco and Valko. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 12 June 2019
doi: 10.3389/fchem.2019.00402

GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism

Christina de Bruyn Kops[1†], Conrad Stork[1†], Martin Šícho[1,2†], Nikolay Kochev[3,4†], Daniel Svozil[2†], Nina Jeliazkova[3†] and Johannes Kirchmair[1,5,6]*

[1] Department of Computer Science, Center for Bioinformatics (ZBH), Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, Hamburg, Germany, [2] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Prague, Czechia, [3] Ideaconsult Ltd., Sofia, Bulgaria, [4] Department of Analytical Chemistry and Computer Chemistry, University of Plovdiv, Plovdiv, Bulgaria, [5] Department of Chemistry, University of Bergen, Bergen, Norway, [6] Computational Biology Unit (CBU), University of Bergen, Bergen, Norway

Computational prediction of xenobiotic metabolism can provide valuable information to guide the development of drugs, cosmetics, agrochemicals, and other chemical entities. We have previously developed FAME 2, an effective tool for predicting sites of metabolism (SoMs). In this work, we focus on the prediction of the chemical structures of metabolites, in particular metabolites of xenobiotics. To this end, we have developed a new tool, GLORY, which combines SoM prediction with FAME 2 and a new collection of rules for metabolic reactions mediated by the cytochrome P450 enzyme family. GLORY has two modes: MaxEfficiency and MaxCoverage. For MaxEfficiency mode, the use of predicted SoMs to restrict the locations in the molecule at which the reaction rules could be applied was explored. For MaxCoverage mode, the predicted SoM probabilities were instead used to develop a new scoring approach for the predicted metabolites. With this scoring approach, GLORY achieves a recall of 0.83 and can predict at least one known metabolite within the top three ranked positions for 76% of the molecules of a new, manually curated test set. GLORY is freely available as a web server at https://acm.zbh.uni-hamburg.de/glory/, and the datasets and reaction rules are provided in the **Supplementary Material**.

Keywords: metabolism prediction, metabolite structure prediction, rule-based approach, sites of metabolism, xenobiotic metabolism, cytochrome P450, metabolites

OPEN ACCESS

Edited by:
Jose L. Medina-Franco, National Autonomous University of Mexico, Mexico

Reviewed by:
Anastasia Rudik, Institute of Biomedical Chemistry, Russian Academy of Medical Sciences (RAMS), Russia
Angelica Mazzolari, University of Milan, Italy

***Correspondence:**
Johannes Kirchmair
johannes.kirchmair@uib.no
orcid.org/0000-0003-2667-5877

†Christina de Bruyn Kops
orcid.org/0000-0001-8890-2137
Conrad Stork
orcid.org/0000-0002-5499-742X
Martin Šícho
orcid.org/0000-0002-8771-1731
Nikolay Kochev
orcid.org/0000-0001-6547-3675
Daniel Svozil
orcid.org/0000-0003-2577-5163
Nina Jeliazkova
orcid.org/0000-0002-4322-6179

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *28 March 2019*
Accepted: *17 May 2019*
Published: *12 June 2019*

Citation:
de Bruyn Kops C, Stork C, Šícho M, Kochev N, Svozil D, Jeliazkova N and Kirchmair J (2019) GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism. Front. Chem. 7:402. doi: 10.3389/fchem.2019.00402

INTRODUCTION

Metabolism is responsible for creating metabolites with different physicochemical and pharmacological properties compared to those of the original parent molecule. Xenobiotic metabolism in particular is directly relevant for humans, especially as it relates to, for example, the development of drugs, cosmetics, and agrochemicals. In fact, it is supposed that metabolism is the main clearance pathway for the vast majority of all xenobiotics (Kirchmair et al., 2015). However, metabolism can also result in pharmacologically active metabolites as well as toxic metabolites (Testa et al., 2012).

The cytochrome P450 (CYP) family of enzymes plays an important role in the metabolism of xenobiotics, especially in the formation of first-generation metabolites, of which roughly 60% are formed by CYPs (Testa et al., 2012). The importance of CYPs to drug discovery is clear from the observation that many drugs are metabolized by CYPs; common estimates range from 50% (Di, 2014) to 80% (Testa et al., 2012). A detailed meta-analysis of the metabolites of over 1,000 different xenobiotic substrates carried out by Testa et al., showed that 40% of all metabolites are formed by CYPs, including a substantial proportion of all toxic or highly reactive metabolites (Testa et al., 2012).

There are 57 known human CYP enzymes, the majority of which are primarily involved in endogenous metabolism. The CYP2 and CYP3 subfamilies are mainly responsible for metabolizing xenobiotics (Testa et al., 2012), and the key CYP isozymes for drug metabolism are CYP3A4, 3A5, 2D6, 2C8, 2C9, 2C19, 1A1, 2B6, and 2E1 (Di, 2014). Among the xenobiotic-metabolizing CYP isozymes, the binding pockets vary greatly; in some cases the binding pocket of a single isozyme is highly flexible and can accommodate a broad range of substrates with widely varying sizes (Kirchmair et al., 2015).

Computational methods can make a significant contribution to predicting xenobiotic metabolism, because they can be used to quickly make predictions that can focus the experimental aspects of the drug development process. Such a focusing effect is both cost-effective and time-effective (Kirchmair et al., 2015).

One relatively well-developed aspect of the computational prediction of xenobiotic metabolism is the identification of the metabolically labile atom positions, also known as sites of metabolism (SoMs) (Kirchmair et al., 2012). Being able to predict SoMs is important because knowing an atom position in a molecule at which a metabolizing reaction is likely to occur usually provides a chemist with a good idea of the ensuing metabolite structure. Besides a range of commercial offerings, several freely available tools, such as SMARTCyp (Olsen et al., 2019), SOMP (Rudik et al., 2015), Xenosite (Zaretzki et al., 2013), and FAME 2 (Šícho et al., 2017), are able to predict SoMs with high accuracy (Tyzack and Kirchmair, 2018). FAME 2, which is used in the present work for SoM prediction, is a machine learning-based tool developed recently in our group. The extra trees classifier models of FAME 2, which are based on a set of 2D circular descriptors, were developed specifically to predict SoMs of metabolic reactions catalyzed by the CYP family of enzymes in humans. FAME 2 is highly accurate, achieving, on an independent test set, a Matthews correlation coefficient of 0.57 and an area under the receiver operating characteristic curve (AUC) of 0.91.

In contrast to *in silico* SoM prediction, computational prediction of the structures of metabolites lags behind with respect to prediction accuracy. In general, existing methods for predicting metabolite structures for xenobiotics are dominated by rule-based approaches. There are a number of well-established commercial tools for metabolite structure prediction, including Meteor Nexus (Lhasa Ltd.), a rule-based metabolite prediction software (Marchant et al., 2008). Meteor Nexus offers three different reasoning methods to prioritize the plethora of generated metabolites. The current default reasoning method is SoM scoring, which compares the SoM identified by the reaction rule to experimental data in order to assign scores to the predicted metabolites[1]. Other rule-based computational tools include TIMES (LMC; Mekenyan et al., 2004), which uses a heuristic algorithm to generate possible metabolic maps, and MetabolExpert (CompuDrug; Darvas, 1987).

In addition to commercial metabolite structure prediction tools, there is an increasing number of freely available options. Again, many of the available options rely primarily on a set of reaction rules to generate structures of possible metabolites. One well-known approach that has been around for some time is SyGMa (Ridder and Wagener, 2008), which in this work is used as a reference method. SyGMa predicts metabolites using knowledge-based reaction rules, some of which were derived from common knowledge of metabolism reactions and some of which were developed using the Metabolite Database (MDL Metabolite Database, Elsevier, 2001), for a total of 144 reaction rules covering both phase I and phase II metabolism. The predicted metabolites are ranked by empirical probability scores calculated based on the fraction of predicted metabolites produced by the particular reaction rule that match reported metabolites in the database. Using all 144 phase I and phase II reaction rules in up to three successive reaction steps, SyGMa was able to predict 68% of all known metabolites in the test set. In terms of ranking, SyGMa ranked 45% of the known metabolites in the test set in the top 10. The authors additionally examined SyGMa's potential usefulness for predicting CYP-mediated metabolism by evaluating its performance on a set of 127 single-step CYP-mediated reactions. Using only the 118 phase I reaction rules, which include but are not specific to CYP-mediated reactions, SyMGa was able to predict 84% of all known CYP-formed metabolites and predict 66% of the known metabolites within the top three ranked predicted metabolites. However, the proprietary nature of the dataset that was used to derive SyGMa's reaction rules and validate the method, not to mention the current unavailability of the dataset, hinders the reproducibility of the results as well as further use of the models derived from the data.

A recent, free software designed to predict metabolites from multiple sources and enzyme families is BioTransformer (Djoumbou-Feunang et al., 2019), which in this work is used as the second reference method. BioTransformer is a comprehensive metabolite prediction tool that contains a CYP metabolite prediction module (in addition to four other metabolite prediction modules). BioTransformer predicts CYP-formed metabolites using a knowledge-based approach combined with built-in CYP selectivity prediction by CypReact (Tian et al., 2018), a machine learning-based tool, as a precursor to metabolite prediction. Aside from the initial CYP isoform-specificity prediction, the basis of BioTransformer's CYP450 metabolite prediction module is a rule-based method whose

Abbreviations: AUC, area under the receiver operating characteristic curve; CYP, cytochrome P450; ROC, receiver operating characteristic; SoM, site of metabolism.

[1]Meteor Reasoning Methodologies, Lhasa Limited, https://www.lhasalimited.org/products/meteor-reasoning-methodologies.htm

reaction rules are derived partly from the metabolic reactions in MetXBioDB (Djoumbou-Feunang et al., 2019), a freely available database of metabolism reactions that was established in the context of developing BioTransformer. In the current version of BioTransformer, the predicted metabolites are not ranked. BioTransformer also offers an option for identifying metabolites based on masses from mass spectrometry data. On a test dataset of 60 parent molecules with a total of 180 known metabolites, BioTransformer's CYP450 metabolite prediction module achieved a recall of 0.90 and a precision of 0.46.

Another freely available metabolite prediction tool is MetaTox (Rudik et al., 2017), which encompasses both phase I and phase II metabolism and combines the prediction of the reaction class and the reacting atom in order to predict metabolites. Additionally, the open-source software Toxtree (Patlewicz et al., 2008) contains a metabolism prediction module called "SMARTCyp—Cytochrome P450-Mediated Drug Metabolism" that predicts SoMs using SMARTCyp (Rydberg et al., 2010) and then applies a small set of reaction rules to the predicted SoMs in order to predict metabolites.

Common to all modern approaches for metabolite prediction is that they remain challenged by the combinatorial explosion of predictions, in particular when looking at several generations of metabolites (Judson, 2014). It is not unusual for metabolite structure predictors to produce several pages full of predicted metabolites, a fact which is often and not without reason criticized, particularly by experts in metabolism. The key to tackling this problem lies in the development of approaches for the accurate ranking of metabolites according to their relevance in terms of metabolic rates and biological properties. A number of methods attempt to get a handle on the immense number of predicted metabolites by ranking their predictions according to various approaches.

Another option, which has primarily been implemented in commercial tools to date, is to use SoM prediction as a preliminary step to reduce the number of generated metabolites. Commercial tools for metabolite prediction that incorporate SoM prediction include ADMET Predictor (SimulationsPlus)[2], which predicts SoMs and the corresponding metabolite structures for nine CYP isoforms, and StarDrop (Optibrium; Tyzack et al., 2016), whose "P450 metabolism" module predicts SoMs using quantum mechanical simulations and displays the structures of the metabolites corresponding to the predicted SoMs. In addition, META Ultra (MultiCASE Inc.; Klopman et al., 1994) predicts SoMs and metabolites, and MetaSite (Cruciani et al., 2005) was a SoM and CYP isoform selectivity prediction software that now also predicts metabolite structures[3].

Few freely available metabolite prediction methods combine information on predicted SoMs with a rule set. MetaTox predicts reaction classes and reacting atoms (i.e., SoMs, in principle) separately for each parent molecule, then combines the predictions to generate metabolites. The probability that the metabolite is formed is calculated based on the predicted probabilities of the reaction class and of the SoM predicted with the SOMP method (Rudik et al., 2015). However, the validation of MetaTox considers the performance of the reaction class prediction and the reacting atom prediction separately, without evaluating the prediction of the metabolite structures themselves, and it is unclear how exactly the reaction class and reacting atom predictions are combined to generate a metabolite structure (Rudik et al., 2017). On the other hand, it is clear that SoM prediction is used directly as a prefilter before applying reaction rules in the SMARTCyp Toxtree module. However, a validation of this method has not been published.

In terms of the availability of rule sets for metabolite structure prediction, there are a few existing freely available collections of reaction rules described in an easily accessible, computer-readable format such as SMIRKS[4], a reaction transform language within the Daylight system. One source of CYP reaction rules is the SMARTCyp Toxtree module, which uses 16 reaction rules and makes the SMIRKS freely available as part of the source code. A larger selection of reaction rules is provided in the freely available SyGMA Python package. The reaction rules are clearly separated into phase I and phase II rules; however, there is no indication of which of the 118 phase I reaction rules specifically describe CYP-mediated reactions. In addition, these rules were derived from a proprietary and no longer distributed dataset. BioTransformer offers a large number of CYP-specific biotransformation rules in SMIRKS format as well as additional constraint(s) for each rule as part of its Reaction Knowledgebase.

In this work, we present a multipronged approach to the prediction of metabolites formed by the CYP enzyme family in humans. In reference to FAME, we name this approach GLORY. One fundamental aspect of GLORY is a new, easily interpretable rule base for CYP metabolism that was developed solely from the scientific literature and basic chemistry knowledge, without relying on any dataset of metabolic reactions. In addition, we have examined the effect of using SoM prediction as a preliminary filter for the positions at which reaction rules are allowed to be applied and also as part of a new approach to ranking the predicted metabolites. GLORY therefore has two modes: MaxCoverage, which focuses solely on recall, and MaxEfficiency, which focuses more on precision. Further, we have validated GLORY on a new, high quality, manually curated dataset that is provided in the **Supplementary Material**.

RESULTS AND DISCUSSION

Two key aspects are at the core of GLORY, which aims to predict metabolites within the context of human, CYP-mediated metabolism: reaction rules and predicted SoMs. In terms of the rule-based aspect, GLORY uses reaction rules to convert parent molecules into their possible metabolites. To this end, we developed a collection of rules based entirely on the scientific literature to ensure that the rule set was not biased by any particular metabolism dataset. The information on the CYP-mediated reactions from the literature was combined with

[2]ADMET Predictor Metabolism Module, SimulationsPlus, https://www.simulations-plus.com/software/admetpredictor/metabolism/

[3]MetaSite, https://www.moldiscovery.com/software/metasite/

[4]SMIRKS—A Reaction Transform Language, Daylight, http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html

basic chemistry knowledge to develop SMIRKS to describe each reaction type. In some cases, such as for O-dearylation, multiple SMIRKS were required for a single reaction type, resulting in a total of 73 SMIRKS for the 61 reaction types present in our collection (**Supplementary Table 1**). We additionally use a simple binary distinction between common and uncommon reaction types, which were thoroughly discussed and distinguished from each other in Guengerich (2001), and which distinction we were able to extrapolate to the CYP-mediated reactions found elsewhere in the literature (see Methods for details). We do not use occurrence ratios calculated based on a given dataset in order to rank the predicted metabolites, due to the limited size, quality, and accessibility of existing datasets. Out of our collection of 61 CYP reaction types, 22 have been designated as common.

The second key aspect of GLORY is its use of the SoM probabilities predicted by FAME 2 for each heavy atom in a molecule to (i) reduce the false-positive prediction rate while maintaining an acceptable recovery rate and (ii) augment the ranking of predicted metabolites. In order to reduce the false-positive prediction rate, the possibility of utilizing a hard cutoff based on SoM probabilities was explored. This cutoff was used to determine at which atom positions the rules were allowed to be applied. In the context of GLORY, we have called this approach, in which SoM prediction is used as a preliminary filter, MaxEfficiency mode. In contrast, we designate the approach in which SoM probabilities are used for ranking metabolites derived for all positions in a molecule regardless of SoM probability the MaxCoverage mode. The difference in workflow between the two modes is illustrated in **Figure 1**.

Datasets

To choose a SoM probability cutoff for the MaxEfficiency mode and develop a priority score to rank predicted metabolites, a large reference dataset was generated by combining the CYP metabolism data extracted from DrugBank (Wishart et al., 2018) and MetXBioDB. MetXBioDB is a recently published database of metabolic reactions, whose substrates are mainly comprised of xenobiotics and also include a few sterol lipids and mammalian primary metabolites, and whose reaction data came from the scientific literature as well as publicly available databases (Djoumbou-Feunang et al., 2019). In addition, a manually curated, high-quality dataset was compiled from the scientific literature for the validation of GLORY. This test dataset contains 29 parent molecules and a total of 81 metabolites, resulting in 2.79 metabolites per parent molecule on average. Importantly, any parent compounds that are in the test dataset were removed from the reference dataset before any analysis occurred. In total, the reference dataset contains 848 parent molecules and a total of 1,588 metabolites, for an average of 1.87 metabolites per parent molecule. Predictions could be made for 847 of 848 molecules in the reference dataset (one molecule could not be processed successfully with FAME 2; see Methods for details).

MaxEfficiency Mode: Selection of a Cutoff for Metabolite Structure Generation Based on SoM Probability

In order to determine the effect of a SoM prediction-based prefilter on predicting preferably only the most relevant metabolites and reducing the number of false positive predictions, we tried several different cutoffs for the SoM probability that must be achieved by at least one atom involved in the reaction (as defined by the reaction's SMIRKS). For each heavy atom in a molecule, FAME 2 reports a probability between 0 and 1, corresponding to the fraction of trees of the extra trees classifier that predict that a particular atom is a SoM. The decision threshold in FAME 2 for whether or not an atom is considered likely enough to be a SoM to be designated as such was determined by the trained model to be 0.4 (Šícho et al., 2017).

We examined the effect of different SoM probability cutoffs using the reference dataset and selected the cutoff to be used

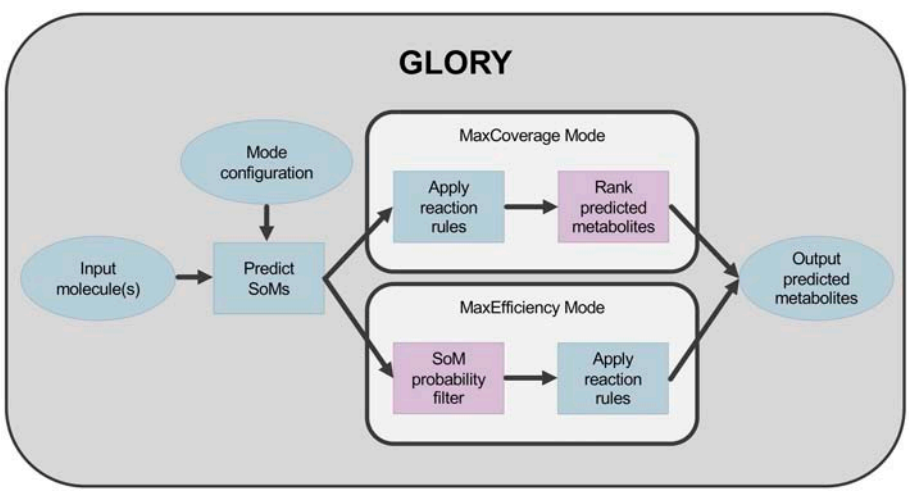


FIGURE 1 | Workflow for GLORY indicating the difference between MaxCoverage mode and MaxEfficiency mode.

in MaxEfficiency mode based on these results. In particular, we inspected the effect of the SoM probability cutoffs on precision and recall, which are defined as follows:

$$Recall = TP / (TP + FN)$$
$$Precision = TP / (TP + FP)$$

where TP is the number of true positive predictions, FP is the number of (putative) false positive predictions, and FN is the number of false negative predictions. In other words, recall measures the portion of known metabolites that were reproduced by the method and precision measures the fraction of all predicted metabolites that are represented in the dataset.

Here it is worth noting that the number of false positives, and the designation of a prediction as false positive, is especially dependent on the dataset that is being used for comparison. Many metabolites that are formed in humans have not yet been discovered, or their structures have not yet been exactly elucidated. Since even the highest-quality dataset is limited by the available experimental data, the reality is that the distinction between a real false positive prediction and the true positive prediction of an as yet unknown or unconformed metabolite may not be possible. Nevertheless, with this caveat, we evaluate our method based on the available data, including the putative false positives.

The purpose of the MaxEfficiency mode is to use the SoM probability cutoff to predict metabolites with increased precision compared to no cutoff (i.e., MaxCoverage mode). At the same time, however, we did not want to sacrifice too much in terms of recall, as it is still important to predict a molecule's actual metabolites even while reducing the number of putative false positive predictions.

For the purpose of metabolite prediction, we found that using FAME 2's decision threshold of 0.4 as the cutoff for SoM probability resulted in a relatively low recall of 0.65 (especially when compared to the recall of 0.83 achieved in MaxCoverage mode, as will be discussed later in this work). Hence, despite the increased precision afforded by a cutoff of 0.4, it was determined that this cutoff too greatly reduced the achieved recall. We therefore additionally tested lower SoM probability cutoffs (**Table 1**). Observing the trade-off between precision and recall with cutoffs ranging from 0.4 to 0.1 and comparing them to MaxCoverage mode, we determined that a SoM probability cutoff of 0.2, which resulted in a precision of 0.19 and a still-high recall of 0.75, offered the best compromise. A SoM probability cutoff of 0.2 for MaxEfficiency mode was therefore fixed based on the results shown in this section. Note that although all of the precision values shown in **Table 1** are quite low, the precision of GLORY using a SoM probability cutoff is comparable to the precision of existing methods for metabolite structure prediction (see below for the results on the test dataset).

TABLE 1 | Effect of different SoM probability cutoffs on precision and recall over the entire reference dataset.

SoM Probability Cutoff[a]	0.4	0.3	0.2	0.1	None
Precision	0.24	0.22	0.19	0.13	0.07
Recall	0.65	0.71	0.75	0.80	0.83

[a]*Note that 0.4 is the default decision threshold in FAME 2, a cutoff of none corresponds to MaxCoverage mode, and a cutoff of 0.2 was chosen for MaxEfficiency mode.*

Development of a Priority Score to Rank Predicted Metabolites for MaxCoverage Mode

In order to rank the predicted metabolites for a particular molecule, we developed a priority score for each predicted metabolite based on the SoM probability of the atoms involved in the transformation and whether the reaction type is common or not. Specifically, the SoM probability calculated by FAME 2 for all atoms in the parent molecule that are involved in a reaction as defined by the SMIRKS is considered, and the maximum SoM probability among these atoms is then incorporated into the score, as illustrated in **Figure 2**. The priority score was calculated using a simple formula:

$$score_{\text{predictedmetabolite}} = P \times F$$

where P is the maximum SoM probability out of the atoms in the parent molecule that were matched by the applied transformation and F is the factor according to whether the reaction type was designated as common or uncommon. In case the same predicted metabolite resulted from multiple transformations, the maximum priority score over all transformations leading to that prediction was used. A higher priority score is intended to indicate a higher likelihood of the prediction being true. For all uncommon reaction types, $F = 1$. The factor F for common reaction types affects the early enrichment of the predictions. Specifically, the early enrichment improves when common reaction types are given more weight in the score than uncommon reaction types, i.e. $F_{common} > 1$ (**Figure 3**). Based on an analysis of the receiver operating characteristic (ROC) curves and area under the ROC curves (AUC) for varying F_{common}, shown in **Figure 3**, a factor of 5, resulting in an AUC of 0.90, was chosen. All subsequent results based on ranking the predicted metabolites therefore used $F_{common} = 5$ in the calculation of the priority score, and the priority score can therefore range from 0 to 5.

Comparison of Performance on a New, Manually Curated Test Set

The performance of the MaxEfficiency and MaxCoverage modes of GLORY was evaluated on the curated test set of 29 parent molecules with a total of 81 metabolites. This evaluation includes a comparison with BioTransformer and SyGMa as well as an analysis of how well the scoring and ranking aspects of the different approaches work. Specifically, we employed the CYP450 module of BioTransformer and the phase I metabolism reactions of SyGMa (SyGMa does not feature a dedicated module for CYP metabolism, but phase I metabolism is carried out to a significant extent by CYP enzymes) for the comparison.

Analysis of MaxEfficiency Mode

GLORY's MaxEfficiency mode was designed to address the problem of low precision caused by a high number of putative

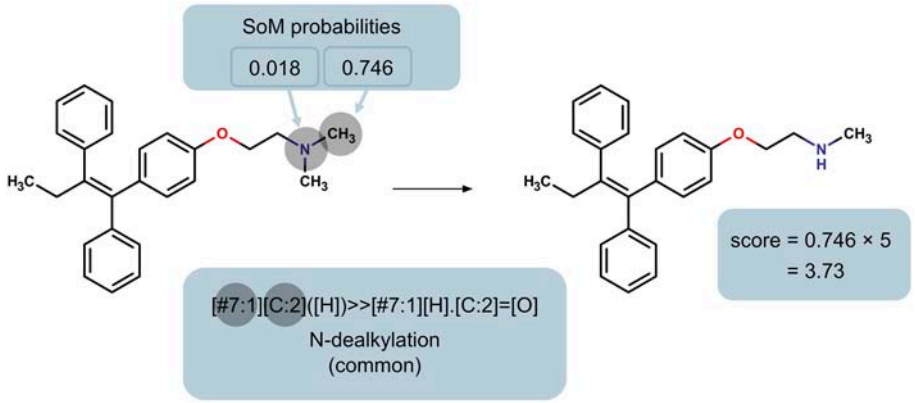


FIGURE 2 | Illustration of the determination of the maximum SoM probability of all heavy atoms in the parent molecule that are matched by the reaction rule, using the N-dealkylation reaction rule (common reaction type; factor $F = 5$) as an example. This maximum probability is used to calculate the priority score of the product.

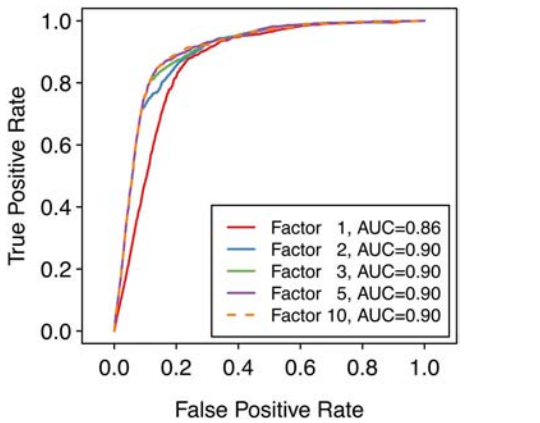


FIGURE 3 | Receiver operating characteristic (ROC) curves over the entire reference dataset of 848 compounds with 1,588 known metabolites, with varying values of the factor used for common reaction types when calculating the priority score for each metabolite. Note that a factor of 1 means that only the SoM probability (i.e., the maximum SoM probability for all atoms that are matched by the SMIRKS) affects the priority score of the predicted metabolite, regardless of the reaction type. Note also that a ROC curve can be calculated despite there being no "true negative" predictions overall (all predicted metabolites are "positive" predictions). To generate the ROC curve, the false positive rate (FPR) is calculated at each score threshold. At each point, predictions with scores below the threshold are considered "negative" predictions and predictions with scores above the threshold are considered "positive" predictions. Hence the number of "true negative" predictions and therefore the FPR can be calculated for each point of the ROC curve.

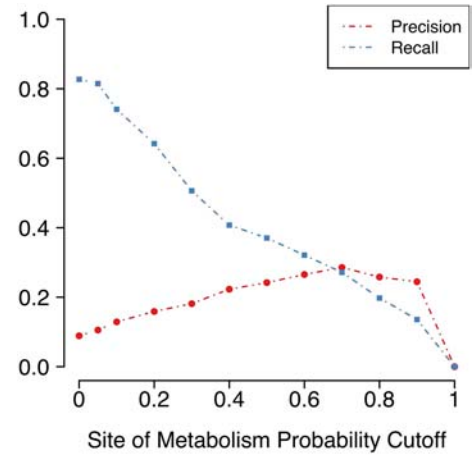


FIGURE 4 | Precision (portion of predictions that are true positives) and recall (portion of known metabolites that are predicted) vary according to the cutoff for FAME 2's predicted SoM probability. A SoM probability cutoff of 0.4 corresponds to the decision threshold used in FAME 2. The SoM probability cutoff chosen for the MaxEfficiency mode of GLORY was 0.2.

false positive metabolite predictions. This general problem of an excess of predictions is well-documented for metabolite prediction tools (Judson, 2014). However, as mentioned above, it is important to note that the designation of predictions as false positive is particularly dataset-dependent.

As described previously, the MaxEfficiency mode uses a cutoff based on the SoM probabilities that FAME 2 predicts for each heavy atom in order to restrict the locations in the molecule at which the reaction rules are allowed to be applied. This SoM probability cutoff was set to 0.2 based on the analysis on the reference dataset; however, we also examine the effect of different SoM probability cutoffs using the high-quality test dataset in order to get a more complete picture of how much can be gained by a cutoff-based approach.

As expected, using SoM predictions to confine the application of reaction rules to certain positions does involve a trade-off between precision and recall (**Figure 4**). Recall measures the portion of known metabolites that the method was able to reproduce, and precision measures the fraction of all predicted metabolites that are actually known metabolites (see previous section for definitions). The larger the SoM probability required to be present among the atoms involved in the transformation, the lower the recall but the higher the precision as measured across the entire test dataset. In addition, the larger the SoM probability cutoff, the more parent molecules there are for which

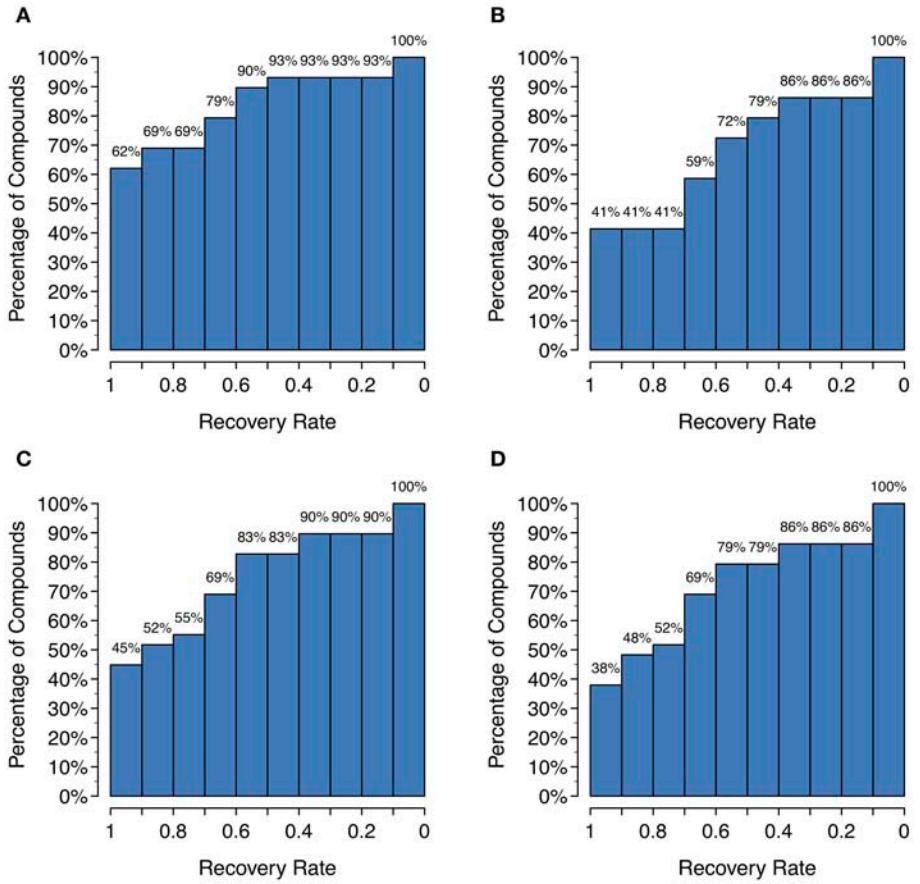


FIGURE 5 | Histograms of the recovery rate of known metabolites broken down by parent compound: **(A)** GLORY in MaxCoverage mode, **(B)** GLORY in MaxEfficiency mode, **(C)** SyGMa, **(D)** BioTransformer. For example, a recovery rate of 0.5 indicates that for x% of all parent molecules, at least half of all recorded metabolites from the test dataset were predicted.

no metabolite predictions can be made. Without any such cutoff and even up to a SoM probability cutoff of 0.2, metabolites can be predicted for all parent molecules in the test dataset. However, with a SoM probability cutoff of 0.3, no metabolites are predicted for two parent molecules, and this number increases to three for a cutoff of 0.4 (**Supplementary Table 2**). The number of molecules affected is small in this case, yet is approximately 10% of the size of the test dataset. Overall, as the cutoff increases, the total number of predicted metabolites decreases drastically (**Supplementary Table 2**).

Unfortunately, as **Figure 4** shows, there is a large decrease in recall for a small increase in precision when using SoM probability cutoffs of 0.1 or greater. Looking more closely at the recovery rates per parent molecule, we see that GLORY's MaxEfficiency mode (using the selected cutoff of 0.2 as described above) can predict at least half of the known metabolites for 72% of the parent molecules in the test dataset, as opposed to 83% for SyGMa and 79% for BioTransformer (**Figure 5**). GLORY's MaxEfficiency mode can predict all known metabolites for 41% of the parent molecules in the test dataset, as opposed to 45% for SyGMa and 38% for BioTransformer. On the other hand, the number of putative false positives per parent molecule is brought to within the same range as was measured for SyGMa and BioTransformer (**Figure 6**). Using MaxEfficiency mode, most parent molecules have fewer than 10 putative false positives, which is also the case for BioTransformer but not quite the case for SyGMa (however, as mentioned above, SyGMa's rule base also includes rules for non-CYP-mediated phase I reactions).

Based on these results, it appears that using FAME 2's predicted SoM probabilities as a hard cutoff for metabolite prediction may not be sufficient for many use cases. However, the SoM predictions are useful for more than just as a hard cutoff, namely to rank the predicted metabolites, as will be shown in the next section.

Comparison of MaxCoverage Mode to SyGMa and BioTransformer

Neither SyGMa nor BioTransformer uses regioselectivity prediction as a prefilter before applying reaction rules. The same is true of MaxCoverage mode, which only uses SoM prediction in order to score and rank the predicted metabolites. Hence, we compared SyGMa and BioTransformer to GLORY's

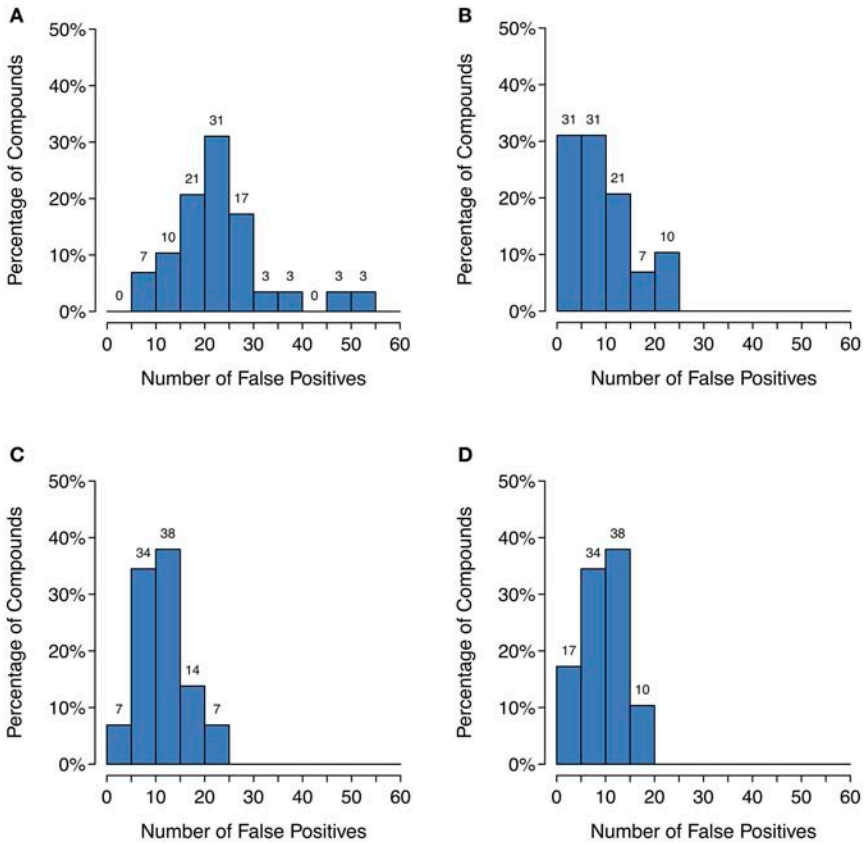


FIGURE 6 | Histograms of the number of putative false positive predictions: **(A)** GLORY in MaxCoverage mode, **(B)** GLORY in MaxEfficiency mode, **(C)** SyGMa, **(D)** BioTransformer. These histograms use right-closed intervals.

MaxCoverage mode in terms of recall, precision, and ability to rank the predicted metabolites.

A high recall is important for any use case of a metabolite structure predictor, but even more so for applications in which it is of utmost importance to not miss any physically existing metabolites, such as, for example, when attempting to identify metabolites based on MS data. GLORY's MaxCoverage mode performs well in terms of recall, with a recall of 0.83 compared to 0.74 and 0.72 for SyGMa and BioTransformer, respectively, across the entire test dataset (**Table 2**). A closer look at recall broken down to the level of the recovery rate of known metabolites for each parent molecule shows that GLORY is able to predict all known metabolites for 62% of the parent molecules, whereas SyGMa and BioTransformer achieve only 45% and 38%, respectively, in this regard (**Figure 5**). The number of parent molecules for which GLORY is able to predict at least half of the known metabolites is 90%, compared to 83% for SyGMa and 79% for BioTransformer (**Figure 5**).

Precision can be a useful metric for measuring how well a method is able to keep the number of putative false positive predictions under control. Precision was low across the board for metabolite prediction on the test dataset, with BioTransformer reaching the highest precision of the three tools at 0.17. SyGMa was close behind at 0.15, and GLORY's MaxCoverage mode lagged further behind at a precision of only 0.08 (**Table 2**). Again breaking this down to a slightly more detailed overview, we see that BioTransformer and SyGMa both always produce fewer than 25 putative false positives per parent molecule and, for the majority of parent molecules, fewer than 15 putative false positives or even, in the case of BioTransformer, fewer than 10 (**Figure 6**). GLORY in MaxCoverage mode, on the other hand, often produces so many predictions per parent molecule that there are up to 53 putative false positives per parent molecule in the test dataset and on average a relatively high number of putative false positive predictions compared to the other two tools (**Figure 6**).

In the case of the low precision observed for SyGMa, it is important to note that SyGMa's rule set is not specific to CYP-mediated metabolism but rather covers phase I metabolism in general. This could indicate that SyGMa might achieve higher precision if only the CYP-specific rules were used.

BioTransformer's CYP450 prediction module, which has the highest precision of all three methods, uses isoform prediction as a preliminary filter. Only the relevant reactions for the predicted metabolizing CYP isoform(s) are applied to the parent molecule, which could contribute to the observed precision.

Although the precision of MaxCoverage mode (as well as SyGMa and BioTransformer) was found to be low and high rates

TABLE 2 | Evaluation results for SyGMa, BioTransformer, and GLORY's MaxCoverage and MaxEfficiency modes on the manually curated test dataset.

	SyGMa	BioTransformer	GLORY, MaxCoverage mode	GLORY, MaxEfficiency mode[b]
Precision	0.15	0.17	0.08	0.16
Recall	0.74	0.72	0.83	0.64
Total number of predicted metabolites	406	344	793	327
Number of successfully predicted reported metabolites[a]	60	58	67	52
Top-1	0%	N/A	68.97%	68.97%
Top-2	48.28%	N/A	72.41%	72.41%
Top-3	68.97%	N/A	75.86%	75.86%

[a] *The total number of reported metabolites in the dataset was 81.*

[b] *The SoM probability cutoff used for MaxEfficiency mode is 0.2, chosen based on the results of the analysis on the reference dataset. Data on the performance of MaxEfficiency mode with different SoM probability cutoffs are reported in* ***Supplementary Table 2****.*

of false positive predictions are problematic in general, in the case of metabolite structure predictors a low precision is only problematic if there is no way to distinguish between the true positive and putative false positive predicted metabolites. This distinction can be achieved with a well-working ranking of the predicted metabolites, which circumvents the need to reduce the total number of predicted metabolites. Hence it is important that a metabolite prediction tool can rank the predicted metabolites in terms of likelihood of occurrence.

GLORY scores its predicted metabolites based partly on the maximum SoM probability of all the atoms involved in the reaction and also takes the type of reaction into account (see above for a more detailed description of the priority score). SyGMa uses empirical probability scores calculated based on the percentage of all predictions for each reaction rule that are found in the training dataset. SyGMa's scoring system thereby relies entirely on the discontinued Metabolite dataset. The scores generated by GLORY or by SyGMa can be used to rank the predicted metabolites for a given parent compound in terms of their likelihood of occurring. The current version of BioTransformer, on the other hand, does not score or rank its predictions.

We compared the ranking capability of GLORY's MaxCoverage mode with that of SyGMa. SyGMa was able to predict a known metabolite within the top three ranked positions for 69% of the parent molecules in the test dataset, whereas GLORY's MaxCoverage mode predicted a known metabolite within the top three predictions for 76% of the parent molecules (**Table 2**).

To look at the overall quality of the scoring as well as the ranking ability of SyGMa compared to GLORY, we generated ROC curves for each method using the score of each predicted metabolite as well as the rank of each predicted metabolite for a given molecule. The rank-based analysis corresponds better to the actual use case, in which it is desired to prioritize the predicted metabolites for a particular parent molecule, as opposed to over an entire dataset [note that SyGMa was originally only evaluated in terms of ranking per parent molecule (Ridder and Wagener, 2008)]. However, we additionally used the score-based ROC curve to visualize the performance of GLORY's priority score across the whole test dataset. To better allow for comparison of the ROC curves, false negatives were included in the ROC curves and thereby in the calculated AUCs by adding those molecules to the set of data points and artificially assigning them a score of 0 or rank of 1,000, as applicable, for the purpose of this evaluation.

Though the AUC values are low, due in part to the inclusion of false negative data points in the ROC curves, the ROC curves show a much better earlier enrichment for GLORY than for SyGMa (**Figure 7**). SyGMa does not rank a known metabolite in the best-ranked position for any parent molecule in the test dataset (**Table 2**), which is reflected in the ROC curve. This decent early enrichment with GLORY, which is corroborated by the top-3 value, is a highly encouraging result indicating that the most likely predictions are closer to the top of the ranked list than the putative false positive predictions are.

One possible explanation for why SyGMa performs poorly in terms of scoring could be that its scoring scheme was derived from occurrence ratios in the Metabolite database and therefore optimized to predict the metabolites in that particular dataset. Although the Metabolite database was large, the authors of SyGMa report that the database was nevertheless biased toward compounds with one known metabolite and postulate that many of the metabolite profiles were incomplete (Ridder and Wagener, 2008). Our manually curated test dataset consists of parent molecules with metabolites that have been published since 2014, while SyGMa was developed using the 2001 version of Metabolite, so we assume that the overlap, if any, between SyGMa's training dataset and our test dataset is low. Without access to the dataset that was used to develop SyGMa's scoring methodology, it remains unclear how well the types of the reactions that lead to the metabolites in the test dataset were represented in their training dataset. Related to that, an additional downside of SyGMa's approach of basing their scoring approach on a database of metabolic reactions is that, since reaction rules can only be included if the database contains enough examples of a specific reaction type to calculate

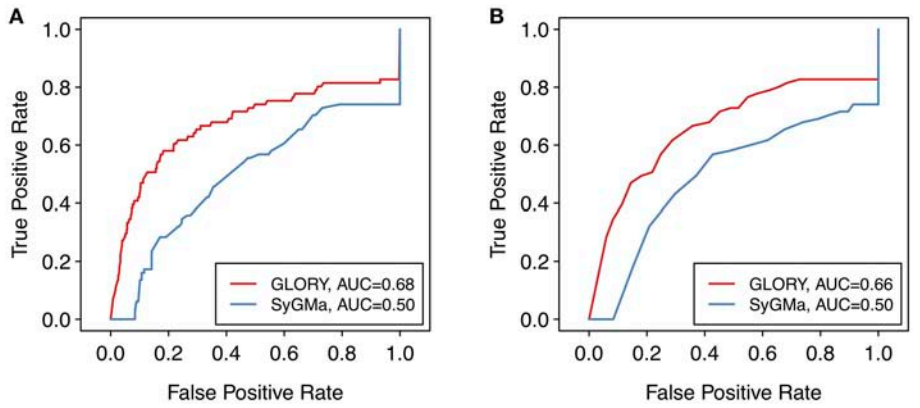


FIGURE 7 | ROC curves over the entire test dataset comparing the **(A)** scoring and **(B)** ranking approaches of SyGMa to GLORY's MaxCoverage mode. For a better comparison of the two methods, false negatives were included in the ROC curve by assigning those data points a score of 0 or rank of 1,000, as applicable.

a probability score, more unusual reaction types or reaction types that are for some reason not well enough represented in the database may be missing from SyGMa's rule base (Ridder and Wagener, 2008).

There are several other differences in methodology between GLORY and SyGMa that could contribute to the difference in performance. Firstly, SyGMa does not specifically predict CYP-mediated metabolism but rather phase I metabolism in general, meaning that it could predict other phase I metabolites that are simply not present in the test dataset because they are not formed by CYPs. Second, in the current Python package implementation that was used for this validation, SyGMa does not appear to require its predicted metabolites to have a certain minimum size. Unlike GLORY, which does not output a potential metabolite if it has fewer than three heavy atoms, SyGMa predicts a handful of metabolites (across the whole test dataset) with only one or two heavy atoms.

Computation Time

The run time for GLORY was measured on a workstation equipped with eight Intel(R) Core(TM) i7-4790 CPUs, 32 GB of main memory, and a Linux operating system. For the test dataset, the total run time (using eight cores) was 4.6 min in MaxCoverage mode and 4.3 min in MaxEfficiency mode (averaged over three runs). On average, the computation time per molecule required to predict metabolites was 10.9 s for MaxCoverage mode and 10.3 for MaxEfficiency mode (averaged over three runs).

METHODS

Development of a Collection of Transformations

A collection of transformations, defined by SMIRKS and representing reaction types, was assembled based on known CYP-mediated reactions found in the literature (see **Supplementary Material** for details). The SMIRKS were defined to be as general as possible while being restricted to reasonable reaction chemistry, as indicated by the literature and common chemical knowledge. Therefore, if a reaction was found in the literature but it was not clear how the reaction would apply to other molecules besides the provided example, the reaction was excluded from the collection. This was the case for most reactions involving large ring systems as well as ring fusions and ring contractions. Specifically, the following types of reactions were excluded from our collection: reactions that appeared to be singleton reactions, reactions involving more than two fused rings that are not part of a steroid backbone, ring fusions, ring contractions, reactions in which the substrate or product is a radical, and reactions specifically indicated to have been found only in the case of plant CYP isozymes.

A few of the SMIRKS used to describe the transformations were taken from the Toxtree SMARTCyp module[5]. Most of the SMIRKS, however, were newly developed specifically for GLORY. When developing the SMIRKS expressions, care was taken to include as few atoms as possible in the explicit mapping, since SoM probabilities were considered for all atoms in the mapping.

Each reaction type was designated as either "common" or "uncommon." Whenever possible, this label was assigned according to the reaction's classification by Guengerich in his 2001 review of CYP-mediated reactions (Guengerich, 2001), which explicitly divided the reactions into these two categories. If the reaction type was not described in that publication, a "common" or "uncommon" label was chosen based on extrapolation (on the basis of empirical similarity to reaction types present in the publication).

Our collection of CYP reaction rules consists of 61 reaction types. In some cases, multiple transformations were required to describe the same reaction type, leading to a total of 73 transformations in the collection of defined reactions. A full list of the reaction types and their SMIRKS can be found in **Supplementary Table 1**.

Metabolite Prediction Program

Predicting the structures of the metabolites involves applying the reaction rules at all relevant positions. The relevant positions

[5]Toxtree Module: SMARTCyp—Cytochrome P450-Mediated Drug Metabolism, http://toxtree.sourceforge.net/smartcyp.html

are determined by the reaction rule itself and, in the case of the MaxEfficiency mode, by the SoM probability predicted for each heavy atom. In MaxCoverage mode, the SoM probabilities are also used to score the predicted metabolites.

SoM Prediction With FAME 2

The SoM predictions were carried out using the FAME 2 software (Šícho et al., 2017), which included preprocessing of the molecules. The circCDK_ATF_6 trained model, which had the best average performance during the independent test set validation in Šícho et al. (2017), was used for the SoM prediction within GLORY.

Application of Transformations

The transformations of parent molecules into predicted metabolites based on the defined SMIRKS strings were performed using Ambit-SMIRKS [Kochev et al., 2018; Ambit-SMARTS Java Library, version 3.1.0. http://ambit.sourceforge.net/smirks.html (accessed Oct 4, 2017)]. Some transformations may result in multiple products. Products that contain fewer than three heavy atoms are not included in the set of predicted metabolites generated by GLORY.

When SoM prediction is used as a preliminary filter, a transformation rule is only applied at a particular location in the parent molecule if one of the heavy atoms involved is predicted to be a SoM with a probability over a certain threshold (see Results for more information on this threshold).

Scoring of Predicted Metabolites

The scoring of the predicted metabolites was based on SoM probability predictions and whether the reaction type was designated as common or uncommon. Each atom in the parent molecule was assigned a likelihood of being a SoM by FAME 2. When applying the transformations defined by SMIRKS, Ambit-SMIRKS maps the reactant portion of the defined transformation to any matching set of atoms in the parent molecule. Within this mapping, the maximum SoM probability was calculated and used to score the predicted metabolite that resulted from this particular transformation and mapping.

For each predicted metabolite, the priority score is calculated by multiplying the maximum SoM probability within the mapping with a factor F depending on whether the reaction type was classified as "common" or "uncommon." Priority scores for the predicted metabolites therefore range from 0 to F_{common}. The higher the score, the more likely the predicted metabolite is considered to be. See Results for further details on the selection of values for F.

If multiple transformations of a given parent molecule lead to the same metabolite structure, the priority score is calculated separately in each case and the highest score is retained. To calculate top-k values and rank-based ROC curves, it was necessary to rank the predicted metabolites for each parent molecule based on their priority scores. If different metabolites of the same parent compound have the same priority score, then they receive the same rank. In the case of a tie, one or more rank numbers, according to the number of tied predictions, following the tied rank are skipped. For example, if the highest score is 2.5 and two predicted metabolites both have this score, then both of these metabolites are assigned a rank of 1, no predicted metabolite is assigned a rank of 2, and the predicted metabolite(s) with the next highest score are assigned the rank of 3.

Program Output

The predicted metabolites are provided as an SD file with the following information for each predicted metabolite: rank (out of all predicted metabolites for a particular parent molecule), priority score, reaction name, and the InChI, SMILES, and ID of the parent molecule. If multiple transformations led to the same product, the highest priority score and the corresponding reaction name are reported. If the input consists of multiple molecules, the ID of a parent molecule is set to the molecule's position in the ordered list of input molecules (i.e., its position in the input file).

Creation of the Reference Dataset

The reference dataset was made by combining the CYP metabolism data from DrugBank and MetXBioDB. The total size of the combined reference dataset, not including any metabolism information for any of the parent molecules contained in the manually curated test dataset, is 848 parent molecules and 1588 metabolites (an average of 1.87 metabolites per parent molecule).

DrugBank Dataset

The DrugBank database (DrugBank, version 5.1.2. https://www.drugbank.ca/ [accessed Jan 14, 2019]) was downloaded from the website. In addition to the database in XML format, the structures of all of the molecules, both parents and metabolites, were downloaded in SD format from the website (drug group "All" for the parent molecules).

Any parent or metabolite molecule without an available structure was ignored. One parent compound (DrugBank ID: DB09327) was ignored because its SMILES had two components of which the main component could not be unambiguously identified. All available generations of metabolism reactions were considered, as long as the reaction was annotated as mediated by one or more CYP isozymes. The enzymes for the reactions listed in DrugBank do not have any apparent species information, so all were assumed to be human and thereby relevant for this dataset.

For all CYP-mediated reactions, the reactant was considered to be the parent molecule and the product was considered to be a first-generation metabolite of that particular parent molecule. Any metabolite with the same InChI, ignoring stereochemistry information, as its parent molecule was removed from the set of metabolites for that parent molecule. Only those parent molecules with at least one valid metabolite were included in the final dataset.

Finally, the six parent molecules that are also present in the manually curated test dataset were removed from the DrugBank dataset prior to any evaluation, along with their corresponding metabolism information. These parent compounds were bupropion, ticlopidine, imipramine, ifosfamide, bosentan, and olanzapine.

After preprocessing, including removal of the overlap with the manually curated test dataset, the DrugBank dataset contained

364 parent molecules and 702 metabolites in total, with an average of 1.93 metabolites per parent molecule in the dataset.

MetXBioDB Dataset

The human, CYP-mediated reactions were extracted from the MetXBioDB dataset (MetXBioDB, version 1.0. https://bitbucket.org/djoumbou/biotransformerjar/src/master/database/ [accessed Jan 11, 2019]). As the only structural information provided in the MetXBioDB is in the form of InChIs and InChIKeys, any substrate or product without a reported InChI could not be considered. A lacking InChI was only the case for one out of 1468 CYP-mediated, human reactions in MetXBioDB.

Stereochemistry information was removed by generating InChIs without a stereochemistry layer, resulting in 751 CYP, human parent compounds in total. Of these, 259 are also present in the DrugBank dataset. For these overlapping parent compounds, 512 of 569 DrugBank metabolites are also in MetXBioDB, and MetXBioDB has an additional 93 metabolites for these overlapping parent compounds.

Eight parent compounds (olanzapine, bupropion, metoclopramide, bosentan, imipramine, ticlopidine, ifosfamide, and atomoxetine) from the manually curated test dataset were also present in the MetXBioDB dataset, only two of which (metoclopramide and atomoxetine) were not also present in the DrugBank dataset. These parent compounds and the corresponding metabolism data were removed from the MetXBioDB dataset.

After preprocessing, including removal of the overlap with the manually curated test dataset, the MetXBioDB dataset contained 743 parent molecules and 1385 metabolites in total, with an average of 1.86 metabolites per parent molecule in the dataset.

Merger of the DrugBank and MetXBioDB Datasets

The DrugBank dataset and the MetXBioDB dataset were combined to form the reference dataset via a straightforward consolidation of the parent and metabolite information. All molecule comparisons occurred using InChIs generated without stereochemistry information. For any parent molecule that was present in both the DrugBank and the MetXBioDB datasets, which was the case for 259 parent molecules, the sets of metabolites from both datasets were combined, disregarding stereochemistry, to yield the final set of metabolites for that parent molecule in the reference dataset.

Creation of the Manually Curated Test Dataset

A new dataset for testing GLORY was manually assembled from the scientific literature. The data were extracted from publications on metabolism that were found in two journals: *Xenobiotica* and *Drug Metabolism and Disposition*. The time frame considered was from January 2014 to June 2018 for *Xenobiotica* and from January 2014 to June 2017 for *Drug Metabolism and Disposition*.

Publications were chosen and the metabolism information they contain included in the dataset if the following criteria were fulfilled:

1. The publication must contain a figure that depicts the metabolism scheme and includes the chemical structures of the parent compound and the first-generation metabolites.
2. The metabolism data must have been experimentally determined from a human source (i.e., either humans, human cells, or recombinant human CYP enzymes). If some but not all of the data were from humans, any non-human metabolites in the metabolism scheme were excluded from the dataset.
3. For at least 75% of all of the first-generation human metabolites depicted in the metabolism scheme (note that any metabolite that is depicted as merely being an intermediate is not considered), the following two criteria must be satisfied. First, the identity of the enzyme(s) responsible for the formation of the metabolite must be known. For this purpose, it is sufficient to know whether or not this metabolite is formed by CYPs. Second, the exact chemical structure, including the connectivity of all atoms, of the metabolite must be known. There is one exception to this rule: If the metabolite is known to not be CYP-formed, then the exact structure is not relevant and the metabolite is counted anyway.

Based on these criteria, 29 metabolism schemes containing at least one human, CYP-formed first-generation metabolite with a fully defined structure were found and included in the dataset. For these 29 parent molecules, there are 81 metabolites in total that fulfill the criteria (first-generation, human, CYP, fully defined structure) for inclusion in the dataset. Note that only first-generation metabolites are included in the dataset. Note also that intermediates, as depicted in the metabolism scheme, are not included in the dataset. Instead, the first non-intermediate metabolite in the pathway is used.

The SMILES for the metabolites were generated using ChemSpider (ChemSpider. http://www.chemspider.com/ [accessed Feb 13, 2019]). Consistency of stereochemistry information between parents and their metabolites was maintained.

Validation of Metabolite Structure Predictors

Predicted metabolites were compared to known metabolites from the reference and test datasets using their InChIs. The InChIs used for this comparison were generated without stereochemistry information using CDK (Willighagen et al., 2017; Chemistry Development Kit, version 2.0. https://cdk.github.io/ [accessed Nov 3, 2017]).

During the validation, a predicted aldehyde metabolite was considered equivalent to the corresponding carboxylic acid, because there is evidence that some percentage of an aldehyde metabolite acts as an intermediate that is further oxidized to a carboxylic acid without leaving the CYP enzyme active site (Bell-Parikh and Guengerich, 1999).

In the case of one parent molecule in the reference dataset, no predictions could be made because the parent molecule contains boron. FAME 2 is unable to make predictions for molecules containing boron because no boron-containing molecules were present in the dataset used to train the model.

The SyGMa predictions were carried out in Python using the SyGMa Python package (SyGMa, version 1.1.0), and RDKit (RDKit: Open-Source Cheminformatics, version 2017_03_01, 2017). Only the phase I reaction rule set was used and one reaction cycle was applied.

The BioTransformer predictions were performed using the CYP450 mode of the BioTransformer (BioTransformer, version 1.0.8. https://bitbucket.org/djoumbou/biotransformerjar/src/master/ [accessed Feb 5, 2019]) command line tool. BioTransformer was run individually for each parent compound using single SMILES input.

The ROC curves were generated using the ROCR R package (Sing et al., 2005; ROCR, version 1.0-7, 2015). When false negative data points were added to the curve, these data points were assigned a score of 0 or a rank of 1,000, respectively, depending on whether the ROC curve represented scores or ranks.

CONCLUSIONS

We have developed GLORY, a new tool for predicting the structures of human metabolites formed by CYPs. GLORY incorporates two key ideas: a literature-based collection of CYP-mediated reaction rules and SoM prediction, which was used particularly auspiciously to develop a new scoring approach for the predicted metabolites.

For GLORY, we developed a new collection of 73 reaction rules, describing 61 reaction types, for CYP-mediated metabolism. In developing this collection, we prioritized the reproducibility of our rule set and therefore based the rules on the scientific literature rather than on any dataset. In addition to the rules themselves, each reaction type was designated as either common or uncommon, again based on the scientific literature rather than on any dataset.

In addition, we have devised a priority score for predicted metabolites based on predicted SoM probabilities and the simple, literature-based distinction between common and uncommon reaction types. Hence neither our rule set nor our scoring approach is directly based on any dataset of metabolic reactions, setting our approach apart from other tools, for example SyGMa, which uses reaction rules and occurrence ratios derived from a proprietary dataset, and BioTransformer, whose rules were to some extent based on a freely available dataset.

GLORY has two modes: MaxEfficiency, which uses SoM prediction as a prefilter for the positions in a molecule at which reactions are allowed to occur, and MaxCoverage, which does not use a prefilter and instead focuses on high recall and an accurate ranking of the predicted metabolites. Using SoM prediction as a preliminary filter, i.e., in MaxEfficiency mode, does not work as well as might be expected in terms of reducing the number of putative false positive predictions while still keeping a high rate of recovery of reported metabolites. However, by developing a priority score for the predicted metabolites using SoM prediction combined with a simple binary distinction between common and uncommon reaction types, we are able to rank the metabolites predicted by MaxCoverage mode to the extent that GLORY can predict at least one known metabolite within the top three ranked positions for 76% of the molecules in the independent test set while achieving a recall of 0.83. GLORY's MaxCoverage mode outperforms both SyGMa and BioTransformer in terms of recall and outperforms SyGMa in terms of ranking (BioTransformer does not currently rank its metabolite predictions). One use case for the MaxCoverage mode could be, for example, identifying metabolites from mass spectrometry data.

Along with the collection of reaction rules, we provide a new, manually curated test dataset for free use as a benchmark dataset. In addition, GLORY is freely available as a web server at https://acm.zbh.uni-hamburg.de/glory/.

Importantly, the concept of GLORY is such that it can be extended to predict metabolites formed by enzymes not belonging to the CYP family. The enzymes that this approach can be expanded to is limited, in principle, only by the extent of the available data and the coverage of the relevant metabolic reactions by SoM prediction tools.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://bitbucket.org/djoumbou/biotransformerjar/ and https://www.drugbank.ca/.

AUTHOR'S NOTE

The GLORY web service is available at the following address: https://acm.zbh.uni-hamburg.de/glory/.

AUTHOR CONTRIBUTIONS

CdBK and JK: conceptualization; CdBK, CS, and JK: methodology; CdBK, CS, MŠ, NK, and NJ: software development; CdBK: validation; JK and DS: resources; CdBK: data curation; all authors: writing—original draft preparation; CdBK: visualization; DS, NJ, and JK: supervision; JK: project administration; DS and JK: funding acquisition.

FUNDING

CS and JK are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–project number KI 2085/1-1. DFG is a German research funding organization supporting research in science, engineering and the humanities. JK is also supported by the Bergen Research Foundation (BFS) [BFS2017TMT01]. BFS gives grants toward research and research supporting activities at the University of Bergen (UiB) and Haukeland University Hospital (HUS), and other Norwegian research institutions that cooperate with institutions in Bergen. The foundation also gives grants to support research at UiB and HUS at the interface between basic research and clinical research. MŠ and DS are supported

by the Ministry of Education, Youth and Sports (MEYS)—project numbers MSMT No 20-SVV/2017, LM2015063 and RVO: 68378050-KAV-NPUI. MEYS is responsible for public administration in education, for developing educational, youth and sport policies and international cooperation in these fields. It is also the central administrative office responsible for research and development and one of the main public research coordinating and funding bodies of the Czech Republic.

ACKNOWLEDGMENTS

We thank Craig Knox, Michael Wilson, and David Wishart from the University of Alberta, AB, Canada and OMx Personal Health Analytics Inc., AB, Canada for sending us the structures of the metabolites in the DrugBank database prior to their being made freely available online.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00402/full#supplementary-material

Supplementary Tables | Reaction rules and additional evaluation results (DOCX).

Supplementary Data Sheet 1 | Test dataset containing SMILES for parent molecules and metabolites as well as publication references (CSV).

Supplementary Data Sheet 2 | Reference dataset containing SMILES (with stereochemistry information), InChI (without stereochemistry information), and DrugBank and MetXBioDB identifiers for parent molecules and metabolites (JSON).

REFERENCES

Bell-Parikh, L. C., and Guengerich, F. P. (1999). Kinetics of cytochrome P450 2E1-catalyzed oxidation of ethanol to acetic acid via acetaldehyde. *J. Biol. Chem.* 274, 23833–23840. doi: 10.1074/jbc.274.34.23833

Cruciani, G., Carosati, E., De Boeck, B., Ethirajulu, K., Mackie, C., Howe, T., et al. (2005). MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem.* 48, 6970–6979. doi: 10.1021/jm050529c

Darvas, F. (1987). "Metabolexpert: an expert system for predicting metabolism of substances," in *QSAR in Environmental Toxicology - II*, ed K. L. E. Kaiser, (Dordrecht: Springer), 71–81.

Di, L. (2014). The role of drug metabolizing enzymes in clearance. *Expert Opin Drug Metabol. Toxicol.* 10, 379–393. doi: 10.1517/17425255.2014.876006

Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la-Fuente, A., Greiner, R., Manach, C., and Wishart, D. S. (2019). BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminform.* 11:2. doi: 10.1186/s13321-018-0324-5

Guengerich, F. P. (2001). Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. *Chem. Res. Toxicol.* 14, 611–650. doi: 10.1021/tx0002583

Judson, P. N. (2014). "Knowledge-based approaches for predicting the sites and products of metabolism," in *Drug Metabolism Prediction*, ed J. Kirchmair, Methods and Principles in Medicinal Chemistry (Weinheim: Wiley-VCH), 293–318.

Kirchmair, J., Göller, A. H., Lang, D., Kunze, J., Testa, B., and Wilson, I. D. (2015). Predicting drug metabolism: experiment and/or computation? *Nat. Rev. Drug Disc.* 14, 387–404. doi: 10.1038/nrd4581

Kirchmair, J., Williamson, M. J., Tyzack, J. D., Tan, L., Bond, P. J., Bender, A., et al. (2012). Computational prediction of metabolism: sites, products, S. A. R., P450 enzyme dynamics, and mechanisms. *J . Chem. Inform. Model.* 52, 617–48. doi: 10.1021/ci200542m

Klopman, G., Dimayuga, M., and Talafous, J. (1994). META. 1. a program for the evaluation of metabolic transformation of chemicals. *J. Chem. Inform. Comp. Sci.* 34, 1320–1325.

Kochev, N., Avramova, S., and Jeliazkova, N. (2018). Ambit-SMIRKS: a software module for reaction representation, reaction search and structure transformation. *J. Cheminform.* 10:42. doi: 10.1186/s13321-018-0295-6

Marchant, C. A., Briggs, K. A., and Long, A. (2008). *In silico* tools for sharing data and knowledge on toxicity and metabolism: Derek for Windows, Meteor, and Vitic. *Toxicol. Mechanisms Methods.* 18, 177–187. doi: 10.1080/15376510701857320

Mekenyan, O. G., Dimitrov, S. D., Pavlov, T. S., and Veith, G. D. (2004). A systematic approach to simulating metabolism in computational toxicology. I. the TIMES heuristic modelling framework. *Curr. Pharmaceutical Design.* 10, 1273–1293. doi: 10.2174/1381612043452596

Olsen, L., Montefiori, M., Tran, K. P., and Jørgensen, F. S. (2019). SMARTCyp 3.0: Enhanced cytochrome P450 site-of-metabolism prediction server. *Bioinformatics.* doi: 10.1093/bioinformatics/btz037. [Epub ahead of print].

Patlewicz, G., Jeliazkova, N., Safford, R., J., Worth, A., P., and Aleksiev, B. (2008). An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *SAR QSAR Environ. Res.* 19, 495–524. doi: 10.1080/10629360802083871

Ridder, L., and Wagener, M. (2008). SyGMa: combining expert knowledge and empirical scoring in the prediction of metabolites. *Chem. Med. Chem.* 3, 821–832. doi: 10.1002/cmdc.200700312

Rudik, A., Dmitriev, A., Lagunin, A., Filimonov, D., and Poroikov, V. (2015). SOMP: web server for *in silico* prediction of sites of metabolism for drug-like compounds. *Bioinformatics* 31, 2046–2048. doi: 10.1093/bioinformatics/btv087

Rudik, A. V., Bezhentsev, V. M., Dmitriev, A. V., Druzhilovskiy, D. S., Lagunin, A. A., Filimonov, D. A., et al. (2017). MetaTox: web application for predicting structure and toxicity of xenobiotics' metabolites. *J. Chem. Inform. Model.* 57, 638–642. doi: 10.1021/acs.jcim.6b00662

Rydberg, P., Gloriam, D. E., Zaretzki, J., Breneman, C., and Olsen, L. (2010). SMARTCyp: a 2D method for prediction of cytochrome p450-mediated drug metabolism. *ACS Med. Chem. Lett.* 1, 96–100. doi: 10.1021/ml100016x

Šícho, M., de Bruyn Kops, C., Stork, C., Svozil, D., and Kirchmair, J. (2017). FAME 2: simple and effective machine learning model of cytochrome P450 regioselectivity. *J. Chem. Inform. Model.* 57, 1832–1846. doi: 10.1021/acs.jcim.7b00250

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* 21, 3940–3941. doi: 10.1093/bioinformatics/bti623

Testa, B., Pedretti, A., and Vistoli, G. (2012). Reactions and enzymes in the metabolism of drugs and other xenobiotics. *Drug Disc. Today* 17, 549–560. doi: 10.1016/j.drudis.2012.01.017

Tian, S., Djoumbou-Feunang, Y., Greiner, R., and Wishart, D. S. (2018). CypReact: a software tool for *in silico* reactant prediction for human cytochrome P450 enzymes. *J. Chem. Inform. Model.* 58, 1282–1291. doi: 10.1021/acs.jcim.8b00035

Tyzack, J. D., Hunt, P. A., and Segall, M. D. (2016). Predicting regioselectivity and lability of cytochrome p450 metabolism using quantum mechanical simulations. *J. Chem. Inform. Model.* 56, 2180–2193. doi: 10.1021/acs.jcim.6b00233

Tyzack, J. D., and Kirchmair, J. (2018). Computational methods and tools to predict cytochrome P450 metabolism for drug discovery. *Chem. Biol. Drug Design* 93, 377–386. doi: 10.1111/cbdd.13445

Willighagen, E. L., Mayfield, J. W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., et al. (2017). The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminform.* 9:33. doi: 10.1186/s13321-017-0220-4

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Zaretzki, J., Matlock, M., and Swamidass, S. J. (2013). XenoSite: accurately predicting CYP-mediated sites of metabolism with neural networks. *J . Chem. Inform. Model.* 53, 3373–3383. doi: 10.1021/ci400518g

Conflict of Interest Statement: NJ is a founder and co-owner of Ideaconsult Ltd. and has been the technical manager of the company since 2009. NK works for Ideaconsult Ltd. on a part-time basis.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 de Bruyn Kops, Stork, Šícho, Kochev, Svozil, Jeliazkova and Kirchmair. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 11 July 2019
doi: 10.3389/fchem.2019.00496

Docking, Interaction Fingerprint, and Three-Dimensional Quantitative Structure–Activity Relationship (3D-QSAR) of Sigma1 Receptor Ligands, Analogs of the Neuroprotective Agent RC-33

José Luis Velázquez-Libera[1], *Giacomo Rossino*[2], *Carlos Navarro-Retamal*[1], *Simona Collina*[2] *and Julio Caballero*[1]*

[1] Centro de Bioinformática y Simulación Molecular, Facultad de Ingeniería, Universidad de Talca, Talca, Chile,
[2] Pharmaceutical and Medicinal Chemistry Section, Drug Sciences Department, Università di Pavia, Pavia, Italy

OPEN ACCESS

Edited by:
Teodorico Castro Ramalho, Universidade Federal de Lavras, Brazil

Reviewed by:
Andrei I. Khlebnikov, Tomsk Polytechnic University, Russia
Marco Tutone, University of Palermo, Italy

***Correspondence:**
Julio Caballero
jcaballero@utalca.cl

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: 12 April 2019
Accepted: 27 June 2019
Published: 11 July 2019

Citation:
Velázquez-Libera JL, Rossino G, Navarro-Retamal C, Collina S and Caballero J (2019) Docking, Interaction Fingerprint, and Three-Dimensional Quantitative Structure–Activity Relationship (3D-QSAR) of Sigma1 Receptor Ligands, Analogs of the Neuroprotective Agent RC-33. Front. Chem. 7:496. doi: 10.3389/fchem.2019.00496

The human Sigma1 receptor (S1R), which has been identified as a target with an important role in neuropsychological disorders, was first crystallized 3 years ago. Since S1R structure has no relation with another previous crystallized structures, the presence of the new crystal is an important hallmark for the design of agonists and antagonists against this important target. Some years ago, our group identified RC-33, a potent and selective S1R agonist, endowed with neuroprotective properties. In this work, drawing on new structural information, we studied the interactions of RC-33 and its analogs with the S1R binding site by using computational methods such as docking, interaction fingerprints, and receptor-guided alignment three dimensional quantitative structure–activity relationship (3D-QSAR). We found that RC-33 and its analogs adopted similar orientations within S1R binding site, with high similitude with orientations of the crystallized ligands; such information was used for identifying the residues involved in chemical interactions with ligands. Furthermore, the structure-activity relationship of the studied ligands was adequately described considering classical QSAR tests. All relevant aspects of the interactions between the studied compounds and S1R were covered here, through descriptions of orientations, binding interactions, and features that influence differential affinities. In this sense, the present results could be useful in the future design of novel S1R modulators.

Keywords: sigma1 receptor ligands, RC-33, arylalkylamine derivates, docking, quantitative structure–activity relationships, interaction fingerprints

INTRODUCTION

The Sigma receptors (SR) have attracted the interest of the scientific community thoroughly in the last decades owing to their potential role in cell survival and function modulation (Walker et al., 1990; Chu and Ruoho, 2016). They were originally misclassified as a subtype of opioid receptors (Martin et al., 1976), but they were later classified as unique class of intracellular proteins,

distinct from other receptors such as GPCRs (G protein-coupled receptors). Sigma receptors (SRs), comprise two subtypes σ_1 and σ_2 receptors (S1R and S2R, respectively) associated with aging- and mitochondria-associated disorders (Tesei et al., 2018). Both subtypes are highly expressed in the central nervous system, but they are derived from completely different genes. S1R was cloned in 1996 (Hanner et al., 1996) and was crystallized for the first time 3 years ago, in 2016 (Schmidt et al., 2016), whereas S2R was cloned only very recently, in 2017, by Alon et al. (2017).

S1R is an intracellular modulator between the endoplasmic reticulum and the mitochondria, the cell nuclei, the membrane, and it also modulates intracellular signaling. It plays a key role in neuropsychological disorders such as depression, enhances the glutamatergic neurotransmission (DeCoster et al., 1995; Meyer et al., 2002), and modulates second messenger systems, such as the phospholipase C/protein kinase C/inositol 1,4,5-trisphosphate system (Morin-Surun et al., 1999). Multiple biological roles of S1R have been identified, which made this protein a relevant target for the future treatment of epilepsy, schizophrenia, sclerosis, Alzheimer, and Parkinson's diseases, cancer, etc. (Mishina et al., 2005; Hashimoto, 2009; Furuse and Hashimoto, 2010; Mavlyutov et al., 2015; Vavers et al., 2017; Tesei et al., 2018). Moreover, S1R agonists enhanced neuroplasticity, and may be effective in amyotrophic lateral sclerosis (Peviani et al., 2014) and multiple sclerosis (Collina et al., 2017b).

Not less important, preclinical studies carried out on different models of memory impairment have revealed that S1R ligands could be promising drugs to treat cognitive dysfunctions (Hayashi and Su, 2004; Monnet and Maurice, 2006; Yagasaki et al., 2006; Collina et al., 2017a). Therefore, the identification of potent and selective S1R modulators is of great interest to develop novel therapeutic strategies focused mainly in the treatment of central nervous system disorders. The list of S1R ligands in the last years includes thioxanthene-derived compounds (Glennon et al., 2004), fenpropimorph-derived analogs (Hajipour et al., 2010), 2(3*H*)-benzothiazolones (Yous et al., 2005), cyclopropylmethylamines (Prezzavento et al., 2007), benzo[d]oxazol-2(3H)-one derivatives (Zampieri et al., 2009), etc. All these compounds were developed when the three-dimensional (3D) structure of S1R was unknown. Despite this, the pharmacophoric features of S1R were identified and these compounds comply with the general accepted pharmacophoric pattern. It was demonstrated that at least one N positively charged atom is important for binding at sigma receptors and this atom must be flanked by two hydrophobic regions of different sizes (Ablordeppey et al., 2000; Glennon, 2005; Caballero et al., 2012).

In the last years, we designed and synthesized compounds that comply with the proposed pharmacophore model and evaluated them as S1R ligands (Collina et al., 2007; Urbano et al., 2007; Rossi et al., 2010, 2011), leading to the finding of compound RC-33 as a potent and selective S1R agonist (Rossi et al., 2013a; Marra et al., 2016). The structure-activity relationship (SAR) of the majority of these compounds was previously described by us by using 2D-QSAR methodologies (Quesada-Romero et al., 2015). With the recent report of the S1R 3D structure (Schmidt et al., 2016), structure-based molecular modeling methods could be used to investigate S1R ligands with a new glance. With this in mind, we propose in this work the analysis of the SAR of RC-33 and its analogs (in total there were 80 compounds) by combining docking and a 3D-QSAR methodology. This is the first study focused on describing the SAR of S1R ligands by using structure-based molecular modeling methods, after the report of the crystallographic structure of this important biological target.

MATERIALS AND METHODS

Dataset Preparation

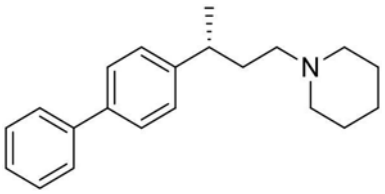
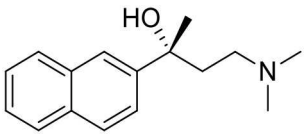
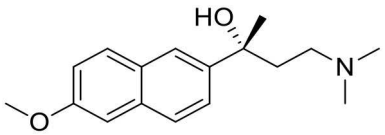
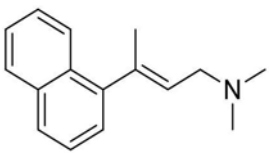
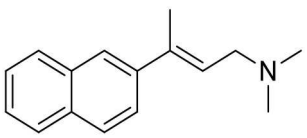
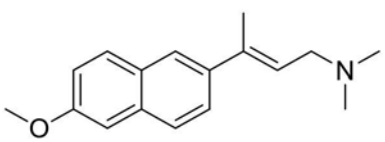
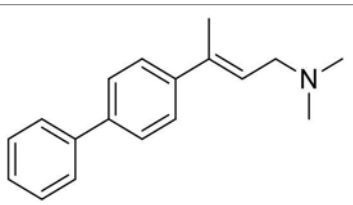
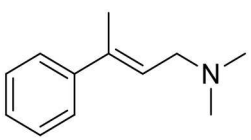
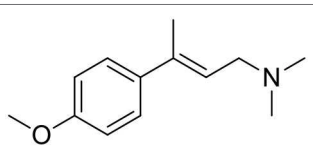
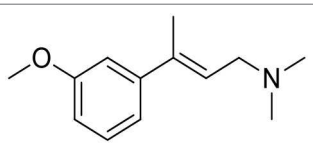
The studied compounds were extracted from references (Collina et al., 2007; Urbano et al., 2007; Rossi et al., 2010, 2011, 2015, 2017; Rui et al., 2016). This dataset yielded a total of 80 compounds with reported activities as Ki ranging from 0.00069 to 1 μM. Ki values were converted into logarithmic pKi values prior 3D-QSAR models' elaboration. The compound chemical structures and their pKi values are depicted in **Table 1**. The molecular structures were sketched using Maestro's molecular editor (Maestro 10.2.011, Schrödinger LLC). Thereafter, the 3D structures were obtained with the help of the LigPrep module (LigPrep, Maestro 10.2.011, Schrödinger LLC); ionization states were generated at pH 7.0 ± 2.0 using Epik (Shelley et al., 2007). For compounds containing two possible enantiomers which are reported in racemic form, the *R* enantiomer was chosen for QSAR experiments because it was determined that both RC-33 enantiomers showed similar affinities for the S1R and they are almost equally effective as S1R agonists (Rossi et al., 2013b). However, both enantiomers were chosen for docking experiments to explore the interactions in the S1R binding site.

Molecular Docking

The ligand-receptor molecular docking experiments of RC-33 analogs into the active site of S1R were performed by using the software Glide from the Schrödinger suite (Friesner et al., 2004). Glide is one of the most effective docking programs at this moment with many successful applications relating to rational design of novel drugs and investigation of protein-ligand interactions. Such applications encompass *in silico* search of novel drugs (Osguthorpe et al., 2012; Amaning et al., 2013), analysis of the SAR of congeneric series of compounds (Almerico et al., 2012; Quesada-Romero and Caballero, 2014; Quesada-Romero et al., 2014; Mena-Ulecia et al., 2015), evaluation of enzymatic reaction pathways (Wu et al., 2011; Batra et al., 2013), etc.

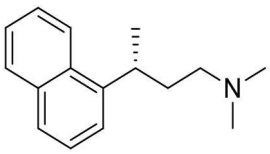
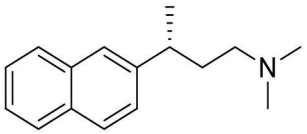
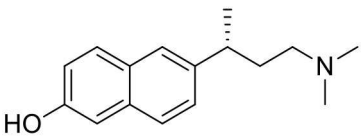
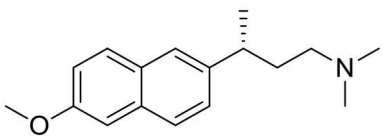
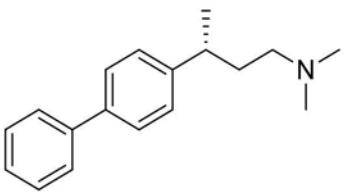
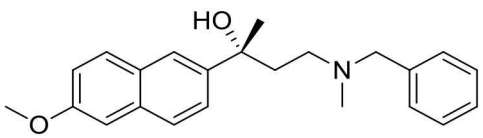
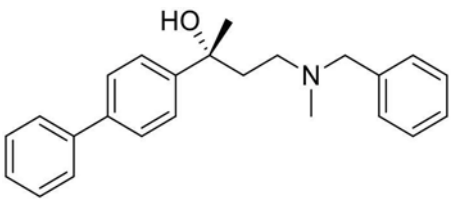
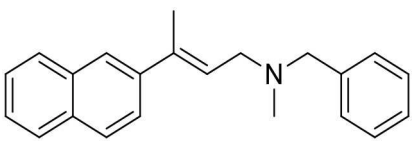
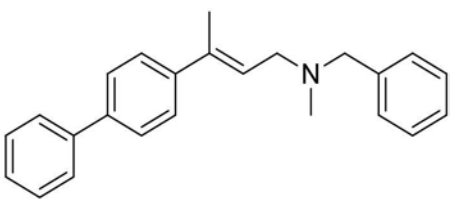
Protein coordinates were extracted from the crystal structure of S1R bound to the selective antagonist PD144418 (code 5HK1 in Protein Data Bank) (Schmidt et al., 2016). A grid box of 20 × 20 × 20Å was centered on the center of mass of the ligand in this crystal structure covering the binding site of S1R. Glide standard (SP) and extra-precision (XP) modes were employed with the same protocol and parameters that were used by us in previous works (Quesada-Romero and Caballero, 2014; Quesada-Romero et al., 2014; Mena-Ulecia et al., 2015). Glide SP was used to evaluate the capability of the Glide method to obtain poses that fit the known pharmacophore of S1R ligands,

TABLE 1 | Structures of RC-33 analogs as S1R ligands.

ID	Structure	Experimental pKi[b]	Predicted pKi[b]	References
1 (RC-33)		9.16	9.37	Rossi et al., 2011
2		7.60	7.54	Collina et al., 2007
3		7.41	7.49	Rossi et al., 2010
4		6.99	6.96	Collina et al., 2007
5		7.70	7.62	Collina et al., 2007
6		7.67	7.44	Collina et al., 2007
7		8.85	8.73	Rossi et al., 2011
8[a]		6.00	6.79	Rossi et al., 2011
9		6.09	6.23	Rossi et al., 2011
10		6.00	6.14	Rossi et al., 2011

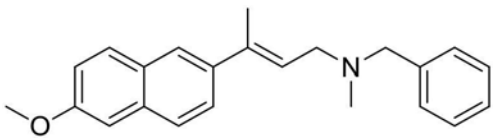
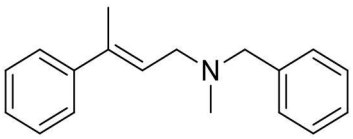
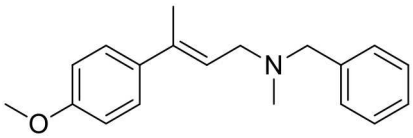
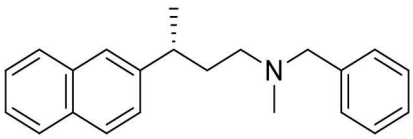
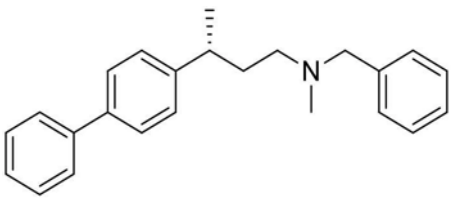
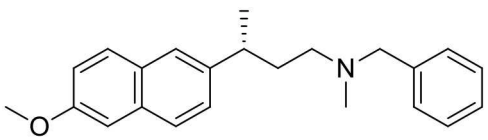
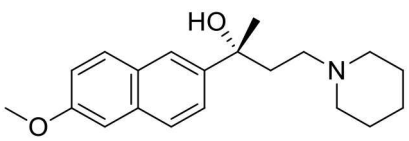
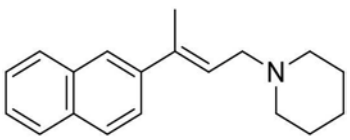
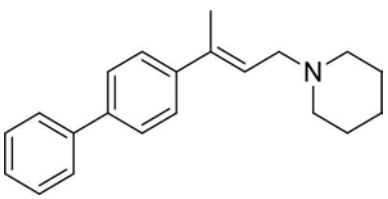
(Continued)

TABLE 1 | Continued

ID	Structure	Experimental pKi[b]	Predicted pKi[b]	References
11		7.33	7.11	Collina et al., 2007
12[a]		8.71	7.99	Collina et al., 2007
13		7.72	7.65	Collina et al., 2007
14		8.64	8.41	Collina et al., 2007
15		8.99	9.05	Collina et al., 2007
16		8.22	8.10	Rossi et al., 2010
17		8.62	8.26	Rossi et al., 2010
18		8.10	8.09	Collina et al., 2007
19		8.20	8.29	Collina et al., 2007

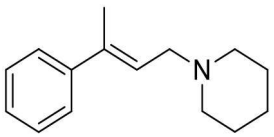
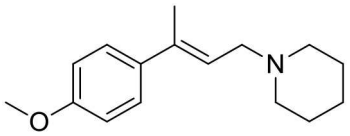
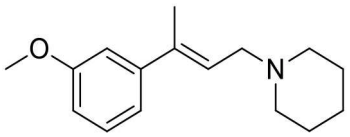
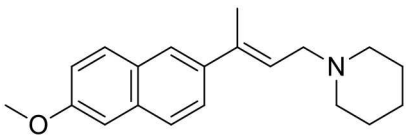
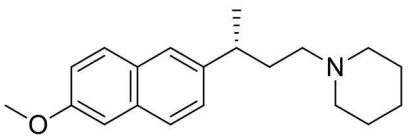
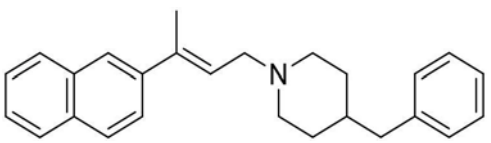
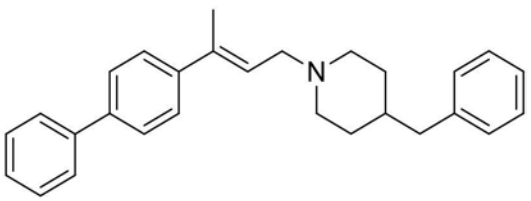
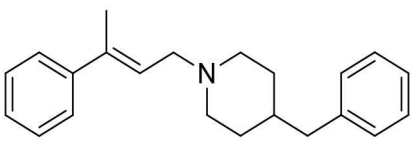
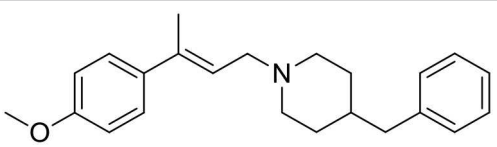
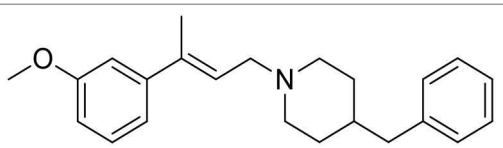
(Continued)

TABLE 1 | Continued

ID	Structure	Experimental pKi[b]	Predicted pKi[b]	References
20		7.86	8.22	Rossi et al., 2010
21		7.04	7.06	Collina et al., 2007
22		8.28	8.11	Collina et al., 2007
23[a]		8.27	8.96	Collina et al., 2007
24		8.24	8.30	Collina et al., 2007
25		8.64	8.52	Rossi et al., 2010
26		7.98	8.01	Rossi et al., 2010
27		9.01	8.79	Rossi et al., 2011
28		9.07	9.22	Rossi et al., 2011

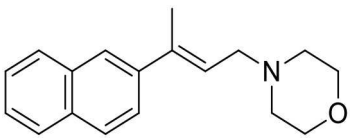
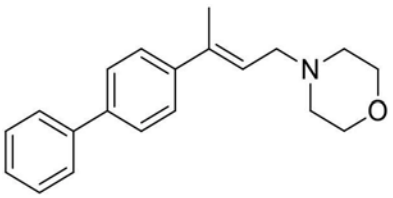
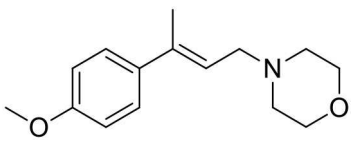
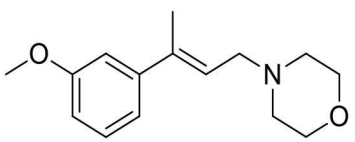
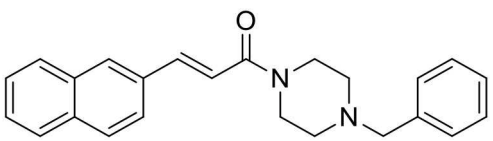
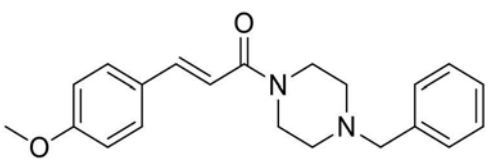
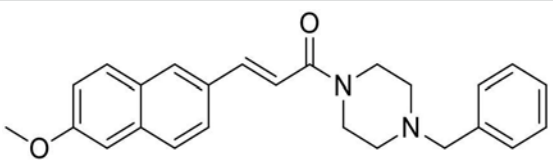
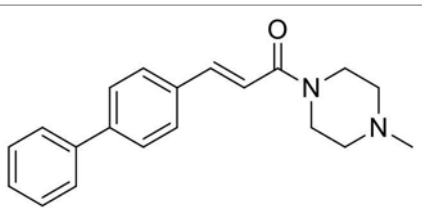
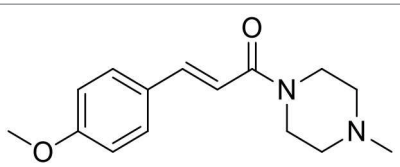
(Continued)

TABLE 1 | Continued

ID	Structure	Experimental pKi[b]	Predicted pKi[b]	References
29[a]		6.34	6.23	Rossi et al., 2011
30		7.85	7.67	Rossi et al., 2011
31		7.04	7.10	Rossi et al., 2011
32		8.36	8.62	Rossi et al., 2010
33		8.89	8.99	Rossi et al., 2010
34[a]		7.64	7.65	Rossi et al., 2011
35		8.15	8.20	Rossi et al., 2011
36[a]		8.97	8.56	Rossi et al., 2011
37[a]		8.38	8.28	Rossi et al., 2011
38[a]		8.11	7.40	Rossi et al., 2011

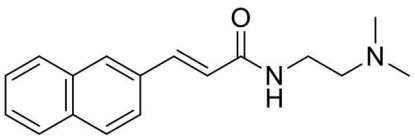
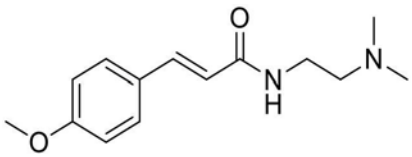
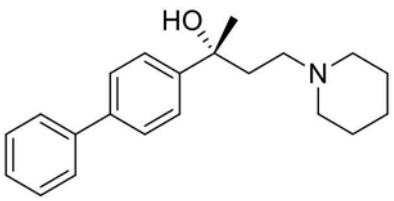
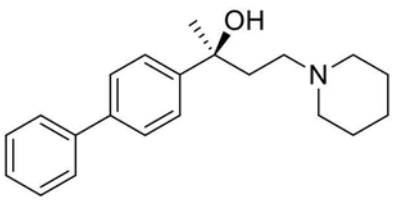
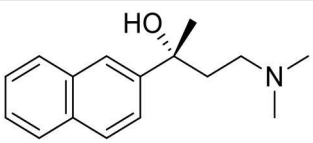
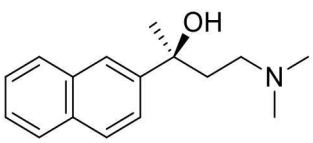
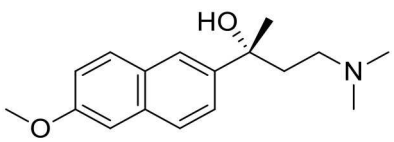
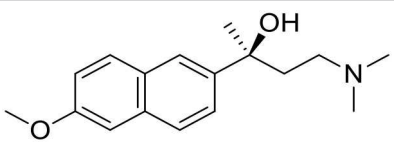
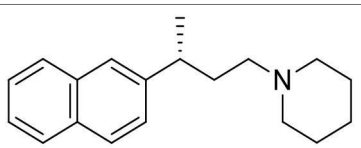
(Continued)

TABLE 1 | Continued

ID	Structure	Experimental pKi[b]	Predicted pKi[b]	References
39		8.02	8.04	Rossi et al., 2011
40		7.94	7.83	Rossi et al., 2011
41		6.00	6.38	Rossi et al., 2011
42		6.00	5.98	Rossi et al., 2011
43[a]		6.00	7.29	Urbano et al., 2007
44		6.00	6.03	Urbano et al., 2007
45		6.00	6.16	Urbano et al., 2007
46		6.00	5.78	Urbano et al., 2007
47[a]		6.00	6.36	Urbano et al., 2007

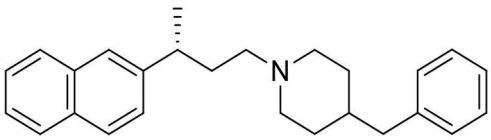
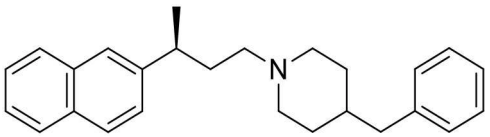
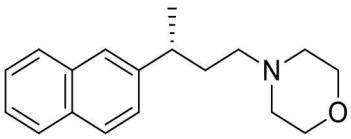
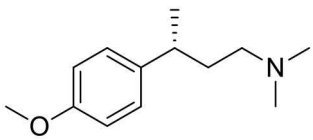
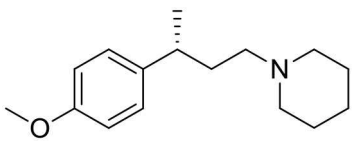
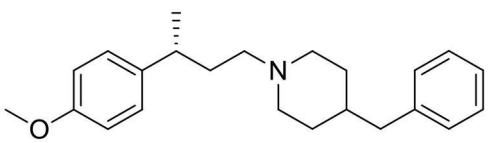
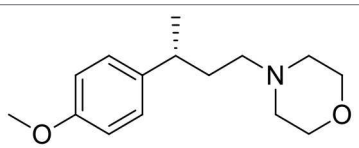
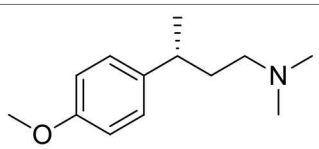
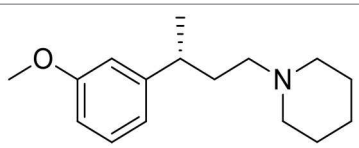
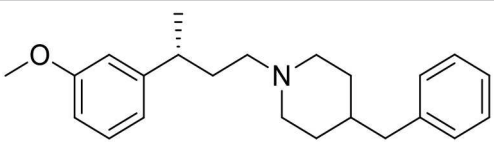
(Continued)

TABLE 1 | Continued

ID	Structure	Experimental pKi[b]	Predicted pKi[b]	References
48		6.00	5.89	Urbano et al., 2007
49		6.00	5.75	Urbano et al., 2007
50		7.41	7.52	Rossi et al., 2015
51		8.33	8.37	Rossi et al., 2015
52		6.69	6.81	Rossi et al., 2015
53		7.20	7.40	Rossi et al., 2015
54[a]		7.29	7.44	Rossi et al., 2015
55		7.60	7.90	Rossi et al., 2015
56		8.82	8.57	Rossi et al., 2017

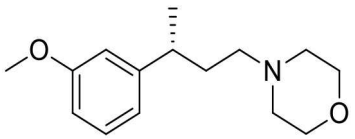
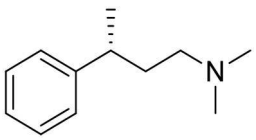
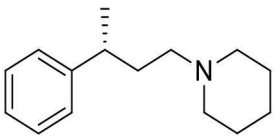
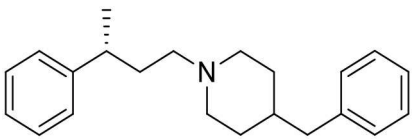
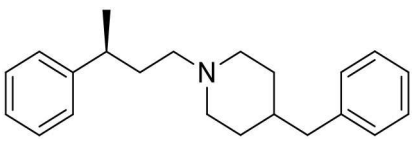
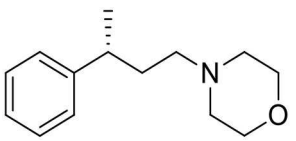
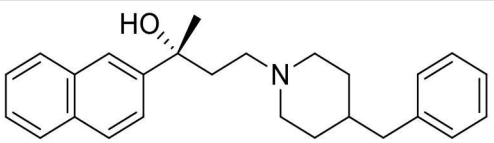
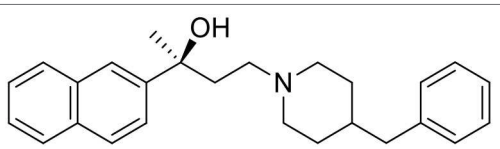
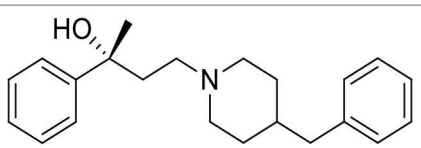
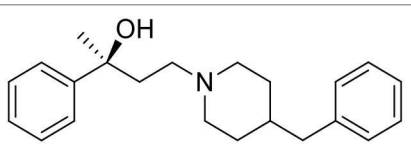
(Continued)

TABLE 1 | Continued

ID	Structure	Experimental pKi[b]	Predicted pKi[b]	References
57		8.22	8.29	Rossi et al., 2017
58		8.16	8.14	Rossi et al., 2017
59		8.27	8.20	Rossi et al., 2017
60		6.94	6.91	Rossi et al., 2017
61		7.70	7.68	Rossi et al., 2017
62		8.46	8.37	Rossi et al., 2017
63		7.12	7.20	Rossi et al., 2017
64		6.62	6.75	Rossi et al., 2017
65[a]		7.44	6.42	Rossi et al., 2017
66		8.54	8.56	Rossi et al., 2017

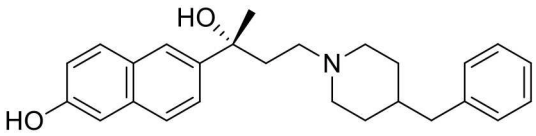
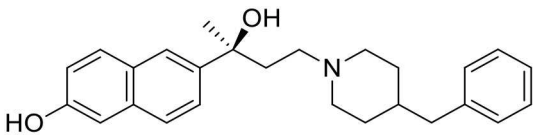
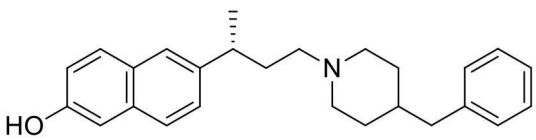
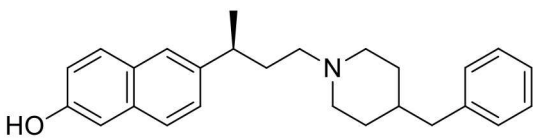
(Continued)

TABLE 1 | Continued

ID	Structure	Experimental pKi[b]	Predicted pKi[b]	References
67[a]		6.86	7.25	Rossi et al., 2017
68		6.37	6.35	Rossi et al., 2017
69		7.34	7.37	Rossi et al., 2017
70[a]		8.54	8.62	Rossi et al., 2017
71		8.52	8.45	Rossi et al., 2017
72[a]		7.07	6.56	Rossi et al., 2017
73		8.00	8.01	Rui et al., 2016
74		7.96	7.92	Rui et al., 2016
75		7.57	7.73	Rui et al., 2016
76		7.40	7.45	Rui et al., 2016

(Continued)

TABLE 1 | Continued

ID	Structure	Experimental pKi[b]	Predicted pKi[b]	References
77		7.15	7.09	Rui et al., 2016
78		7.21	7.18	Rui et al., 2016
79		7.46	7.42	Rui et al., 2016
80[a]		7.89	7.19	Rui et al., 2016

[a] *Test set compounds.*
[b] *Experimental and predicted pKi values using Model SE.*

and the more precise Glide XP was used for finding the final docking poses.

After several poses were found for each compound, the ones that showed the best scoring energies were considered. The information of PD144418, 4-IBP, haloperidol, NE-100, and (+)-pentazocine in the crystallographic structures recently reported (Schmidt et al., 2016, 2018) was considered for the selection of the best solutions; these compounds show how the previously reported pharmacophoric pattern (Glennon, 2005) is oriented inside the S1R binding site. The essential chemical interactions described for analog ligands (ECIDALs) (Muñoz-Gutierrez et al., 2016; Ramírez and Caballero, 2018) defined for S1R ligands were identified using this information. The most obvious essential chemical interaction is that charged amino group of the ligands must be close to the side chain carboxylate group of the residue Glu172, forming an electrostatic interaction. Therefore, the best docking solution for each compound was the pose that had the best scoring energy and complies with this essential chemical interaction.

The "Interaction Fingerprints Panel" of Maestro (Maestro 10.2.011, Schrödinger LLC) was used for deriving the Interaction fingerprints (IFPs) as described in Singh et al. reports (Deng et al., 2004; Singh et al., 2006). The method accounts for the presence of different types of chemical interactions between ligands and the binding site residues of the target receptor by using bits. For this purpose, distance cutoffs are defined for the binding site, and the interacting set encompasses the residues that contain atoms within the specified cutoff distance from ligand atoms. An interaction matrix is constructed including the bits with relevant information of the defined chemical interactions.

QSAR Modeling

After docking experiments, 3D-QSAR models were performed to explain the SAR of the RC-33 analogs. Their bioactive conformations predicted by using docking were used as the alignment rule for deriving the models. The structural features that affect their activities against the S1R were identified by describing steric and electrostatic fields.

The 80 compounds dataset was randomly partitioned into training (64 compounds) and external (16 compounds) sets. A homogenous distribution of the activities was granted in both training and test sets. 3D-QSAR models were generated using Open3DQSAR (Tosco and Balle, 2011), an open access tool with all the capacities to construct 3D-QSAR models. Steric and electrostatic fields were computed according to classical molecular mechanics equations using the Merck Molecular Force Field (Halgren, 1996).

The field variables were calculated by describing the interaction energies between probe atoms (sp^3 carbon atoms with a charge +1) and structures in a 1.0 Å step size grid box surrounding the whole set. Variables were processed as follows: (i) high energies adopted the top value of 30 kcal/mol, (ii) energy values very close to zero (below 0.05 kcal/mol) were set to zero in order to reduce noise, (iii) variables which only assumed a few different values (*n*-level variables) were removed. Thereafter, variables were scaled using the Block Unscaled Weighting procedure (Kastenholz et al., 2000; Boháč et al., 2002) and the predictive power of the models was improved by using the Smart Region Definition algorithm (Pastor et al., 1997).

Partial Least Square (PLS) regression was used to construct 3D-QSAR models, including from one to five Principal Components (PCs) and different combinations of fields. Models

were derived by using one field and by combining them; the best model was selected by considering the higher value of the internal leave-one-out (LOO) cross-validation Q^2.

RESULTS AND DISCUSSION

Docking Predictions

We have a structural information of the binding poses of S1R ligands such as PD144418, 4-IBP, haloperidol, and NE-100 that similar in shape to RC-33. This information was used for evaluating the quality of the obtained docking results for RC-33 and its analogs. It is known that S1R ligands contain a charged nitrogen central atom flanked by two hydrophobic regions of different size (Glennon, 2005). The above mentioned S1R ligands form electrostatic interactions between the ligand charged nitrogen atoms and the side chain carboxylate of Glu172. In addition, their larger hydrophobic groups locate near the residues Val84, Met93, Leu95, Leu105, Tyr206, Ile178, Leu182, and Tyr103 (primary hydrophobic site), and their smaller hydrophobic groups locate near the residues Phe107, Trp164, His154, and Ile124 (secondary hydrophobic site). It is expected that the studied compounds establish such interactions.

Docking orientations of RC-33 and its analogs are represented in **Figure 1**. The best docking pose obtained for RC-33 was compared with the orientations of PD144418, 4-IBP, haloperidol, NE-100, and (+)-pentazocine in the reference crystallographic structures 5HK1, 5HK2, 6DJZ, 6DK0, and 6DK1, respectively. (+)-Pentazocine is an agonist as RC-33, but it is shorter than RC-33 and the other crystallized ligands; therefore, it is the least suitable ligand for the structural comparison between the crystallized ligands and the docked RC-33 analogs. **Figure 1A** shows that the docked structure of RC-33 was similarly oriented as the other crystallized ligands. On the other hand, **Figure 1B** shows that suitable binding modes of the ligands were found for all the RC-33 analogs. All of them form the conserved salt bridge between the charged N atom of the ligands and the residue Glu172 of the S1R. They also oriented their large hydrophobic groups to the primary hydrophobic site, and oriented their small hydrophobic groups to the secondary hydrophobic site. Representations in **Figure 1** show that our docking poses are similar to the S1R-ligand X-ray structures reported to date.

We calculated RMSD values for the studied compounds with respect to the docking result of RC-33 inside the S1R by using an *in-house* script (Velázquez-Libera et al., 2018). These calculations show the similarity in orientations between RC-33 and its analogs in an easy way. Since the RC-33 derivatives are different from the reference compound, RMSD values were calculated by considering only the common graphs between molecules. %RefMatch and %MolMatch values were defined, where %RefMatch refers to the percent of common graphs between the docked compound and RC-33 regarding the total number of atoms of RC-33; meanwhile, %MolMatch refers to the percent of common graphs between the docked compound, and RC-33 regarding the total number of atoms of the docked compound. These values allow identifying the maximal similitude between the docked compound and RC-33; therefore, an RMSD value with high %RefMatch and %MolMatch values reflects that the compound under analysis bears a strong resemblance with RC-33.

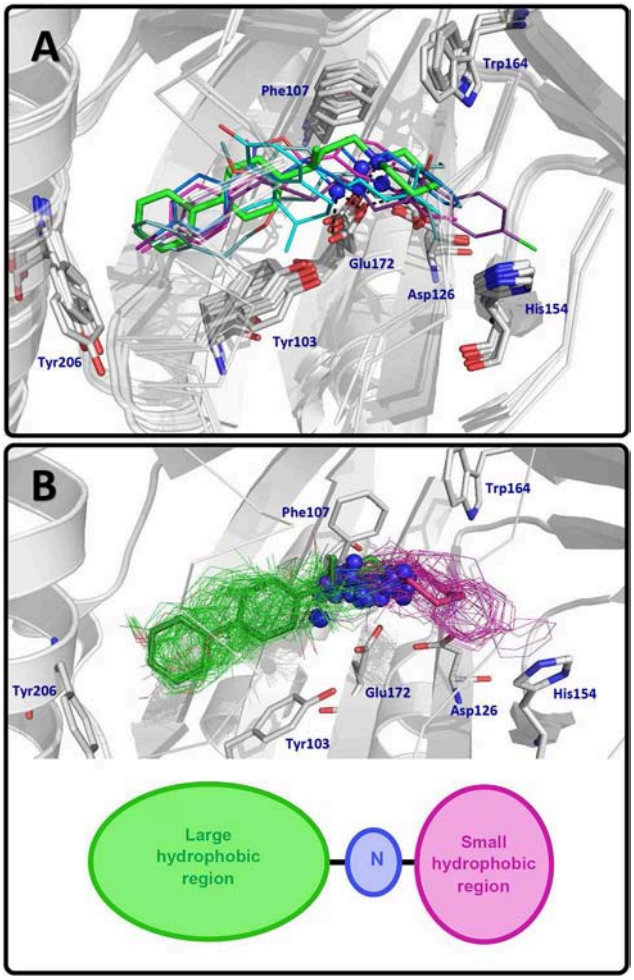


FIGURE 1 | Docking results for RC-33 and its analogs. **(A)** Docking pose obtained for RC-33 (stick representation in green) and comparison with X-ray crystallographic structures of the antagonist PD144418 (thin stick representation in purple, PDB code 5HK1), the ambiguous ligand 4-IBP (thin stick representation in light blue, PDB code 5HK2), the antagonist haloperidol (thin stick representation in lilac, PDB code 6DJZ), the antagonist NE-100 (thin stick representation in teal, PDB code 6DK0), and the agonist (+)-pentazocine (thin stick representation in cyan, PDB code 6DK1). N positively charged atom for each compound is represented by a blue sphere. **(B)** (top) Docking of RC-33 (in sticks representation) and comparison with its analogs (in lines representation); for each compound large hydrophobic group is in green at the left, small hydrophobic group is in purple at the right, and N positively charged atom is a sphere in blue. (bottom) Pharmacophoric model for S1R ligands: N positively charged atom (blue) flanked by large hydrophobic (green) and small hydrophobic (purple) regions.

The majority of the compounds under study have the 1-(3-phenylbutyl)piperidine or parts of this group in common with RC-33. Their RMSD values are reported in **Table 2**. In general, RMSD values reflect that the majority of compounds had the 1-(3-phenylbutyl)piperidine (or part of this group) similarly oriented with respect to RC-33 (RMSD $<$ 2 Å). However,

TABLE 2 | RMSD values of the obtained docking pose common fragments for the studied compounds with respect to the docking result of RC-33 inside the S1R.

ID	RMSD (Å)[a]	%RefMatch[b]	%MolMatch[c]	RMSD (Å)[a,d]	%RefMatch[b,d]	%MolMatch[c,d]
1 (RC-33)				0.86	100	100
2	1.70	50	61	2.30	50	61
3	1.40	50	55	1.82	50	55
4	2.61	36	47			
5	2.20	36	47			
6	1.42	36	42			
7	0.46	64	74			
8	2.94	36	62			
9	1.67	36	53			
10	2.50	36	53			
11	2.80	50	65	2.89	50	65
12	2.05	50	65	2.16	50	65
13	2.21	50	61	1.50	50	61
14	1.43	50	58	1.46	50	58
15	0.43	77	89	0.86	77	89
16	1.45	50	42	1.07	50	42
17	0.97	77	65	0.69	77	65
18	1.58	36	35			
19	0.42	64	56			
20	1.29	36	32			
21	2.85	36	42			
22	0.38	36	38			
23	2.09	50	48	1.94	50	48
24	0.94	77	68	0.83	77	68
25	1.22	50	44	0.82	50	44
26	1.23	73	70	1.98	73	70
27	1.38	36	40			
28	0.59	64	64			
29	2.24	36	50			
30	0.85	36	44			
31	1.15	36	44			
32	1.11	36	36			
33	2.15	73	73	1.78	73	73
34	2.14	36	30			
35	0.45	64	48			
36	3.70	36	35			
37	4.17	36	32			
38	2.41	36	32			
39	1.16	36	40			
40	0.63	64	64			
41	0.77	36	44			
42	0.72	36	44			
43	2.10	32	26			
44	1.10	32	28			
45	1.66	32	24			
46	0.67	59	57			
47	1.05	32	37			
48	2.28	32	35			
49	2.23	32	39			
50	0.96	100	96			
51	0.64	100	96			

(Continued)

TABLE 2 | Continued

ID	RMSD (Å)[a]	%RefMatch[b]	%MolMatch[c]	RMSD (Å)[a,d]	%RefMatch[b,d]	%MolMatch[c,d]
52	1.50	50	61			
53	1.87	50	61			
54	1.48	50	55			
55	1.07	50	55			
56	2.19	73	80	1.96	73	80
57	2.63	73	59			
58	2.22	73	59			
59	1.95[e]	91	100	1.98[e]	91	100
60	1.80	50	73	2.54	50	73
61	2.63	73	89	1.52	73	89
62	3.02	73	64	6.97	73	64
63	2.38[e]	82	100	2.51[e]	82	100
64	2.11	50	73	2.58	50	73
65	2.45	73	89	1.01	73	89
66	2.68	73	64	4.26	73	64
67	2.19[e]	82	100	1.98[e]	82	100
68	0.63	50	85	2.96	50	85
69	2.65	73	100	2.73	73	100
70	3.87	73	70			
71	6.66	73	70			
72	2.58[e]	73	100	1.49[e]	73	100
73	2.29	73	57			
74	2.20	73	57			
75	4.65	73	67			
76	4.39	73	67			
77	2.72	73	55			
78	2.34	73	55			
79	1.74	73	57			
80	2.08	73	57			

[a] *RMSD values considering only the common chemical fragments between the docked compound and the reference compound RC-33.*
[b] *%RefMatch refers to the percent of common graphs between the docked and reference compound RC-33 concerning the total number of atoms of the reference compound RC-33.*
[c] *%MolMatch refers to the percent of common graphs between the docked and reference compound RC-33 regarding the total number of atoms of the docked compound.*
[d] *RMSD, %RefMatch, and %MolMatch values for the S enantiomer of the compounds reported as racemic pairs.*
[e] *In this case, difference in ring heavy atoms were not considered between the docked compound and the reference compound RC-33.*

Table 2 reports compounds with RMSD > 2.5 Å (for instance, compounds **11** (R and S), **57**, **60** (S), and **77**). The 1-(3-phenylbutyl)piperidine group of these compounds is displaced toward the helices α4 and α5; however, their amine groups keep the salt bridge interaction with the residue Glu172. In addition, we found in **Table 2** compounds with RMSD > 4 Å (for instance, compounds **37**, **62** (S), **66** (S), **75**, and **76**). The 1-(3-phenylbutyl)piperidine group of these compounds is oriented to the reverse with respect to this group in RC-33; their amine groups also keep the salt bridge interaction with the residue Glu172. These compounds have larger hydrophobic substituents at position 4 of the piperidine, increasing the size of this group. The changed groups fit better inside the bigger hydrophobic cavity close to the helices α4 and α5 when their orientations are opposed to the orientation of the 1-(3-phenylbutyl)piperidine group in RC-33. In this way, these compounds are also adapted to the previous described pharmacophore pattern for S1R ligands (Ablordeppey et al., 2000; Glennon, 2005; Caballero et al., 2012) (the N positively charged atom flanked by two hydrophobic groups of different sizes), where the charged atom is salt-bridged to Glu172, the bigger hydrophobic group is placed near the helices α4 and α5 at the membrane proximal, and the smaller hydrophobic group is placed near the narrow end of the cupin barrel that is further from the membrane.

The chemical interactions between the RC-33 analogs and the residues at the S1R binding site can be described in detail by using IFPs. This method has been commonly used for identifying the relevant residues involved in protein-ligand affinities (Caballero et al., 2018; Navarro-Retamal and Caballero, 2018; Velázquez-Libera et al., 2018). IFPs capture and label the chemical contacts between a target protein and a set of its ligands as a whole. The chemotypes are identified with the following labels: P (polar groups), H (hydrophobic groups), A (hydrogen bonds where the residue is the acceptor), D (hydrogen bonds

where the residue is the donor), Ar (aromatic groups), and Ch (electrostatic interactions with charged groups). IFPs also differentiate between contacts with backbone and contacts with side-chain functional groups. We calculated IFPs by considering the S1R-ligand complexes formed by our docked structures.

The calculated IFPs are reported in **Figure 2**. The IFP analysis applied to the complexes between S1R and the RC-33 analogs obtained by docking revealed that 29 S1R residues had contacts with ligands. These residues and their positions in the S1R secondary structure are depicted in **Figure 2A**. The S1R binding site is mainly hydrophobic; in fact, the vast majority of the observed interactions are hydrophobic or aromatic when analyzing the occurrence of chemical contacts in the studied structures (**Figure 2C**).

The residues with polar interactions were identified in the plots of percent of occurrence obtained from IFP calculations (**Figure 2**). The residue E172 at the sheet β10 has polar contributions in 100% of the total structures, forming a salt-bridge and it also acts as HB acceptor in 80% of the studied structures. The residue D126 at the sheet β5 was identified with polar contributions in more than 50% of the studied structures. The residue T181 at the helix α4 has polar contributions in more than 80% of the studied structures. Finally, the residues S117 (backbone and side chain), H154 (side chain), and T202

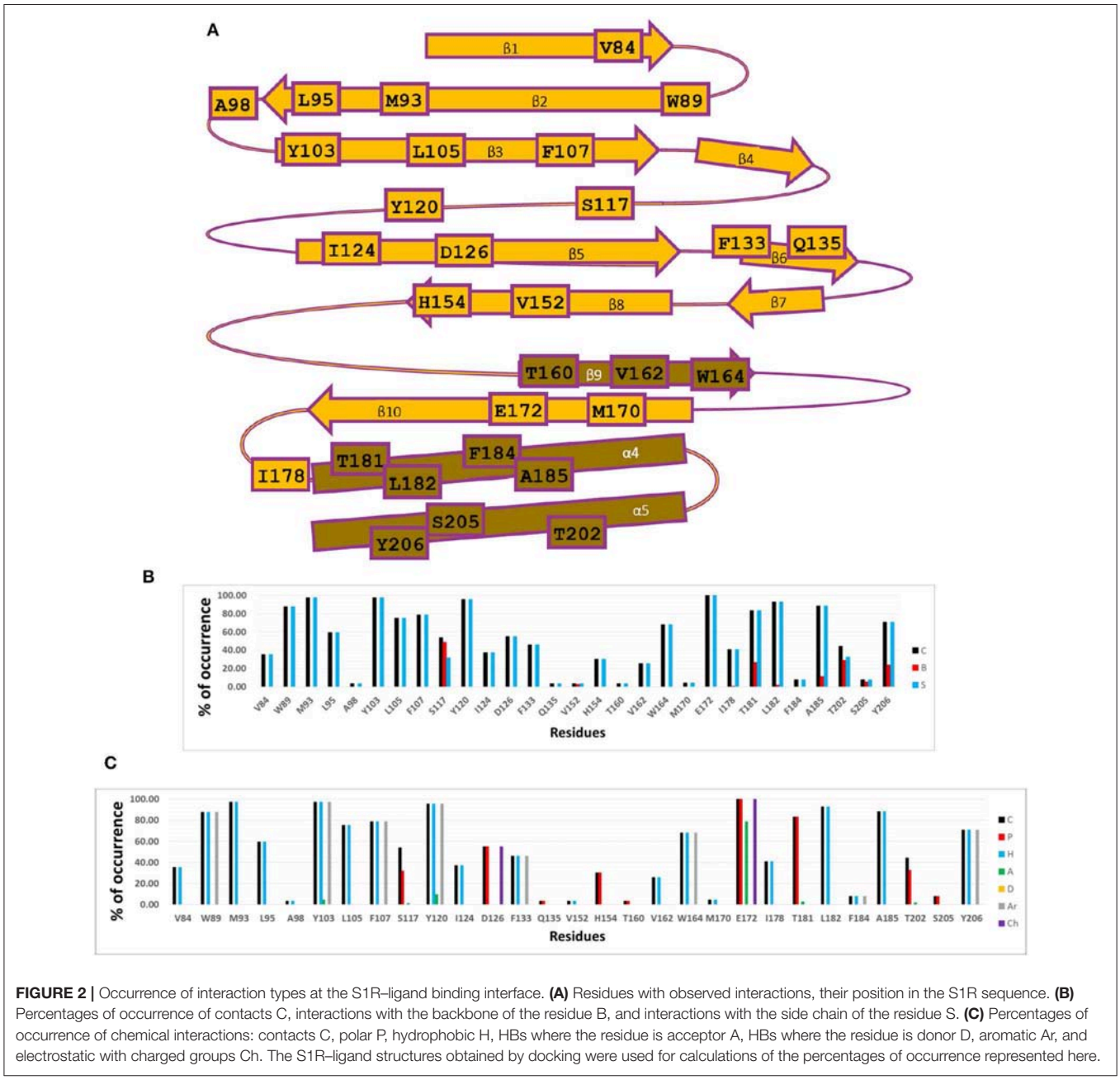


FIGURE 2 | Occurrence of interaction types at the S1R–ligand binding interface. **(A)** Residues with observed interactions, their position in the S1R sequence. **(B)** Percentages of occurrence of contacts C, interactions with the backbone of the residue B, and interactions with the side chain of the residue S. **(C)** Percentages of occurrence of chemical interactions: contacts C, polar P, hydrophobic H, HBs where the residue is acceptor A, HBs where the residue is donor D, aromatic Ar, and electrostatic with charged groups Ch. The S1R–ligand structures obtained by docking were used for calculations of the percentages of occurrence represented here.

TABLE 3 | Statistical information of the 3D-QSAR models.

Fields	NC	R^2	S	Q^2	S_{LOO}	R^2_{test}	S_{test}	%S	%E
S	8	0.98	0.13	0.64	0.54	0.34	0.81	1	
E	4	0.81	0.39	0.54	0.62	0.59	0.63	–	1
SE	7	0.97	0.15	0.70	0.50	0.61	0.62	0.88	0.12

NC is the number of components; S is the standard deviation of the fitted activity of the training set; R^2, Q^2, and R^2_{test} are the coefficients of correlation of the training set, LOO cross validation, and test set, respectively; S_{LOO} is the standard deviation of the LOO cross validation, and S_{test} is the standard deviation of the test predictions. %S and %E are the relative contributions of the steric (S) and the electrostatic (E) fields, respectively.

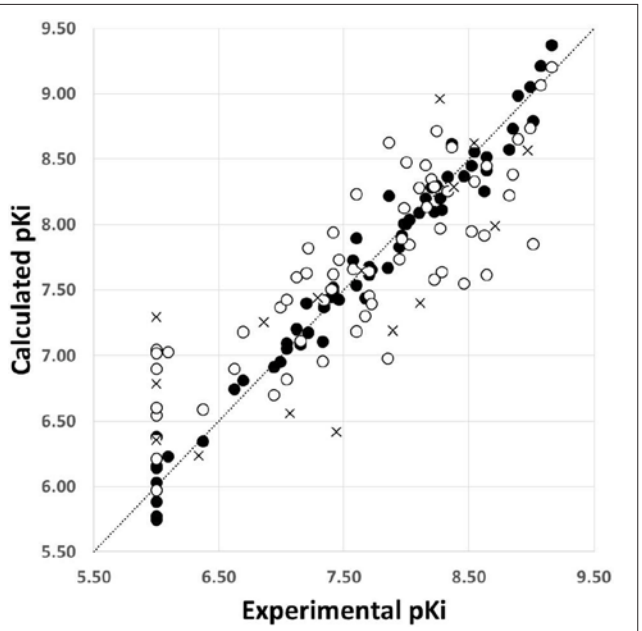


FIGURE 3 | Scatter plot of the experimental activities vs. predicted activities for the model SE: (•) training set predictions, (○) LOO cross-validation predictions, and (×) test set predictions.

(backbone and side chain) have polar contributions in around 30% of the studied structures.

Several residues with aromatic interactions were also identified in the plots of percent of occurrence obtained from IFP calculations (**Figure 2**). The residues with aromatic interactions were important for the shape of the S1R binding site because they restrict the space of the pockets. Four aromatic residues located at the center of the binding site (W89, Y103, F107, and Y120) were identified by the IFP calculations with percent of occurrence values above 80%. These residues cause a bottleneck just in front of the residue E172; therefore they could help to orient the positively charged N of the ligands to form the salt bridge. At the same time, they could stabilize the presence of the positive charge by means of π-cation interactions. The aromatic residues F133 at the sheet β6 and W164 the sheet β9, located close to the narrower end of the cupin β-barrel, have percent of occurrence values of 50 and 70%, respectively. On the other hand, the residue Y206, located at the helix α5, has a percent of occurrence value of 70%.

The remaining residues with hydrophobic interactions were also identified in the plots of percent of occurrence obtained from IFP calculations (**Figure 2**). The residues identified with percent of occurrence above 75% M93 (at β2), L105 (at β3), and L182/A185 (at α4) are located at the bigger hydrophobic pocket. The residues V84 (at β1), L95 (at β2), and I178 (at the loop between β10 and α4) are also located at the bigger hydrophobic pocket and were identified by IFP calculations with lower percent of occurrences, and the residue I124 at β5, located at the smaller hydrophobic pocket, had a percent of occurrence below 40%.

In general, the reported IFPs identify the most important S1R residues which establish chemical interactions with RC-33 analogs. Furthermore, it could be useful for the understanding of the interactions between S1R and its ligands.

3D-QSAR Results

We constructed the 3D-QSAR models based on docking alignment; therefore, the docked structures were included in a box for creating the relevant fields, since they are models of the ligand conformations inside the S1R binding site. The docking-based or receptor-guided alignment 3D-QSAR is a well-documented method in literature (Guasch et al., 2012; Navarro-Retamal and Caballero, 2016; Muñoz-Gutiérrez et al., 2017). Three 3D-QSAR models were trained using the steric field (Model S), the electrostatic field (Model E), and the combination of both fields (Model SE). The most reliable models were selected by measuring the LOO cross-validation performance ($Q^2 > 0.5$) and the test set predictions ($R^2_{test} > 0.5$).

Table 3 lists the description and statistical information of the best 3D-QSAR models. This report proved that model S has better (LOO) cross-validation Q^2 than model E. However, when both steric and electrostatic fields are tied together in the more complex model SE, the Q^2 value increases; therefore, this model, which had a $Q^2 = 0.70$ including seven components, containing a major contribution of the steric field (88%), was identified as the model best describing the structure-activity relationship of the studied RC-33 analogs. These results reflect that the steric features are mandatory for modulating the agonistic activities of the studied compounds. This is reasonable considering that the S1R binding site is mostly hydrophobic.

The model SE explains 97% of the variance and has a low standard deviation ($S = 0.15$). The predictions of pKi values for the 64 RC-33 analogs in the training set using the model SE are reported in **Table 1** and the correlations between the predicted and experimental pKi values (from training and LOO cross validation) are shown in **Figure 3**. It is possible to observe that the selected model fitted adequately the whole dataset; it is noteworthy that the more potent compounds had an outstanding performance. When the model SE was used to predict the pKi values of the test set compounds, well results were also found, reflected by the value of $R^2_{test} = 0.61$. The predicted pKi values for the test set are listed in **Table 1**, and the correlation between the calculated and experimental pKi values are plotted in **Figure 3**.

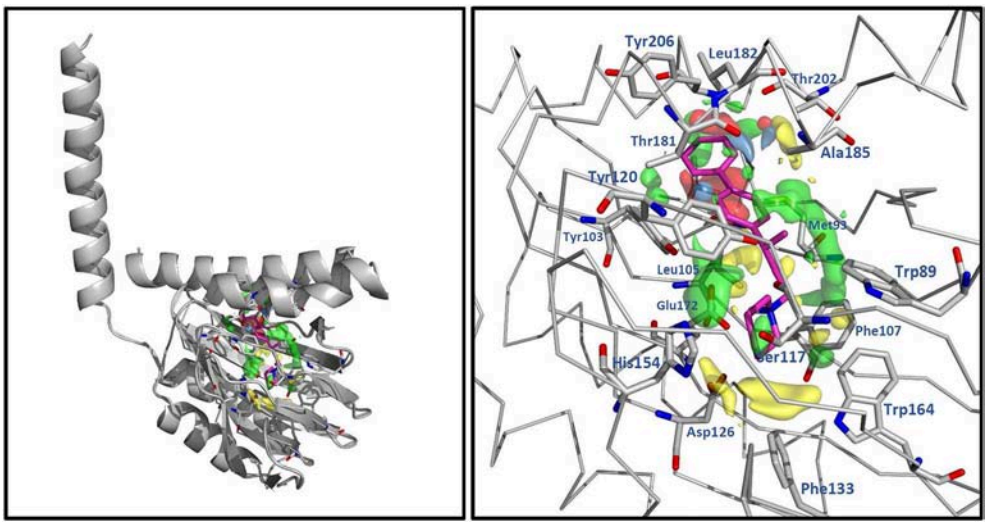


FIGURE 4 | 3D-QSAR contour maps for the RC-33 analogs (model SE). The steric field is represented by green and yellow isopleths: the green ones indicate regions where bulky groups enhance the activity, and the yellow ones indicate regions where bulky groups disfavor the activity. The electrostatic field is represented by blue and red isopleths: the blue ones indicate regions where an increase of positive charge enhances the activity and the red ones indicate regions where an increase of negative charge enhances the activity. RC-33 is shown inside the fields.

Figure 4 shows contour plots of the steric and electrostatic fields projected onto the docked structure of RC-33 for association between the fields, the compounds of the whole set, and the residues at the S1R binding site. In this figure, green and yellow contours represent regions with positive and negative steric components, respectively. It is noted that positive steric components have a major role. A great green contour in front of the 3-phenylbutylamine, and near the residues V84, W89, F107, and A185, indicates that bulk groups are desired in this region. It is noteworthy that the most active compounds such as RC-33, **7**, **15**, **27**, **28**, and **33** has the methyl group of the 3-phenylbutylamine in this region, but the majority of the less active compounds such as **8**, **29**, **41**, **64**, and **67** have this group deeper into the bigger hydrophobic pocket. Another three green contours are located near the piperidine of RC-33 and the residues Y120, S117, and W164 indicating that this group or another bulky group in this region is needed. In general, compounds with a dimethylamine in this region (compounds **2**–**15**) are less active than similar compounds that contain piperidine. Another green contour near the residues Y103 and E172 reflects that several active compounds contain the methyl group of the 3-phenylbutylamine in this region. Another green contours are located at the bigger hydrophobic pocket near the residues Y103, Y206, and T202, indicating the preference of a bulky group in this region. For instance, the biphenyl group in compound **7** is preferred instead the phenyl group in compound **8** because the former group fills the entire space of the bigger hydrophobic pocket. Several yellow contours were identified near the residues W164, L105, F107, and T202. All of them are close to the green contours both in the bigger and smaller pockets, and reflect the complexity of the steric field inside the S1R binding site.

In **Figure 4**, blue and red contours represent regions with positive and negative electrostatic components, respectively; all of them are small and are located inside the bigger hydrophobic pocket. The blue contours are near the residues T181, A185, L182, and the backbone of Y206, and the red contours are near the residues A92, L95, L105, L182, and T202. The blue contours are located in regions where ligands placed hydroxyl groups and their pKi values are between 7 and 7.8 (moderate activities). For instance, compounds **13** and **77** have hydroxyl close to the backbone of Y206, **75,** and **76** have hydroxyl close to A185, and **53**, 55, and **78** have hydroxyl close to T181. The red contours are located in regions where ligands placed OMe groups and the activity is increased. For instance, compounds **22** and **30** that contain OMe have better activities than compounds **21** and **29** without this group.

The docking-based 3D-QSAR methodology allows establishing a comparison between the chemical features that describe the structure-activity relationship of bioactive ligands and the protein binding site (Alzate-Morales and Caballero, 2010; Caballero et al., 2011; Quesada-Romero et al., 2014; Mena-Ulecia et al., 2015; Muñoz-Gutiérrez et al., 2017). The contour plots in receptor-based 3D-QSAR are not receptor maps, but they solve another key point of the description of the differential activities: different potency in activities is connected with different chemical environments and interactions. The docking and 3D-QSAR methods applied to the study of RC-33 analogs give more information about the structure of S1R-ligand complexes, and identify important chemical features to take into account in the future design of potent S1R ligands. We feel that another similar studies on other series of compounds will be reported during next years.

CONCLUSION

This is the first structure-based molecular modeling investigation a few years after the elucidation of the S1R crystallographic structure; therefore, details of the binding poses and the chemical interactions in the binding site are described. Binding orientations and structure-activity relationship of RC-33 analogs as S1R agonists were studied by using molecular docking and 3D-QSAR methods.

Docking poses obtained for the studied compounds inside the S1R binding site explain the interactions between the well-known theoretical pharmacophore model reported for these compounds (elucidated before the knowledge of the S1R 3D structure) and the residues located at the binding site. They also reproduced structural features reported for complexes between S1R and PD144418, 4-IBP, and other active ligands. The docking analysis, including the IFP calculations, confirmed the preponderant role of E172 forming a salt bridge with the positively charged N of the ligands. Furthermore, docking experiments also identified the importance role of the aromatic residues delimiting the shape of the S1R binding site: specifically, W89, Y103, F107, and Y120 which are at the center of the binding site, F133 and W164 which are close to the narrower end of the cupin β-barrel, and Y206 which is close to the helix α5.

A receptor-guided alignment 3D-QSAR model with adequate statistical significance and acceptable prediction power was obtained. Steric and electrostatic features had contributions to the differential potency of the agonists, with a major role of the steric ones. The 3D-QSAR model demonstrated that an implicit correlation is found in the data under analysis between the chemical features of the compounds in their active conformations and their interactions in the pockets of the S1R binding site.

Overall, the information reported here, derived from the recently reported S1R structure, will be useful for the future research in the design of novel S1R ligands.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

The work was completed by cooperation of all authors. JC was responsible for the study of concept and design of the project. JV-L performed the docking RMSD analysis, IFPs, and 3D-QSAR calculations. GR and CN-R performed the docking calculations. SC and JC drafted and revised the manuscript.

FUNDING

This research was funded by FONDECYT Regular grant number 1170718 (JC) and FONDECYT Postdoc grant number 3170434 (CN-R). The authors gratefully acknowledge MIUR for the doctoral fellowship to GR and thankfully recognize Scuola di Alta Formazione Dottorale of University of Pavia for the mobility research scholarship provided to GR.

REFERENCES

Ablordeppey, S. Y., Fischer, J. B., and Glennon, R. A. (2000). Is a nitrogen atom an important pharmacophoric element in sigma ligand binding? *Bioorg. Med. Chem.* 8, 2105–2111. doi: 10.1016/S0968-0896(00)00148-6

Almerico, A. M., Tutone, M., and Lauria, A. (2012). Receptor-guided 3D-QSAR approach for the discovery of c-kit tyrosine kinase inhibitors. *J. Mol. Model.* 18, 2885–2895. doi: 10.1007/s00894-011-1304-0

Alon, A., Schmidt, H. R., Wood, M. D., Sahn, J. J., Martin, S. F., and Kruse, A. C. (2017). Identification of the gene that codes for the σ2 receptor. *Proc. Natl. Acad. Sci. U.S.A.* 114, 7160–7165. doi: 10.1073/pnas.1705154114

Alzate-Morales, J., and Caballero, J. (2010). Computational study of the interactions between guanine derivatives and cyclin-dependent kinase 2 (CDK2) by CoMFA and QM/MM. *J. Chem. Inf. Model.* 50, 110–122. doi: 10.1021/ci900302z

Amaning, K., Lowinski, M., Vallee, F., Steier, V., Marcireau, C., Ugolini, A., et al. (2013). The use of virtual screening and differential scanning fluorimetry for the rapid identification of fragments active against MEK1. *Bioorg. Med. Chem. Lett.* 23, 3620–3626. doi: 10.1016/j.bmcl.2013.04.003

Batra, J., Szabó, A., Caulfield, T. R., Soares, A. S., Sahin-Tóth, M., and Radisky, E. S. (2013). Long-range electrostatic complementarity governs substrate recognition by human chymotrypsin C, a key regulator of digestive enzyme activation. *J. Biol. Chem.* 288, 9848–9859. doi: 10.1074/jbc.M113.457382

Boháč, M., Loeprecht, B., Damborský, J., and Schüürmann, G. (2002). Impact of Orthogonal Signal Correction (OSC) on the predictive ability of CoMFA models for the ciliate toxicity of nitrobenzenes. *Q. Struct. Activit. Relationsh.* 21, 3–11. doi: 10.1002/1521-3838(200205)21:1<3::AID-QSAR3>3.0.CO;2-D

Caballero, J., Morales-Bayuelo, A., and Navarro-Retamal, C. (2018). *Mycobacterium tuberculosis* serine/threonine protein kinases: structural information for the design of their specific ATP-competitive inhibitors. *J. Comput. Aided Mol. Des.* 32, 1315–1336. doi: 10.1007/s10822-018-0173-3

Caballero, J., Quiliano, M., Alzate-Morales, J. H., Zimic, M., and Deharo, E. (2011). Docking and quantitative structure–activity relationship studies for 3-fluoro-4-(pyrrolo[2,1-f][1,2,4]triazin-4-yloxy)aniline, 3-fluoro-4-(1H-pyrrolo[2,3-b]pyridin-4-yloxy)aniline, and 4-(4-amino-2-fluorophenoxy)-2-pyridinylamine derivatives as c-Met kinase inhibitors. *J. Comput. Aided Mol. Des.* 25, 349–369. doi: 10.1007/s10822-011-9425-1

Caballero, J., Zilocchi, S., Tiznado, W., Rossi, D., and Collina, S. (2012). Models of the pharmacophoric pattern and affinity trend of methyl 2-(aminomethyl)-1-phenylcyclopropane-1-carboxylate derivatives as σ1 ligands. *Mol. Simul.* 38, 227–235. doi: 10.1080/08927022.2011.614243

Chu, U. B., and Ruoho, A. E. (2016). Biochemical pharmacology of the sigma-1 receptor. *Mol. Pharmacol.* 89, 142–153. doi: 10.1124/mol.115.101170

Collina, S., Bignardi, E., Rui, M., Rossi, D., Gaggeri, R., Zamagni, A., et al. (2017a). Are sigma modulators an effective opportunity for cancer treatment? A patent overview (1996–2016). *Expert Opin. Ther. Pat.* 27, 565–578. doi: 10.1080/13543776.2017.1276569

Collina, S., Loddo, G., Urbano, M., Linati, L., Callegari, A., Ortuso, F., et al. (2007). Design, synthesis, and SAR analysis of novel selective [sigma]1 ligands. *Bioorg. Med. Chem.* 15, 771–783. doi: 10.1016/j.bmc.2006.10.048

Collina, S., Rui, M., Stotani, S., Bignardi, E., Rossi, D., Curti, D., et al. (2017b). Are sigma receptor modulators a weapon against multiple sclerosis disease? *Future Med. Chem.* 9, 2029–2051. doi: 10.4155/fmc-2017-0122

DeCoster, M. A., Klette, K. L., Knight, E. S., and Tortella, F. C. (1995). σ receptor-mediated neuroprotection against glutamate toxicity in primary rat neuronal cultures. *Brain Res.* 671, 45–53. doi: 10.1016/0006-8993(94)01294-R

Deng, Z., Chuaqui, C., and Singh, J. (2004). Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* 47, 337–344. doi: 10.1021/jm030331x

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* 47, 1739–1749. doi: 10.1021/jm0306430

Furuse, T., and Hashimoto, K. (2010). Sigma-1 receptor agonist fluvoxamine for delirium in patients with Alzheimer's disease. *Ann. Gen. Psychiatry* 9:6. doi: 10.1186/1744-859X-9-6

Glennon, R. A. (2005). Pharmacophore identification for sigma-1 (sigma1) receptor binding: application of the "deconstruction-reconstruction-elaboration" approach. *Mini Rev. Med. Chem.* 5, 927–940. doi: 10.2174/138955705774329519

Glennon, R. A., Ismaiel, A. M., Ablordeppey, S., El-Ashmawy, M., and Fisher, J. B. (2004). Thioxanthene-derived analogs as sigma(1) receptor ligands. *Bioorg. Med. Chem. Lett.* 14, 2217–2220. doi: 10.1016/j.bmcl.2004.02.018

Guasch, L., Sala, E., Valls, C., Mulero, M., Pujadas, G., and Garcia-Vallvé, S. (2012). Development of docking-based 3D-QSAR models for PPARgamma full agonists. *J. Mol. Graph. Model.* 36, 1–9. doi: 10.1016/j.jmgm.2012.03.001

Hajipour, A. R., Fontanilla, D., Chu, U. B., Arbabian, M., and Ruoho, A. E. (2010). Synthesis and characterization of N,N-dialkyl and N-alkyl-N-aralkyl fenpropimorph-derived compounds as high affinity ligands for sigma receptors. *Bioorg. Med. Chem.* 18, 4397–4404. doi: 10.1016/j.bmc.2010.04.078

Halgren, T. A. (1996). Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.* 17, 520–552. doi: 10.1002/(SICI)1096-987X(199604)17:5/6<520::AID-JCC2>3.0.CO;2-W

Hanner, M., Moebius, F. F., Flandorfer, A., Knaus, H. G., Striessnig, J., Kempner, E., et al. (1996). Purification, molecular cloning, and expression of the mammalian sigma1-binding site. *Proc. Natl. Acad. Sci. U.S.A.* 93, 8072–8077. doi: 10.1073/pnas.93.15.8072

Hashimoto, K. (2009). Can the sigma-1 receptor agonist fluvoxamine prevent schizophrenia? *CNS Neurol. Disord. Drug Targets* 8, 470–474. doi: 10.2174/187152709789824633

Hayashi, T., and Su, T.-P. (2004). Sigma-1 receptor ligands: potential in the treatment of neuropsychiatric disorders. *CNS Drugs* 18, 269–284. doi: 10.2165/00023210-200418050-00001

Kastenholz, M. A., Pastor, M., Cruciani, G., Haaksma, E. E., and Fox, T. (2000). GRID/CPCA: a new computational tool to design selective ligands. *J. Med. Chem.* 43, 3033–3044. doi: 10.1021/jm000934y

Marra, A., Rossi, D., Pignataro, L., Bigogno, C., Canta, A., Oggioni, N., et al. (2016). Toward the identification of neuroprotective agents: g-scale synthesis, pharmacokinetic evaluation and CNS distribution of (R)-RC-33, a promising SIGMA1 receptor agonist. *Future Med. Chem.* 8, 287–295. doi: 10.4155/fmc.15.191

Martin, W. R., Eades, C. G., Thompson, J. A., Huppler, R. E., and Gilbert, P. E. (1976). The effects of morphine- and nalorphine- like drugs in the non-dependent and morphine-dependent chronic spinal dog. *J. Pharmacol. Exp. Ther.* 197, 517–532.

Mavlyutov, T. A., Guo, L.-W., Epstein, M. L., and Ruoho, A. E. (2015). Role of the sigma-1 receptor in Amyotrophic Lateral Sclerosis (ALS). *J. Pharmacol. Sci.* 127, 10–16. doi: 10.1016/j.jphs.2014.12.013

Mena-Ulecia, K., Tiznado, W., and Caballero, J. (2015). Study of the differential activity of thrombin inhibitors using docking, QSAR, molecular dynamics, and MM-GBSA. *PLoS ONE* 10:e0142774. doi: 10.1371/journal.pone.0142774

Meyer, D. A., Carta, M., Partridge, L. D., Covey, D. F., and Valenzuela, C. F. (2002). Neurosteroids enhance spontaneous glutamate release in hippocampal neurons possible role of metabotropic σ1-like receptors. *J. Biol. Chem.* 277, 28725–28732. doi: 10.1074/jbc.M202592200

Mishina, M., Ishiwata, K., Ishii, K., Kitamura, S., Kimura, Y., Kawamura, K., et al. (2005). Function of sigma1 receptors in Parkinson's disease. *Acta Neurol. Scand.* 112, 103–107. doi: 10.1111/j.1600-0404.2005.00432.x

Monnet, F. P., and Maurice, T. (2006). The sigma1 protein as a target for the non-genomic effects of neuro(active)steroids: molecular, physiological, and behavioral aspects. *J. Pharmacol. Sci.* 100, 93–118. doi: 10.1254/jphs.CR0050032

Morin-Surun, M. P., Collin, T., Denavit-Saubié, M., Baulieu, E. E., and Monnet, F. P. (1999). Intracellular sigma1 receptor modulates phospholipase C and protein kinase C activities in the brainstem. *Proc. Natl. Acad. Sci. U.S.A.* 96, 8196–8199. doi: 10.1073/pnas.96.14.8196

Muñoz-Gutierrez, C., Adasme-Carreño, F., Fuentes, E., Palomo, I., and Caballero, J. (2016). Computational study of the binding orientation and affinity of PPARγ agonists: inclusion of ligand-induced fit by cross-docking. *RSC Adv.* 6, 64756–64768. doi: 10.1039/C6RA12084A

Muñoz-Gutiérrez, C., Cáceres-Rojas, D., Adasme-Carreño, F., Palomo, I., Fuentes, E., and Caballero, J. (2017). Docking and quantitative structure-activity relationship of bi-cyclic heteroaromatic pyridazinone and pyrazolone derivatives as phosphodiesterase 3A (PDE3A) inhibitors. *PLoS ONE* 12:e0189213. doi: 10.1371/journal.pone.0189213

Navarro-Retamal, C., and Caballero, J. (2016). Flavonoids as CDK1 inhibitors: insights in their binding orientations and structure-activity relationship. *PLoS ONE* 11:e0161111. doi: 10.1371/journal.pone.0161111

Navarro-Retamal, C., and Caballero, J. (2018). "Molecular modeling of tau Proline-Directed Protein Kinase (PDPK) inhibitors," in *Computational Modeling of Drugs Against Alzheimer's Disease Neuromethods*, ed K. Roy (New York, NY: Humana Press), 305–345. doi: 10.1007/978-1-4939-7404-7_13

Osguthorpe, D. J., Sherman, W., and Hagler, A. T. (2012). Generation of receptor structural ensembles for virtual screening using binding site shape analysis and clustering. *Chem. Biol. Drug Des.* 80, 182–193. doi: 10.1111/j.1747-0285.2012.01396.x

Pastor, M., Cruciani, G., and Clementi, S. (1997). Smart region definition: a new way to improve the predictive ability and interpretability of three-dimensional quantitative structure-activity relationships. *J. Med. Chem.* 40, 1455–1464. doi: 10.1021/jm9608016

Peviani, M., Salvaneschi, E., Bontempi, L., Petese, A., Manzo, A., Rossi, D., et al. (2014). Neuroprotective effects of the Sigma-1 receptor (S1R) agonist PRE-084, in a mouse model of motor neuron disease not linked to SOD1 mutation. *Neurobiol. Dis.* 62, 218–232. doi: 10.1016/j.nbd.2013.10.010

Prezzavento, O., Campisi, A., Ronsisvalle, S., Li Volti, G., Marrazzo, A., Bramanti, V., et al. (2007). Novel sigma receptor ligands: synthesis and biological profile. *J. Med. Chem.* 50, 951–961. doi: 10.1021/jm0611197

Quesada-Romero, L., and Caballero, J. (2014). Docking and quantitative structure–activity relationship of oxadiazole derivates as inhibitors of GSK3beta. *Mol. Divers.* 18, 149–159. doi: 10.1007/s11030-013-9483-5

Quesada-Romero, L., Mena-Ulecia, K., Tiznado, W., and Caballero, J. (2014). Insights into the interactions between maleimide derivates and GSK3β combining molecular docking and QSAR. *PLoS ONE* 9:e102212. doi: 10.1371/journal.pone.0102212

Quesada-Romero, L., Mena-Ulecia, K., Zuñiga, M., De-la-Torre, P., Rossi, D., Tiznado, W., et al. (2015). Optimal graph-based and simplified molecular input line entry System-based descriptors for quantitative structure–activity relationship analysis of arylalkylaminoalcohols, arylalkenylamines, and arylalkylamines as σ1 receptor ligands. *J. Chemom.* 29, 13–20. doi: 10.1002/cem.2650

Ramírez, D., and Caballero, J. (2018). Is it reliable to take the molecular docking top scoring position as the best solution without considering available structural data? *Molecules* 23:1038. doi: 10.3390/molecules23051038

Rossi, D., Marra, A., Picconi, P., Serra, M., Catenacci, L., Sorrenti, M., et al. (2013a). Identification of RC-33 as a potent and selective σ1 receptor agonist potentiating NGF-induced neurite outgrowth in PC12 cells. Part 2: g-scale synthesis, physicochemical characterization and *in vitro* metabolic stability. *Bioorg. Med. Chem.* 21, 2577–2586. doi: 10.1016/j.bmc.2013.02.029

Rossi, D., Marra, A., Rui, M., Laurini, E., Fermeglia, M., Pricl, S., et al. (2015). A step forward in the sigma enigma: a role for chirality in the sigma1 receptor–ligand interaction? *Med. Chem. Commun.* 6, 138–146. doi: 10.1039/C4MD00349G

Rossi, D., Pedrali, A., Gaggeri, R., Marra, A., Pignataro, L., Laurini, E., et al. (2013b). Chemical, pharmacological, and *in vitro* metabolic stability studies on enantiomerically pure RC-33 compounds: promising neuroprotective agents acting as σ1 receptor agonists. *ChemMedChem* 8, 1514–1527. doi: 10.1002/cmdc.201300218

Rossi, D., Pedrali, A., Urbano, M., Gaggeri, R., Serra, M., Fernández, L., et al. (2011). Identification of a potent and selective σ1 receptor agonist potentiating NGF-induced neurite outgrowth in PC12 cells. *Bioorg. Med. Chem.* 19, 6210–6224. doi: 10.1016/j.bmc.2011.09.016

Rossi, D., Rui, M., Di Giacomo, M., Schepmann, D., Wünsch, B., Monteleone, S., et al. (2017). Gaining in pan-affinity towards sigma 1 and sigma 2 receptors. SAR studies on arylalkylamines. *Bioorg. Med. Chem.* 25, 11–19. doi: 10.1016/j.bmc.2016.10.005

Rossi, D., Urbano, M., Pedrali, A., Serra, M., Zampieri, D., Mamolo, M. G., et al. (2010). Design, synthesis and SAR analysis of novel selective sigma1 ligands (Part 2). *Bioorg. Med. Chem.* 18, 1204–1212. doi: 10.1016/j.bmc.2009.12.039

Rui, M., Rossi, D., Marra, A., Paolillo, M., Schinelli, S., Curti, D., et al. (2016). Synthesis and biological evaluation of new aryl-alkyl(alkenyl)-4-benzylpiperidines, novel Sigma Receptor (SR) modulators, as potential anticancer-agents. *Eur. J. Med. Chem.* 124, 649–665. doi: 10.1016/j.ejmech.2016.08.067

Schmidt, H. R., Betz, R. M., Dror, R. O., and Kruse, A. C. (2018). Structural basis for σ1 receptor ligand recognition. *Nat. Struct. Mol. Biol.* 25, 981–987. doi: 10.1038/s41594-018-0137-2

Schmidt, H. R., Zheng, S., Gurpinar, E., Koehl, A., Manglik, A., and Kruse, A. C. (2016). Crystal structure of the human σ1 receptor. *Nature* 532, 527–530. doi: 10.1038/nature17391

Shelley, J. C., Cholleti, A., Frye, L. L., Greenwood, J. R., Timlin, M. R., and Uchimaya, M. (2007). Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. *J. Comput. Aided Mol. Des.* 21, 681–691. doi: 10.1007/s10822-007-9133-z

Singh, J., Deng, Z., Narale, G., and Chuaqui, C. (2006). Structural interaction fingerprints: a new approach to organizing, mining, analyzing, and designing protein-small molecule complexes. *Chem. Biol. Drug Des.* 67, 5–12. doi: 10.1111/j.1747-0285.2005.00323.x

Tesei, A., Cortesi, M., Zamagni, A., Arienti, C., Pignatta, S., Zanoni, M., et al. (2018). Sigma receptors as endoplasmic reticulum stress "Gatekeepers" and their modulators as emerging new weapons in the fight against cancer. *Front. Pharmacol.* 9:711. doi: 10.3389/fphar.2018.00711

Tosco, P., and Balle, T. (2011). Open3DQSAR: a new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields. *J. Mol. Model.* 17, 201–208. doi: 10.1007/s00894-010-0684-x

Urbano, M., Collina, S., Rossi, D., Baraglia, A. C., Alcaro, S., Artese, A., et al. (2007). Design and synthesis of a (N-Alkylaminoalkyl-Substituted)arylalkenylamide drug discovery library. *Lett. Drug Des. Discov.* 4, 605–610. doi: 10.2174/157018007782794581

Vavers, E., Svalbe, B., Lauberte, L., Stonans, I., Misane, I., Dambrova, M., et al. (2017). The activity of selective sigma-1 receptor ligands in seizure models *in vivo*. *Behav. Brain Res.* 328, 13–18. doi: 10.1016/j.bbr.2017.04.008

Velázquez-Libera, J. L., Navarro-Retamal, C., and Caballero, J. (2018). Insights into the structural requirements of 2(S)-amino-6-boronohexanoic acid derivatives as arginase I inhibitors: 3D-QSAR, docking, and interaction fingerprint studies. *Int. J. Mol. Sci.* 19:2956. doi: 10.3390/ijms19102956

Walker, J. M., Bowen, W. D., Walker, F. O., Matsumoto, R. R., De Costa, B., and Rice, K. C. (1990). Sigma receptors: biology and function. *Pharmacol. Rev.* 42, 355–402.

Wu, D., Wang, Q., Assary, R. S., Broadbelt, L. J., and Krilov, G. (2011). A computational approach to design and evaluate enzymatic reaction pathways: application to 1-butanol production from pyruvate. *J. Chem. Inf. Model.* 51, 1634–1647. doi: 10.1021/ci2000659

Yagasaki, Y., Numakawa, T., Kumamaru, E., Hayashi, T., Su, T.-P., and Kunugi, H. (2006). Chronic antidepressants potentiate via sigma-1 receptors the brain-derived neurotrophic factor-induced signaling for glutamate release. *J. Biol. Chem.* 281, 12941–12949. doi: 10.1074/jbc.M508157200

Yous, S., Wallez, V., Belloir, M., Caignard, D. H., McCurdy, C. R., and Poupaert, J. H. (2005). Novel 2(3H)-benzothiazolones as highly potent and selective sigma-1 receptor ligands. *Med. Chem. Res.* 14, 158–168. doi: 10.1007/s00044-005-0131-1

Zampieri, D., Grazia Mamolo, M., Laurini, E., Zanette, C., Florio, C., Collina, S., et al. (2009). Substituted benzo[d]oxazol-2(3H)-one derivatives with preference for the sigma1 binding site. *Eur. J. Med. Chem.* 44, 124–130. doi: 10.1016/j.ejmech.2008.03.011

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Velázquez-Libera, Rossino, Navarro-Retamal, Collina and Caballero. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

PERSPECTIVE
published: 11 July 2019
doi: 10.3389/fchem.2019.00498

Refinement and Rescoring of Virtual Screening Results

Giulio Rastelli and Luca Pinzi*

Department of Life Sciences, University of Modena and Reggio Emilia, Modena, Italy

High-throughput docking is an established computational screening approach in drug design. This methodology enables a rapid identification of biologically active hit compounds, providing an efficient and cost-effective complement or alternative to experimental high-throughput screenings. However, limitations inherent to the methodology make docking results inevitably approximate. Two major Achille's heels include the use of approximated scoring functions and the limited sampling of the ligand-target complexes. Therefore, docking results require careful evaluation and further post-docking analyses. In this article, we will overview our approach to post-docking analysis in virtual screenings. BEAR (Binding Estimation After Refinement) was developed as a post-docking processing tool that refines docking poses by means of molecular dynamics (MD) and then rescores the ligands based on more accurate scoring functions (MM-PB(GB)SA). The tool has been validated and used prospectively in drug discovery applications. Future directions regarding refinement and rescoring in virtual screening are discussed.

Keywords: docking, post-docking, virtual screening, molecular dynamics, BEAR, binding free energy

OPEN ACCESS

Edited by:
Simone Brogi,
University of Pisa, Italy

Reviewed by:
Demet Akten,
Kadir Has University, Turkey
Horacio Pérez-Sánchez,
Catholic University San Antonio of Murcia, Spain
Tingjun Hou,
Zhejiang University, China

****Correspondence:***
Giulio Rastelli
giulio.rastelli@unimore.it

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *23 April 2019*
Accepted: *28 June 2019*
Published: *11 July 2019*

Citation:
Rastelli G and Pinzi L (2019) Refinement and Rescoring of Virtual Screening Results. Front. Chem. 7:498. doi: 10.3389/fchem.2019.00498

INTRODUCTION

High-throughput screening (HTS) is a widely used method for the discovery of biologically active hits. However, the high costs and the low hit rates characterizing such experiments often make HTS not affordable for academic labs or small companies (Sliwoski et al., 2014). As a consequence, high-throughput docking screenings represent an attractive alternative (Irwin and Shoichet, 2016). Structure-based virtual screenings (SBVSs) require the knowledge of the three-dimensional structure of the target of interest, as well as the access to large libraries of small molecules available in public databases (Kar and Roy, 2013; Rastelli, 2013). Docking programs generate binding poses of compounds in the active site of a target and evaluate the ligand binding strength by means of scoring functions (Lengauer and Rarey, 1996; Kitchen et al., 2004). Several docking software relying on different algorithms have been developed for virtual screening so far (Rarey et al., 1996; Morris et al., 1998; Friesner et al., 2004; Sánchez-Linares et al., 2012). However, although remarkable improvements have been obtained along the years, several drawbacks and limitations still exist (Huang and Zou, 2010; Rastelli, 2013). First of all, sampling the conformational space accessible to ligand-target complexes in an induced-fit context is a difficult and target-dependent task. To help overcoming such limitations, several *in silico* strategies including molecular dynamics or induced fit strategies have been introduced (Sherman et al., 2006; Nabuurs et al., 2007; Caporuscio and Rastelli, 2016). Secondly, docking scores and experimental binding affinities usually do not correlate, because screening large numbers of compounds in a reasonable time requires the use of approximate scoring functions. Together, the two effects imply that a variable number of false-positive and false-negative hits populate the ranked lists, which then require careful

evaluation and further post-docking analyses. Hence, it has become general opinion that docking results should be improved by means of more rigorous post-docking processing strategies. Several post-processing strategies have been developed to overcome docking limitations over the past decades. In particular, methods based on binding free energy estimations have demonstrated to provide higher hit rates and to be more suitable for ranking cognate ligands in virtual screening (Hou et al., 2011; Genheden and Ryde, 2015; Pu et al., 2017), the predicted binding free energy usually correlating better with experimental data (Brandsdal et al., 2003). One of the first reported energy-based methods is MM-PB(GB)SA, which was developed to more accurately assess the relative free energy of binding for a given macromolecular system from molecular dynamics simulations (Kollman et al., 2000). This method represented a remarkable step forward to the obtainment of *in silico* predicted binding affinities that are in good agreement with experiments. In fact, it was extensively used to evaluate the free energy of binding for a number of complexes in the last years (Gohlke et al., 2003; Hou and Yu, 2007; Ferri et al., 2009; Yang et al., 2012). For example, it was successfully used for identifying residue *hot-spots* outside the binding interface of the Ras–Raf and Ras–RalGDS protein-protein complexes, discussing also their implications for an allosteric activation of the proteins (Gohlke et al., 2003). More recently, this method was also employed for predicting binding affinities of few inhibitors of HIV-1 protease and to help rationalize drug resistance caused by the mutations on the enzyme binding site (Hou and Yu, 2007). However, it should also be noted that MM-PB(GB)SA results are dependent on the employed parameters and receptor structures used in the calculations (Xu et al., 2013; Sun et al., 2014, 2018; Genheden and Ryde, 2015). More accurate free energy-based methods have also been reported (Brandsdal et al., 2003; Jorgensen and Thomas, 2008; Parenti and Rastelli, 2012; Limongelli et al., 2013; De Vivo et al., 2016). Among them, it is worth mentioning the Free Energy Perturbation (FEP) method, which allows to estimate the free energy of binding of a ligand to a protein by decomposing the system through a series of "alchemical transformations" (Jorgensen and Thomas, 2008; De Vivo et al., 2016). More recently, funnel-metadynamics (FM) methods that use a funnel-shaped potential limiting the sampling space available for a ligand to bind/unbind to a protein have been proposed (Limongelli et al., 2013; De Vivo et al., 2016). However, although these methods demonstrated to accurately estimate ligand binding, they are time-consuming and therefore not suitable for virtual screening rescoring of large databases. With the aim of improving ligand-binding estimations of docking complexes at reasonable computational costs, we developed Binding Estimation After Refinement (BEAR) (Rastelli et al., 2009). BEAR is an automated post-docking tool based on conformational refinement of docking poses with molecular dynamics followed by a more accurate prediction of binding free energies performed with MM-PBSA and MM-GBSA, which take into account desolvation energies (Kuhn et al., 2005; Lyne et al., 2006; Rastelli et al., 2010; Genheden and Ryde, 2015). As it allows accurately rescoring docking poses in reasonable times, BEAR can be considered an efficient tool that could be routinely used for virtual screening. In this article, we will briefly describe the BEAR tool, providing an overview of the validation studies performed so far. Finally, we will describe its prospective applications in drug discovery campaigns and comment on future directions of refinement and rescoring methods.

THE BEAR TOOL

The BEAR workflow (**Figure 1**; Rastelli et al., 2009) consists of an initial pre-processing step in which hydrogen atoms are added to the protein, atomic charges (AM1-BCC) are calculated for the docked molecules, and missing force-field parameters are assigned. Then, topologies for the ligand, the protein, and the ligand-protein complex are built. In particular, ligand atom types are assigned according to the Generalized Amber Force Field (GAFF) (Wang et al., 2004), while, the atom types and charges of amino acids are assigned according to the Amber ff03 force field (Duan et al., 2003). The following iterative three steps procedure is based on molecular mechanics (MM) and molecular dynamics (MD) cycles. In particular, an initial MM energy minimization of the whole protein–ligand complex is performed, followed by a short MD simulation where the ligand is allowed to move, and a final re-minimization of the entire complex. All the minimization tasks are performed through 2000 steps without restraints, and with a distance-dependent dielectric constant $\varepsilon = 4r$ and a cutoff of 12 Å. The MD simulation is performed at 300 K for 100 ps, with the SHAKE parameter turned on and a time-step of 2.0 fs. This protocol allows evaluating the reliability of the predicted docking complex and to establish potential additional ligand-protein interactions resulting from the structural refinement of the complex, thus obtaining more accurate binding energy predictions. After refinement of the complex, the free energy of binding of the ligand is calculated with the MM-PBSA and MM-GBSA methods. These operations are implemented with the use of AMBER modules (Case et al., 2018). Further details about the BEAR tool are described in Rastelli et al. (2009).

BENCHMARKING STUDIES

The post-docking tool BEAR has been extensively validated in various test cases. First of all, the MM/MD protocol described above was investigated on a series of aldose reductase inhibitors with notable chemical diversity. Remarkably, the calculated free energies of binding after refinement of ligand-protein complexes resulted to be highly correlated with experimental affinities. This study demonstrated that different classes of aldose reductase inhibitors could be accurately rescored with our procedure (Ferrari et al., 2007). Extensive validations were also made on *Plasmodium falciparium* dihydrofolate reductase (PfDHFR). These simulations aimed at evaluating the performance of BEAR in virtual screening settings of different size and complexity. Firstly, BEAR performed well in discriminating 14 known inhibitors of PfDHFR from the 1,720 compounds included in the National Cancer Institute diversity database (Rastelli et al., 2009). The achieved performances were clearly superior to those of AutoDock (Morris et al., 1998), demonstrating that rescoring

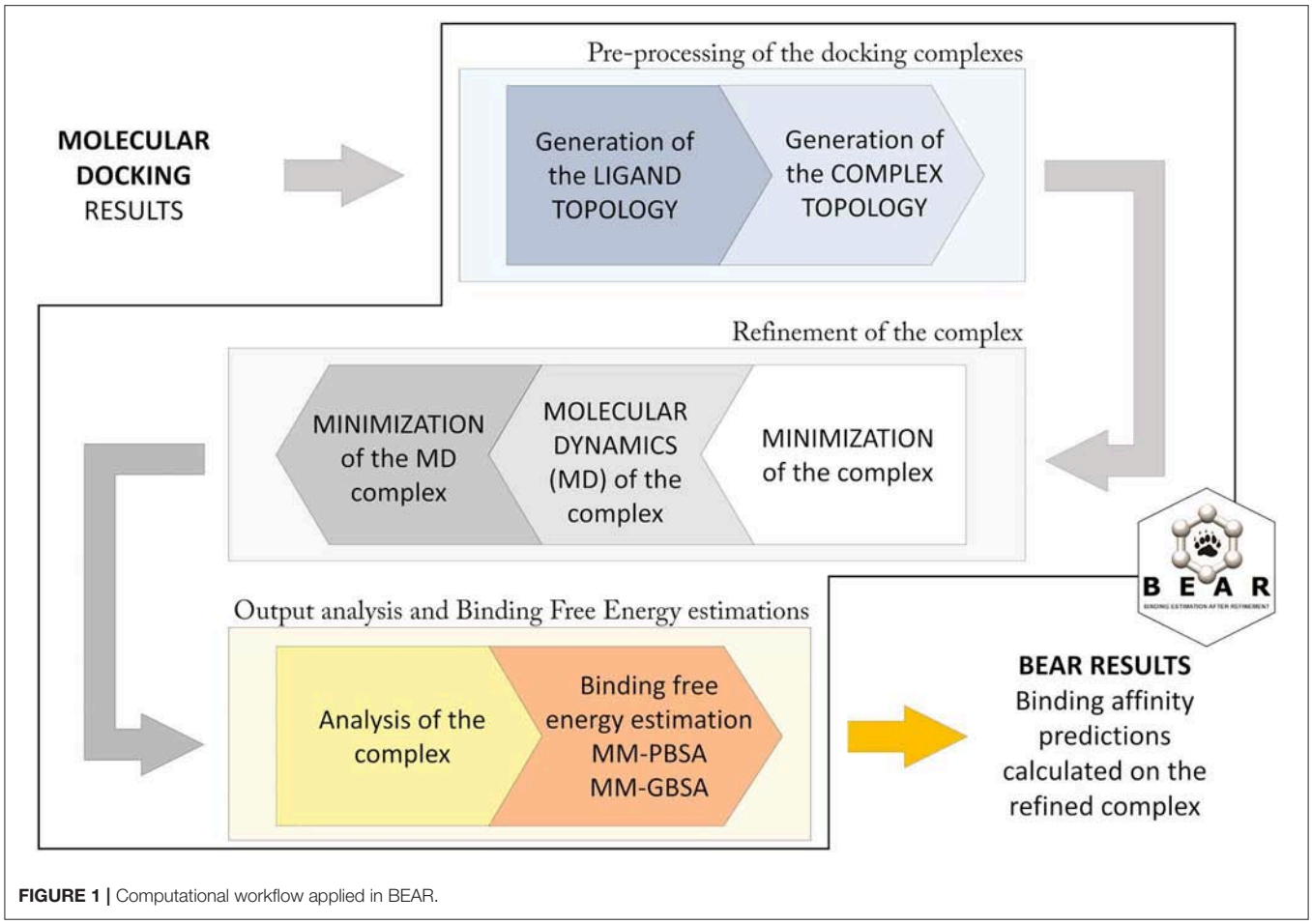


FIGURE 1 | Computational workflow applied in BEAR.

of the predicted docking poses with BEAR heavily improved SBVS results. In a second experiment, enrichment factors (EFs) obtained with BEAR were evaluated by seeding 201 known inhibitors with 7,150 decoys as contained in the DHFR data set of the Directory of Useful Decoys database (Mysinger et al., 2012). Moreover, the same set of ligands was also seeded into the 1.5 million compounds belonging to the lead-like subset of the ZINC database (Irwin et al., 2012), this latter benchmark reflecting a typical virtual screening setting. In both cases, BEAR refinement and rescoring yielded significantly higher EFs compared to docking (Degliesposti et al., 2011). This was also an opportunity for fine-tuning the BEAR parameters, and thus achieving good performances at reasonable computational costs.

The BEAR performance was also assessed on biological targets characterized by flexible binding sites and/or containing water molecules in the binding pocket. Such targets are particularly challenging for SBVS (Elokely and Doerksen, 2013). In fact, certain ligand chemotypes could fit with favorable scores into certain protein conformations but not in others, thus hampering their identification in a virtual screening. To evaluate whether docking into multiple protein conformations (ensemble docking) instead of using a single representative structure would improve BEAR predictions for "difficult" targets (Sgobba et al., 2012), we investigated targets of different families (adenosine deaminase, factor Xa, estrogen receptor, thymidine kinase, aldose reductase, and enoyl ACP reductase). Interestingly, a comparative analysis of the EFs obtained for different proteins and multiple protein conformations revealed that the application of BEAR was able in several cases to yield higher EFs compared to docking. However, in challenging targets such as adenosine deaminase and enoyl ACP reductase, all scoring functions failed in yielding high EFs. This effect was attributed to difficulties in predicting correct ligand binding modes in these two targets. In particular, when the docked pose was completely wrong, for example head-to-tail with respect to the correct binding mode, the MM/MD refinement stage was not enough to turn the binding mode into the correct one. Therefore, the advantage of using MM-PBSA and MM-GBSA in prioritizing active compounds is dependent on the obtainment of correct binding modes, which makes the refinement and rescoring procedures intimately connected.

More recently, BEAR was also applied to screen ligands of G-protein coupled receptors (GPCRs) with known crystal structure, namely β_2-adrenergic (β_2), adenosine A_{2A} (A_{2A}), dopamine D_3 (D_3), and histamine H_1 (H_1) receptors (Anighoro and Rastelli, 2013). Results were analyzed in terms of the ability to recognize known antagonists from decoys, as well as to predict correct binding modes. In all cases except for A_{2A}, significant or dramatic improvements of EFs were

obtained after the application of BEAR. A_{2A} was challenging because antagonists participated to an extended water-mediated hydrogen bond network. Interestingly, explicit consideration of a suitable number of these structural waters significantly improved the predictions. This finding is in line with the fact that MM-PB(GB)SA calculations do not explicitly consider water molecules mediating ligand-protein interactions, and binding mode predictions heavily depend on the presence of bridging water molecules participating to hydrogen bond networks. For all GPCRs, a more accurate account of desolvation effects, such as the one performed by MM-PBSA, is important to accurately predict the affinity of the protonated biogenic amines. We also found that five known H_1 and D_3 receptor antagonists were top-scored and ranked well in each of the two target screenings, prospecting for the first time the utility of post-docking tools in multi-target drug design (Anighoro et al., 2014). Indeed, as the free energy of binding calculated by BEAR allows to more accurately predicting the affinity of ligands for their target(s) with a reasonable computer time investment, we envision that *in silico* strategies embedding this tool can be useful to allow the identification of ligands with the desired multi-target profiles.

PROSPECTIVE VALIDATIONS

The BEAR workflow has also been implemented in the computing GRID infrastructure EGEE, as part of the WISDOM (Wide *in silico* DOcking on Malaria) initiative against malaria (Kasam et al., 2009). Then, it was deployed to perform virtual screenings against antimalarial drug targets. One massive data challenge was performed on Plasmepsin II, an aspartic protease involved in the metabolism of *P. falciparum* (Degliesposti et al., 2009). In this work, BEAR was used to refine and rescore the 5,000 top-scoring compounds docked with FlexX (Rarey et al., 1996). Then, the final step of candidates' selection was performed on the top 200 compounds resulting from both MM-PBSA and MM-GBSA ranked lists. Interestingly, an analysis of the BEAR ranked lists, together with an inspection of the protein-ligand complexes and a similarity-based clustering of the ligands allowed selecting 30 compounds belonging to 5 different chemotypes as potential Plasmepsin II inhibitors. Remarkably, 26 of them were active, resulting in an impressive hit rate of 87%, and some of the compounds displayed nanomolar inhibitory activity.

More recently, BEAR was successfully applied in a virtual screening campaign that allowed the identification of the *first-in-class* allosteric inhibitors of CDK2 (Rastelli et al., 2014). In this work, around 600.000 commercially available compounds were screened against a crystal structure of CDK2 with an open type III allosteric pocket, by using AutoDock for docking and BEAR for post-docking analyses. The adopted virtual screening protocol led to the identification of 7 allosteric ligands of CDK2, providing a hit rate of 20%. Interestingly, the most potent compound was able to selectively inhibit CDK2-mediated Retinoblastoma phosphorylation, confirming that its mechanism of action is fully compatible with a selective inhibition of CDK2 phosphorylation in cells. Moreover, some of these ligands inhibited the proliferation of MDA-MB231 and ZR-75-1 breast cancer cells with IC_{50} values in the low micromolar range (Rastelli et al., 2014).

FINAL REMARKS

Although many progresses have been made in molecular docking, limitations deriving from the use of rigid protein conformations and of approximate scoring functions often impair virtual screening results. Therefore, docking results require careful evaluation and further post-docking analyses. BEAR is a post-processing tool that performs binding free energy estimations after MM and MD refinement of docking complexes. Our previous studies demonstrated that BEAR performed well in a number of benchmarking investigations, as well as in discovering biologically active hits in different prospective virtual screening campaigns. Moreover, as it not computationally demanding as other free energy-based methods, it constitutes a reasonable compromise to obtain accurate rescoring of ligands at reasonable computational costs. One might argue that the application of more accurate workflows would require longer computing times with respect to docking. This is especially true considering the recent contributions provided by high performance computing systems to molecular docking, which enable the screening of millions of compounds in a reasonable time (Perez-Sanchez and Wenzel, 2011; Guerrero et al., 2012; Dong et al., 2018). However, future advances in hardware and software will help circumventing such limitation (De Vivo et al., 2016; Wang et al., 2018). Moreover, further advances in our ability to correctly estimate entropies of binding, which are usually not considered in the calculations, will certainly improve post-docking tools, and binding free energy predictions in general. The implementation of enhanced sampling MD protocols in post-docking protocols is another possibility that may enable a more efficient sampling of ligand-protein complexes. Because free energy predictions are heavily dependent on correct binding modes, this may have dramatic consequences on our ability to predict active ligands in virtual screenings. Another interesting question is how to further increase hit rates while enabling post-docking tools to identify significantly vs. moderately active hits. This is an important aspect that would make the subsequent hit-to-lead optimization much easier. Exponential consensus ranking approaches such as the one developed by Palacio-Rodríguez et al. (2019) could be of help, for example to favorably exploit both MM-PBSA and MM-GBSA ranked lists, which generally differ.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

GR conceived and wrote the study. LP contributed in writing and editing the manuscript.

REFERENCES

Anighoro, A., Bajorath, J. J., and Rastelli, G. (2014). Polypharmacology: challenges and opportunities in drug discovery. *J. Med. Chem.* 57, 7874–7887. doi: 10.1021/jm5006463

Anighoro, A., and Rastelli, G. (2013). Enrichment factor analyses on G-protein coupled receptors with known crystal structure. *J. Chem. Inf. Model.* 53, 739–743. doi: 10.1021/ci4000745

Brandsdal, B. O., Österberg, F., Almlöf, M., Feierberg, I., Luzhkov, V. B., and Åqvist, J. B. (2003). Free energy calculations and ligand binding in protein simulations. *Adv. Protein Chem.* 66, 123–158. doi: 10.1016/S0065-3233(03)66004-3

Caporuscio, F., and Rastelli, G. (2016). Exploiting computationally derived out-of-the-box protein conformations for drug design. *Future Med. Chem.* 8, 1887–1897. doi: 10.4155/fmc-2016-0098

Case, D. A., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., Cheatham, T. E. III., Cruzeiro, V. W. D., et al. (2018). *AMBER 2018*. University of California, San Francisco, CA. Available online at: http://www.ambermd.org/ (accessed April 15, 2019).

De Vivo, M., Masetti, M., Bottegoni, G., and Cavalli, A. (2016). Role of molecular dynamics and related methods in drug discovery. *J. Med. Chem.* 59, 4035–4061. doi: 10.1021/acs.jmedchem.5b01684

Degliesposti, G., Kasam, V., Da Costa, A., Kang, H. K., Kim, N., Kim, D. W., et al. (2009). Design and discovery of plasmepsin II inhibitors using an automated workflow on large-scale grids. *ChemMedChem* 4, 1164–1173. doi: 10.1002/cmdc.200900111

Degliesposti, G., Portioli, C., Parenti, M. D., and Rastelli, G. (2011). BEAR, a novel virtual screening methodology for drug discovery. *J. Biomol. Screen.* 16, 129–133. doi: 10.1177/1087057110388276

Dong, D., Xu, Z., Zhong, W., and Peng, S. (2018). Parallelization of molecular docking: a review. *Curr. Top. Med. Chem.* 18, 1015–1028. doi: 10.2174/1568026618666180821145215

Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., et al. (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* 24, 1999–2012. doi: 10.1002/jcc.10349

Elokely, K. M., and Doerksen, R. J. (2013). Docking challenge: protein sampling and molecular docking performance. *J. Chem. Inf. Model.* 53, 1934–1945. doi: 10.1021/ci400040d

Ferrari, A. M., Degliesposti, G., Sgobba, M., and Rastelli, G. (2007). Validation of an automated procedure for the prediction of relative free energies of binding on a set of aldose reductase inhibitors. *Bioorg. Med. Chem.* 15, 7865–7877. doi: 10.1016/j.bmc.2007.08.019

Ferri, N., Corsini, A., Bottino, P., Clerici, F., and Contini, A. (2009). Virtual screening approach for the identification of New Rac1 inhibitors. *J. Med. Chem.* 52, 4087–4090. doi: 10.1021/jm8015987

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* 47, 1739–1749. doi: 10.1021/jm0306430

Genheden, S., and Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* 10, 449–461. doi: 10.1517/17460441.2015.1032936

Gohlke, H., Kiel, C., and Case, D. A. (2003). Insights into protein–protein binding by binding free energy calculation and free energy decomposition for the Ras–Raf and Ras–RalGDS complexes. *J. Mol. Biol.* 330, 891–913. doi: 10.1016/S0022-2836(03)00610-7

Guerrero, G. D., Perez-S'nchez, H. E., Cecilia, J. M., and Garcia, J. M. (2012). "Parallelization of virtual screening in drug discovery on massively parallel architectures," in *2012 20th Euromicro International Conference on Parallel, Distributed and Network-based Processing* (Garching), 588–595. doi: 10.1109/PDP.2012.26

Hou, T., Wang, J., Li, Y., and Wang, W. (2011). Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. *J. Comput. Chem.* 32, 866–877. doi: 10.1002/jcc.21666

Hou, T., and Yu, R. (2007). Molecular dynamics and free energy studies on the wild-type and double mutant HIV-1 protease complexed with amprenavir and two amprenavir-related inhibitors: mechanism for binding and drug resistance. *J. Med. Chem.* 50, 1177–1188. doi: 10.1021/jm0609162

Huang, S.-Y., and Zou, X. (2010). Advances and challenges in protein-ligand docking. *Int. J. Mol. Sci.* 11, 3016–3034. doi: 10.3390/ijms11083016

Irwin, J. J., and Shoichet, B. K. (2016). Docking screens for novel ligands conferring new biology. *J. Med. Chem.* 59, 4103–4120. doi: 10.1021/acs.jmedchem.5b02008

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. (2012). ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* 52, 1757–1768. doi: 10.1021/ci3001277

Jorgensen, W. L., and Thomas, L. L. (2008). Perspective on free-energy perturbation calculations for chemical equilibria. *J. Chem. Theory Comput.* 4, 869–876. doi: 10.1021/ct800011m

Kar, S., and Roy, K. (2013). How far can virtual screening take us in drug discovery? *Expert Opin. Drug Discov.* 8, 245–261. doi: 10.1517/17460441.2013.761204

Kasam, V., Salzemann, J., Botha, M., Dacosta, A., Degliesposti, G., Isea, R., et al. (2009). WISDOM-II: screening against multiple targets implicated in malaria using computational grid infrastructures. *Malar. J.* 8:88. doi: 10.1186/1475-2875-8-88

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* 3, 935–949. doi: 10.1038/nrd1549

Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., et al. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* 33, 889–897. doi: 10.1021/ar000033j

Kuhn, B., Gerber, P., Schulz-Gasch, T., and Stahl, M. (2005). Validation and use of the MM-PBSA approach for drug discovery. *J. Med. Chem.* 48, 4040–4048. doi: 10.1021/jm049081q

Lengauer, T., and Rarey, M. (1996). Computational methods for biomolecular docking. *Curr. Opin. Struct. Biol.* 6, 402–406. doi: 10.1016/S0959-440X(96)80061-3

Limongelli, V., Bonomi, M., and Parrinello, M. (2013). Funnel metadynamics as accurate binding free-energy method. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6358–6363. doi: 10.1073/pnas.1303186110

Lyne, P. D., Lamb, M. L., and Saeh, J. C. (2006). Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring. *J. Med. Chem.* 49, 4805–4808. doi: 10.1021/jm060522a

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., et al. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* 19, 1639–1662.

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55, 6582–6594. doi: 10.1021/jm300687e

Nabuurs, S. B., Wagener, M., and de Vlieg, J. (2007). A flexible approach to induced fit docking. *J. Med. Chem.* 50, 6507–6518. doi: 10.1021/jm070593p

Palacio-Rodríguez, K., Lans, I., Cavasotto, C. N., and Cossio, P. (2019). Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. *Sci. Rep.* 9:5142. doi: 10.1038/s41598-019-41594-3

Parenti, M.D., and Rastelli, G. (2012). Advances and applications of binding affinity prediction methods in drug discovery. *Biotechnol. Adv.* 30, 244–250. doi: 10.1016/j.biotechadv.2011.08.003

Perez-Sanchez, H., and Wenzel, W. (2011). Optimization methods for virtual screening on novel computational architectures. *Curr. Comput. Aided. Drug Des.* 7, 44–52. doi: 10.2174/157340911793743565

Pu, C., Yan, G., Shi, J., and Li, R. (2017). Assessing the performance of docking scoring function, FEP, MM-GBSA, and QM/MM-GBSA approaches on a series of PLK1 inhibitors. *Medchemcomm* 8, 1452–1458. doi: 10.1039/c7md00184c

Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* 261, 470–489. doi: 10.1006/jmbi.1996.0477

Rastelli, G. (2013). Emerging topics in structure-based virtual screening. *Pharm. Res.* 30, 1458–1463. doi: 10.1007/s11095-013-1012-9

Rastelli, G., Anighoro, A., Chripkova, M., Carrassa, L., and Broggini, M. (2014). Structure-based discovery of the first allosteric inhibitors of cyclin-dependent kinase 2. *Cell Cycle* 13, 2296–2305. doi: 10.4161/cc.29295

Rastelli, G., Degliesposti, G., Del Rio, A., and Sgobba, M. (2009). Binding estimation after refinement, a new automated procedure for the refinement and rescoring of docked ligands in virtual screening. *Chem. Biol. Drug Des.* 73, 283–286. doi: 10.1111/j.1747-0285.2009.00780.x

Rastelli, G., Del Rio, A., Degliesposti, G., Sgobba, M., Rio, A., Del Degliesposti, G., et al. (2010). Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *J. Comput. Chem.* 31, 797–810. doi: 10.1002/jcc.21372

Sánchez-Linares, I., Pérez-Sánchez, H., Cecilia, J. M., and García, J. M. (2012). High-throughput parallel blind virtual screening using BINDSURF. *BMC Bioinformatics* 13:S13. doi: 10.1186/1471-2105-13-S14-S13

Sgobba, M., Caporuscio, F., Anighoro, A., Portioli, C., and Rastelli, G. (2012). Application of a post-docking procedure based on MM-PBSA and MM-GBSA on single and multiple protein conformations. *Eur. J. Med. Chem.* 58, 431–440. doi: 10.1016/j.ejmech.2012.10.024

Sherman, W., Beard, H. S., and Farid, R. (2006). Use of an induced fit receptor structure in virtual screening. *Chem. Biol. Drug Des.* 67, 83–84. doi: 10.1111/j.1747-0285.2005.00327.x

Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E. W. (2014). Computational methods in drug discovery. *Pharmacol. Rev.* 66, 334–395. doi: 10.1124/pr.112.007336

Sun, H., Duan, L., Chen, F., Liu, H., Wang, Z., Pan, P., et al. (2018). Assessing the performance of MM/PBSA and MM/GBSA methods. 7. Entropy effects on the performance of end-point binding free energy calculation approaches. *Phys. Chem. Chem. Phys.* 20, 14450–14460. doi: 10.1039/C7CP07623A

Sun, H., Li, Y., Shen, M., Tian, S., Xu, L., Pan, P., et al. (2014). Assessing the performance of MM/PBSA and MM/GBSA methods. 5. Improved docking performance using high solute dielectric constant MM/GBSA and MM/PBSA rescoring. *Phys. Chem. Chem. Phys.* 16, 22035–22045. doi: 10.1039/C4CP03179B

Wang, C., Greene, D., Xiao, L., Qi, R., and Luo, R. (2018). Recent developments and applications of the MMPBSA method. *Front. Mol. Biosci.* 4:87. doi: 10.3389/fmolb.2017.00087

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., Case, D. A., Wang, J., et al. (2004). Development and testing of a general amber force field. *J. Comput. Chem.* 25, 1157–1174. doi: 10.1002/jcc.20035

Xu, L., Sun, H., Li, Y., Wang, J., and Hou, T. (2013). Assessing the performance of MM/PBSA and MM/GBSA methods. 3. The impact of force fields and ligand charge models. *J. Phys. Chem. B* 117, 8408–8421. doi: 10.1021/jp404160y

Yang, Y., Liu, H., and Yao, X. (2012). Understanding the molecular basis of MK2–p38α signaling complex assembly: insights into protein–protein interaction by molecular dynamics and free energy studies. *Mol. Biosyst.* 8, 2106–2118. doi: 10.1039/C2MB25042J

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Rastelli and Pinzi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 16 July 2019
doi: 10.3389/fchem.2019.00509

Predicting Synergism of Cancer Drug Combinations Using NCI-ALMANAC Data

Pavel Sidorov[1], *Stefan Naulaerts*[1,2], *Jérémy Ariey-Bonnet*[1], *Eddy Pasquier*[1] *and Pedro J. Ballester*[1]*

[1] CRCM, INSERM, Cancer Research Center of Marseille, Institut Paoli-Calmettes, Aix-Marseille Univ, CNRS, Marseille, France, [2] Department of Tumor Immunology, Institut de Duve, Bruxelles, Belgium

Drug combinations are of great interest for cancer treatment. Unfortunately, the discovery of synergistic combinations by purely experimental means is only feasible on small sets of drugs. *In silico* modeling methods can substantially widen this search by providing tools able to predict which of all possible combinations in a large compound library are synergistic. Here we investigate to which extent drug combination synergy can be predicted by exploiting the largest available dataset to date (NCI-ALMANAC, with over 290,000 synergy determinations). Each cell line is modeled using primarily two machine learning techniques, Random Forest (RF) and Extreme Gradient Boosting (XGBoost), on the datasets provided by NCI-ALMANAC. This large-scale predictive modeling study comprises more than 5,000 pair-wise drug combinations, 60 cell lines, 4 types of models, and 5 types of chemical features. The application of a powerful, yet uncommonly used, RF-specific technique for reliability prediction is also investigated. The evaluation of these models shows that it is possible to predict the synergy of unseen drug combinations with high accuracy (Pearson correlations between 0.43 and 0.86 depending on the considered cell line, with XGBoost providing slightly better predictions than RF). We have also found that restricting to the most reliable synergy predictions results in at least 2-fold error decrease with respect to employing the best learning algorithm without any reliability estimation. Alkylating agents, tyrosine kinase inhibitors and topoisomerase inhibitors are the drugs whose synergy with other partner drugs are better predicted by the models. Despite its leading size, NCI-ALMANAC comprises an extremely small part of all conceivable combinations. Given their accuracy and reliability estimation, the developed models should drastically reduce the number of required *in vitro* tests by predicting *in silico* which of the considered combinations are likely to be synergistic.

Keywords: chemoinformatics, drug synergy, machine learning, QSAR (qualitative structure-activity relationships), predictive (QSPR) models

OPEN ACCESS

Edited by:
Jose L. Medina-Franco, National Autonomous University of Mexico, Mexico

Reviewed by:
Oscar Mendez Lucio, Bayer, France
Monica Campillos, Helmholtz Center Munich, Germany

***Correspondence:**
Pedro J. Ballester
pedro.ballester@inserm.fr

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *17 May 2019*
Accepted: *02 July 2019*
Published: *16 July 2019*

Citation:
Sidorov P, Naulaerts S, Ariey-Bonnet J, Pasquier E and Ballester PJ (2019) Predicting Synergism of Cancer Drug Combinations Using NCI-ALMANAC Data. Front. Chem. 7:509. doi: 10.3389/fchem.2019.00509

INTRODUCTION

Drug combinations are a well-established form of cancer treatment (Bayat Mokhtari et al., 2017). Administering more than one drug can provide many benefits: higher efficacy, lower toxicity, and at least delayed onset of acquired drug resistance (Sugahara et al., 2010; Holohan et al., 2013; Crystal et al., 2014). Serendipitous discovery in the clinic has been a traditional source of effective drug

combinations (Zoli et al., 2001; Kurtz et al., 2015). Yet systematic large-scale efforts to identify them have only recently been pursued, with a growing number of preclinical experimental efforts to identify synergistic combinations (Zoli et al., 2001; Budman et al., 2012; Lieu et al., 2013; Kashif et al., 2015; Yu et al., 2015; Kischkel et al., 2017) being reported in literature. The sheer number of available and possible drug-like molecules (Polishchuk et al., 2013) and an exponential number of their combinations, however, make the process of finding new therapeutic combinations by purely experimental means highly inefficient.

An efficient way of discovering molecules with previously unknown activity on a given target is using *in silico* prediction methods. Quantitative Structure-Activity Relationship (QSAR) models establish a mathematical relationship between the chemical structure of a molecule, encoded as a set of structural and/or physico-chemical features (descriptors), and its biological activity on a target. Such methods have been successfully used in a wide variety of pharmacology and drug design projects (Cherkasov et al., 2014), including cancer research (Chen et al., 2007; Mullen et al., 2011; Ali and Aittokallio, 2018). QSAR models are traditionally built using simple linear models (Sabet et al., 2010; Pick et al., 2011; Speck-Planche et al., 2011, 2012) to predict the activity of individual molecules against a molecular target. In the last 15 years, non-linear machine learning methods, such as Neural Network (NN) (González-Díaz et al., 2007), Support Vector Machine (SVM) (Doucet et al., 2007) or Random Forest (RF) (Singh et al., 2015), have also been employed to build QSAR models. More recently, QSAR modeling has also achieved accurate prediction of compound activity on non-molecular targets such as cancer cell lines (Kumar et al., 2014).

To extend QSAR modeling beyond individual molecules, the set of features from each molecule in the combination must be integrated. Various ways exist to encode two or more molecules as a feature vector, e.g., SIRMS descriptors (Kuz'min et al., 2008) for properties of combinations or the CGR approach for chemical reactions (de Luca et al., 2012). Rigorous validation strategies for the resulting models have been developed too (Muratov et al., 2012). The most common representation of a drug pair is, however, the concatenation of features from both molecules (Bulusu et al., 2016). On the other hand, modeling drug combinations requires the quantification of their synergy. Several metrics exist to quantify synergy (Foucquier and Guedj, 2015) (e.g., Bliss independence Bliss, 1939, Loewe additivity Chou and Talalay, 1984, Highest single agent approach Greco et al., 1995 or Chou-Talalay Method Chou, 2010). These are implemented in various commercial and publicly available software kits for the analysis of combination data, e.g., Combenefit (Di Veroli et al., 2016), CompuSyn (http://www.combosyn.com) or CalcuSyn (http://www.biosoft.com/w/calcusyn.htm).

One major roadblock in drug synergy modeling has been the lack of homogeneous data (i.e., datasets generated with the same assay, experimental conditions and synergy quantification). This has been, however, alleviated by the recent availability of large datasets from High-Throughput Screening (HTS) of drug combinations on cancer cell lines. For instance, Merck has released an HTS synergy dataset (O'Neil et al., 2016), covering combinations of 38 drugs and their activity against 39 cancer cell lines (more than 20,000 measured synergies). This dataset has been used to build predictive regression and classification models using multiple machine learning methods (Preuer et al., 2018). AstraZeneca carried out a screening study, spanning 910 drug combinations over 85 cancer cell lines (over 11,000 measured synergy scores), which was subsequently used for a DREAM challenge (Li et al., 2018; Menden et al., 2019). Very recently, the largest publicly available cancer drug combination dataset has been provided by the US National Cancer Institute (NCI). This NCI-ALMANAC (Holbeck et al., 2017) tested over 5,000 combinations of 104 investigational and approved drugs, with synergies measured against 60 cancer cell lines, leading to more than 290,000 synergy scores (ComboScores).

NCI-ALMANAC datasets have recently been modeled to predict the best growth inhibition of a given drug combination—cell line tuple (Xia et al., 2018). However, the question remains of how well ComboScores can be predicted on each NCI-60 cell line, which is important given that ComboScore-based screening has led to the identification of novel synergistic drug combinations *in vivo* (Holbeck et al., 2017). Here we present a large-scale study addressing this question. We build an individual model for each cell line using the popular RF algorithm (Breiman, 2001). We also build a second model per cell line using XGBoost (XGB for short) (Chen and Guestrin, 2016), a recent machine learning method that has helped to win numerous Kaggle competitions (Chen and Guestrin, 2016) as well as to generate highly predictive QSAR models (Sheridan et al., 2016). We validate these models for commonly-encountered prediction scenarios: e.g., unseen drug combination or unseen drug partner. We also introduce and validate reliability estimation techniques to further improve prediction of drug combination synergy. Lastly, we assess the suitability of NCI-ALMANAC datasets for predictive modeling depending on the screening center where they were generated.

METHODS

Data

NCI-ALMANAC is the largest-to-date phenotypic drug combination HTS. It contains the synergy measurements of pairwise combinations of 104 FDA approved drugs on the 60 cancer cell lines forming the NCI-60 panel (Shoemaker, 2006). The drugs include a wide array of small organic compound families, as well as several inorganic molecules (cisplatin and related platinum-organic compounds, arsenic trioxide). A similarity clustering dendrogram (**Figure 1**) shows the high diversity of the drugs in NCI-ALMANAC. Indeed, only 3 clusters comprising 8 drugs are formed with a Tanimoto score threshold of 0.8 (Vinblastine with Vincristine, Sirolimus and Everolimus, and Daunorubicin-Doxorubicin-Idarubicin-Epirubicin clusters), while the remaining 96 drugs have smaller similarity among them.

NCI-ALMANAC aggregates synergy data from three screening centers: NCI's Frederick National laboratory for Cancer Research (screening center code 1A, 11,259 synergy determinations), SRI International (FF, 146,147 determinations),

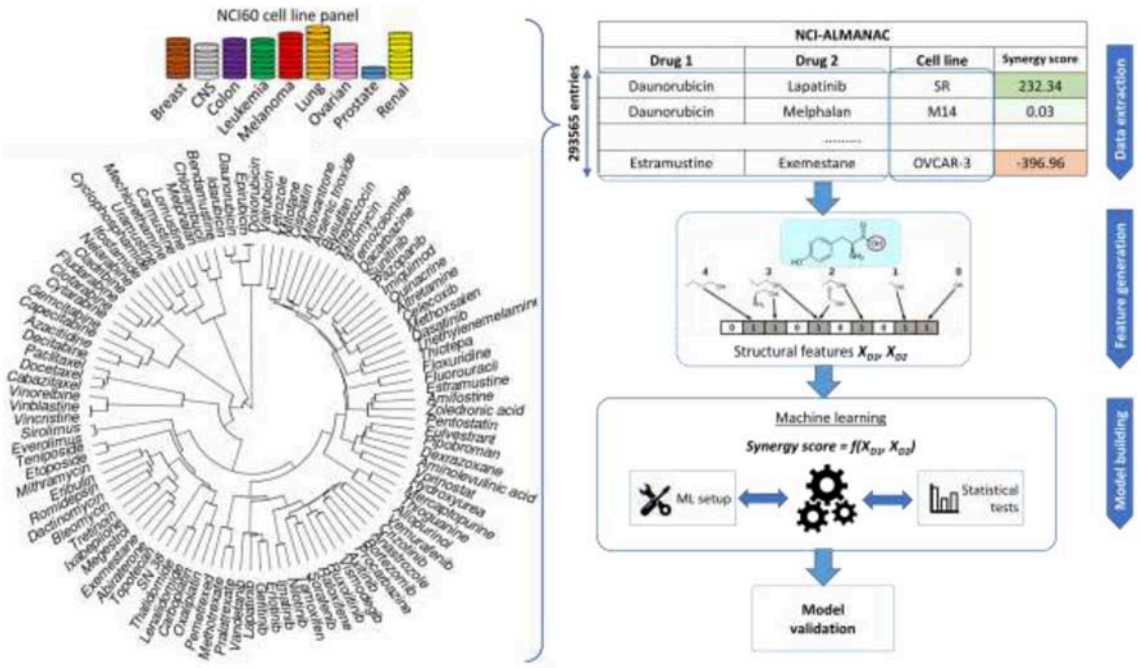


FIGURE 1 | Sketch of the workflow for drug combination modeling. Training data comes from NCI-ALMANAC, which comprises over 290,000 synergy measurements from pairs of 104 drugs tested on the 60 cell lines. Structural and physico-chemical features are calculated for each drug from its chemical structure. Similarity clustering diagram for 104 NCI-ALMANAC drugs is on the left. Each drug is characterized by MFPC features complemented with physico-chemical features, using the Tanimoto score on these features as the similarity metric. Hierarchical agglomerative clustering was carried out (ward.D2 algorithm in the R hclust function). Closely related compounds form tight clusters (e.g., doxorubicin and its analogs, analogs of paclitaxel, etc). By contrast, naturally inorganic compounds such as cisplatin and arsenic trioxide appear as outliers (the highest similarity coefficient to other drugs being 0.156 and 0.125, respectively). The concatenated vectors of the two drugs are the features utilized to build and test predictive models with machine learning techniques. The predictive accuracies of the models are determined by multiple cross-validation experiments.

and University of Pittsburgh (FG, 136,129 determinations). The synergy of drug pairs is measured in these screening centers against the NCI-60 panel, which includes cell lines from nine cancer types: leukemia, melanoma, non-small-cell lung, colon, central nervous system, ovarian, renal, prostate, and breast. In total, synergy is measured for 293,565 drug combination—cell line tuples, which represents a matrix completeness of 91.35%. Each center follows its own protocol, and some drugs are absent from the combination pool depending on the screening center. Since there is no overlap between drug combination—cell line tuples between the three centers, it is not possible to estimate inter-center batch effects, and therefore we must use data from different screening centers separately.

The combination benefit is quantified in NCI-ALMANAC by the so-called ComboScore (a modified version of the Bliss independence model). From the entire dose-response matrix of the considered drug combination and cell line tuple, the gain (or loss) of the effect achieved by the combination over the theoretically expected value if the effect was additive is calculated. Positive values of ComboScore indicate a synergistic effect of the combination, whereas the negative correspond to an antagonistic effect (those purely additive obtain a zero ComboScore).

Further description of NCI-ALMANAC data is available at **Supplementary Information**.

Features

For the use in machine learning, the structures of compounds must be encoded as vectors of numerical features known in chemoinformatics as molecular descriptors (Todeschini and Consonni, 2000). Several types of chemical structure features have been considered in this work: (1) Morgan FingerPrints (MFP) are topological descriptors describing the connectivity of the molecular structure, which take values 0 or 1, depending on whether the pattern is present in the molecule or not (Rogers and Hahn, 2010). They have been calculated with RDKit library (Lamdrum, 2015) using the following parameters—length is 256 bits, radius is 2. (2) Morgan FingerPrint Counts (MFPC) are a non-binary version of MFP that takes integer values equal to the number of times the pattern is detected in the molecule (256 features per drug, also calculated with RDKit). (3) MACCS keys encode presence or absence of 166 predetermined substructural fragments as binary vectors (calculated with RDKit). (4) ISIDA fragments encode structure as a vector of numbers of occurrences of substructural fragments of given nature and topology in the molecule (Varnek et al., 2005), which are calculated with ISIDA/Fragmentor (Ruggiu et al., 2015). Only one type of fragments is considered here: sequences of atoms and bonds of length 2 to 6 (1,325 features per drug in total). (5) SIRMS fragments are the number of occurrences of 4-atom fragments

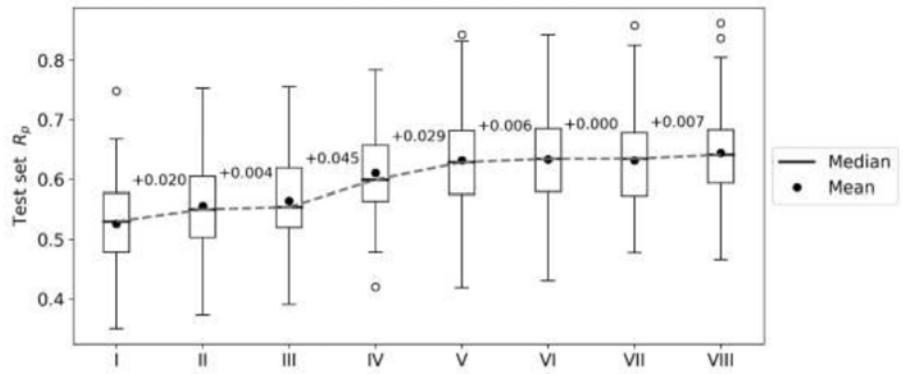


FIGURE 2 | Performance gain across cell lines for each introduced modeling choice during the exploratory analysis of FG data. Each boxplot represents the distribution of the cell line models' test set performances (R_p) at any given step. Analysis steps are carried out sequentially: I—RF, 1,000 trees with all n features tried to split a node, 80% training set, 20% test set, MACCS (Molecular ACCess System) keys as features; II—MFPC (Morgan fingerprint counts) are used as features instead; III—physico-chemical features are added for each drug; IV—training set rows are duplicated with the reverse order of drugs (data augmentation); V-−90% training set, 10% test set are used instead of the initial 80/20 partition; VI—RF with 250 trees with n/3 features tried to split a node; VII—XGB models with recommended settings; VIII—tuned XGB models. Note that I-V employ RF with same values for its hyperparameters (RF tuned in VI) and V–VIII use the same training and test sets. Modeling choices introducing the largest improvements are the choice of molecular features and the data augmentation strategies.

of varying topology in a molecule, including bonded and non-bonded atoms (Kuz'min et al., 2008). Calculated with SiRMS python library (github.com/DrrDom/sirms), it led to 1,454 features per drug. In addition to these sets, 7 physico-chemical features are calculated by RDkit: total polar surface area (TPSA), molecular weight, logP, number of aliphatic and aromatic rings, H-bond donors and acceptors.

Machine Learning (ML) Workflow

Models are built using two ML algorithms: Random Forest (RF) (Svetnik et al., 2003) and Extreme Gradient Boosting (XGBoost; XGB for short) (Sheridan et al., 2016). The entire modeling workflow is sketched in **Figure 1**. Further details about how ML models were built are available at the **Supplementary Information**.

Predictive Performance Metrics

To evaluate the performance of a model, the following metrics are calculated from observed y_{obs} and predicted y_{pred} ComboScore values:

Root Mean Squared Error (*RMSE)*

$$RMSE = \sqrt{\frac{\sum_N \left(y_{i,\,obs} - y_{i,\,pred}\right)^2}{N}}$$

Coefficient of determination (R^2) (Leach and Gillet, 2007)

$$R^2 = 1 - \frac{\sum_N \left(y_{i,\,obs} - y_{i,\,pred}\right)^2}{\sum_N \left(y_{i,\,obs} - \overline{y_{obs}}\right)^2} = 1 - \frac{RMSE^2}{Var\left(y_{obs}\right)};$$

$$\overline{y_{obs}} = \frac{1}{N}\sum_{i=1}^{N} y_{i,obs}$$

Pearson's correlation coefficient (R_p)

$$R_p = \frac{\sum_N (y_{i,\,obs} - \overline{y_{\,obs}})(y_{i,\,pred} - \overline{y_{\,pred}})}{\sqrt{\sum_N (y_{i,\,obs} - \overline{y_{\,obs}})^2}\sqrt{\sum_N (y_{i,\,pred} - \overline{y_{\,pred}})^2}}$$

Spearman's rank-order correlation coefficient (R_s)

$$R_s = R_p(\text{rank}\, y_{obs},\ \text{rank}\, y_{pred})$$

We use R_p between observed and predicted values of ComboScore of a dataset not used to train the model as a primary metric of its accuracy. For proper estimation of the generalization error, these metrics are always calculated here on a test set not used to train or select the model.

RESULTS

Exploratory Modeling of NCI-ALMANAC Data

First, we perform an exploratory modeling on the FG datasets in order to determine optimal settings for synergy prediction by assessing various types of features, data augmentation schemes and machine learning methods. The summary of performance improvements is shown on **Figure 2**. The best median R_p across cell lines for RF was obtained with 250 trees, a third of the features evaluated at each tree node, training data augmentation and MFPC fingerprints complemented by physico-chemical properties (256 and 7 features per drug, respectively). The gain of performance with RF is substantial: the median R_p increases from 0.530 (I) to 0.634 (VI).

XGB models are generated with the same features and data set partitions. Changing the machine learning algorithm from RF to XGB does not improve the median test set R_p, although both minimum and maximum R_p are higher with XGB (boxplots

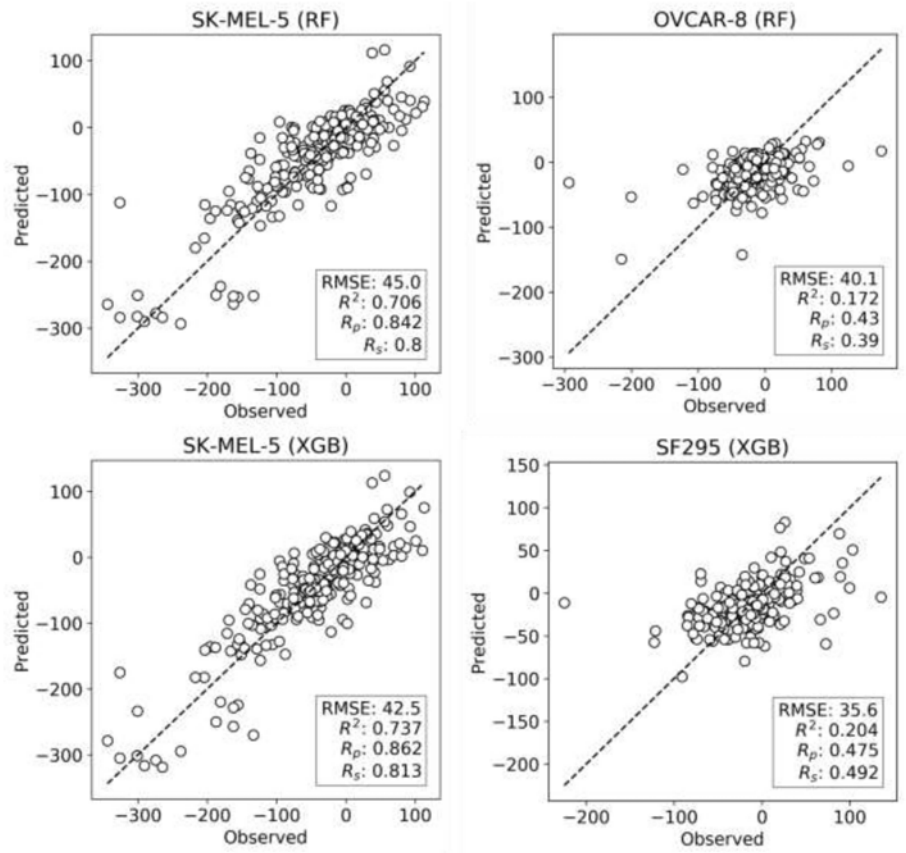


FIGURE 3 | Observed vs. predicted ComboScore for all the drug combinations in the test set. This is presented for the best- and worst-performing models with both ML methods, RF, and XGB (these models correspond hence to the extremes of **Figure 2**'s boxplots VI and VIII, respectively). On the left column, the best-performing cell line models from each method. On the right side, the worst-performing cell line models. All performance metrics are shown. Each point represents a drug combination in that test set.

VI and VII in **Figure 2**, respectively). After tuning of XGB hyperparameters per cell line, a small gain in overall performance is obtained: the median R_p of tuned XGB rises to 0.641 (boxplot VIII). In comparison, Y-randomization (Tropsha et al., 2003) tests using the same learning algorithm did only obtain a median R_p of −0.016 (−0.024 when using RF). **Figure 3** shows the degree of accuracy achieved by each algorithm for the best and the worst predicted cell line. The cell lines with the worst predictions (OVCAR-8 for RF and SF295 for XGB) have substantially smaller variance in observed ComboScore than those with the best predictions (SK-MEL-5 for both algorithms).

Estimating the Reliability of Drug Synergy Predictions

For prospective use of models, it is paramount to calculate not only predicted drug combination synergies, but also how reliable these predictions are (Mathea et al., 2016). With this purpose, we have applied a RF-specific reliability prediction approach, where the degree of agreement between the diverse trees in the forest serves as a reliability score. This is quantified here as the standard deviation (SD) of the RF tree predictions (250 per drug combination and cell line) and referred to as tree_SD. tree_SD has been pointed out as one of the most powerful metrics to assess the reliability of predictions in regression problems (Mathea et al., 2016). We thus assemble test subsets with the 25% most reliable ComboScore predictions per cell line (i.e., combinations with the 25% lowest tree_SD scores). Likewise, we assemble test subsets with 25% least reliable predictions per cell line.

Figure 4 presents the test set performances of each cell line model on the three scenarios: 25% most reliable predictions, all predictions regardless of estimated reliability and 25% least reliable predictions. The top and bottom 25% predictions in terms of reliability obtain the lowest and highest RMSE in every cell line, which demonstrates the accuracy and generality of tree_SD as a reliability score for drug synergy predictions. Test set RMSE varies greatly across cell lines, e.g., models built on leukemia cell lines obtain in general higher error. This, however, comes from the higher range of ComboScores observed in these cell lines. Indeed, the larger this range, the higher the range of predicted ComboScores is, which combined tend to make RMSE larger. Similar RMSE is only obtained on the K-562 leukemia cell line, which is consistent with the fact that it has the lowest range among leukemia cell lines and similar to that of other cancer types.

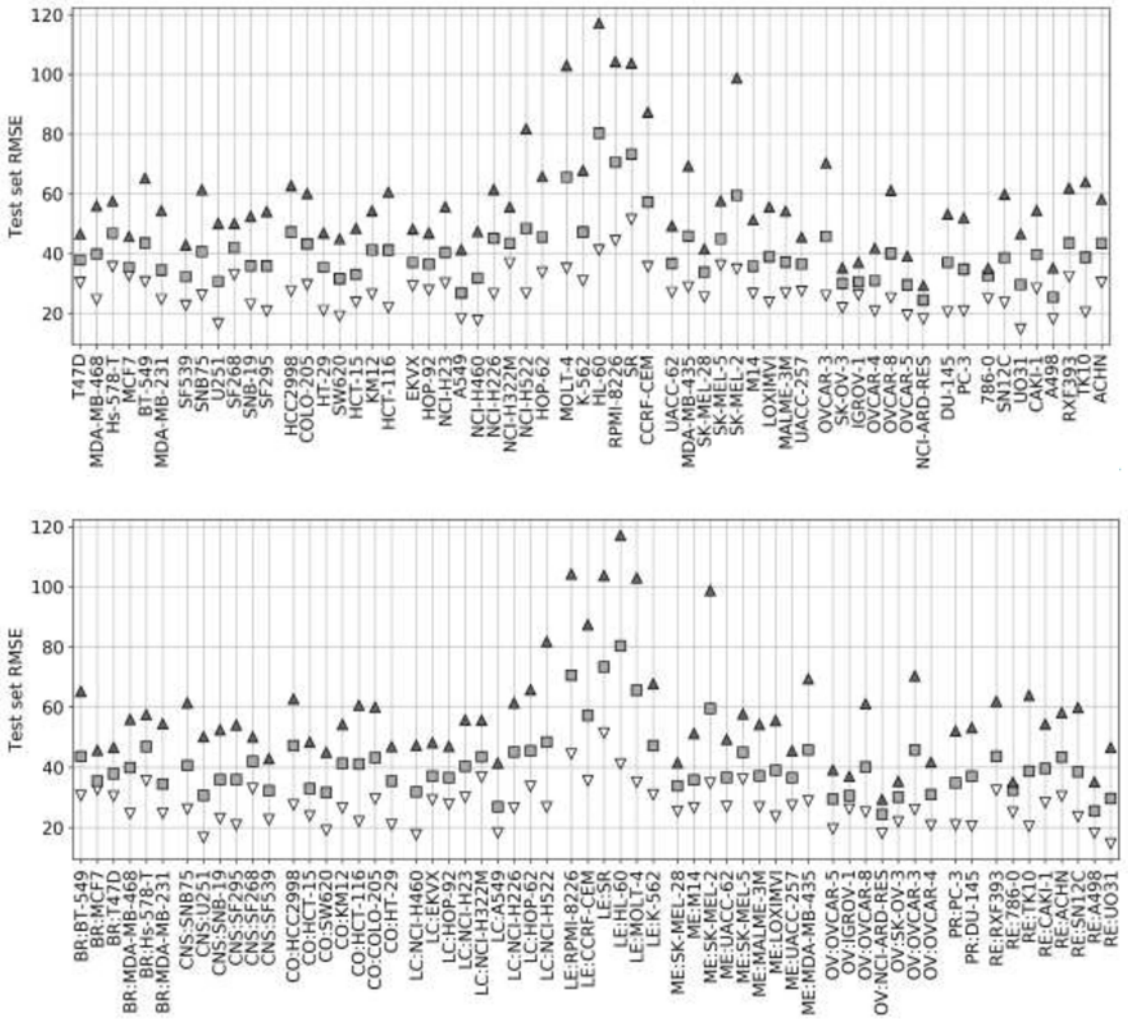


FIGURE 4 | Ten percent test set RMSE of RF cell line models trained on 90% of the FG data. Gray squares represent the model's RMSE on all the test combinations (RF predictions as usual). Black triangles mark the RMSE of the 25% least reliable (highest tree_SD) combos, whereas white inverted triangles correspond to the RMSE of the 25% most reliable (lowest tree_SD). In each cell line, the reliability score tracks test RMSE and hence it can be used to identify *a priori* the most accurate predictions. Each cell line name in the horizontal axis is preceded by its cancer type ID: breast (BR), central nervous system (CNS), colon (CO), non-small-cell lung (LC), leukemia (LE), melanoma (ME), ovarian (OV), prostate (PR), and renal (RE).

Reliability estimation is evaluated in terms of RMSE rather than R_p. While RMSE is not as intuitive as correlation, correlations may be misleading when comparing performances of models across test sets with distinct variances. **Figure 5** illustrates this issue with the test performances of HL-60 models, which benefit the most from reliability estimation. The test set with the most reliable combinations is predicted with half the RMSE of the entire test set (RMSE of 41 vs. 80) and a third of the least reliable combinations (RMSE of 41 vs. 117). This more accurate prediction can be visually observed too, but the other metrics (R^2, R_p, and R_s) do not capture this increase in accuracy due to substantially different ComboScore variance between the compared test sets. Importantly, RF with reliability prediction provides a much larger reduction in RMSE than that introduced by XGB (bottom right), both with respect to RF without reliability prediction (bottom left). These results strongly suggest that, in cases where it is not necessary to test all positive predictions (here synergistic drug combinations), selecting the most reliable predictions is more effective than using the most suitable ML algorithm.

Performance in Predicting Synergies With Drugs Not Included in NCI-ALMANAC

The random data splits that we have used so far may overestimate the model's performance in the case of drug combinations. This would be due to the presence of the two drugs in the combination in both training and test sets, albeit with other partners (Muratov et al., 2012). In order to assess to which extent this is the case, we also carry out Leave-One-Drug-Out (LODO)

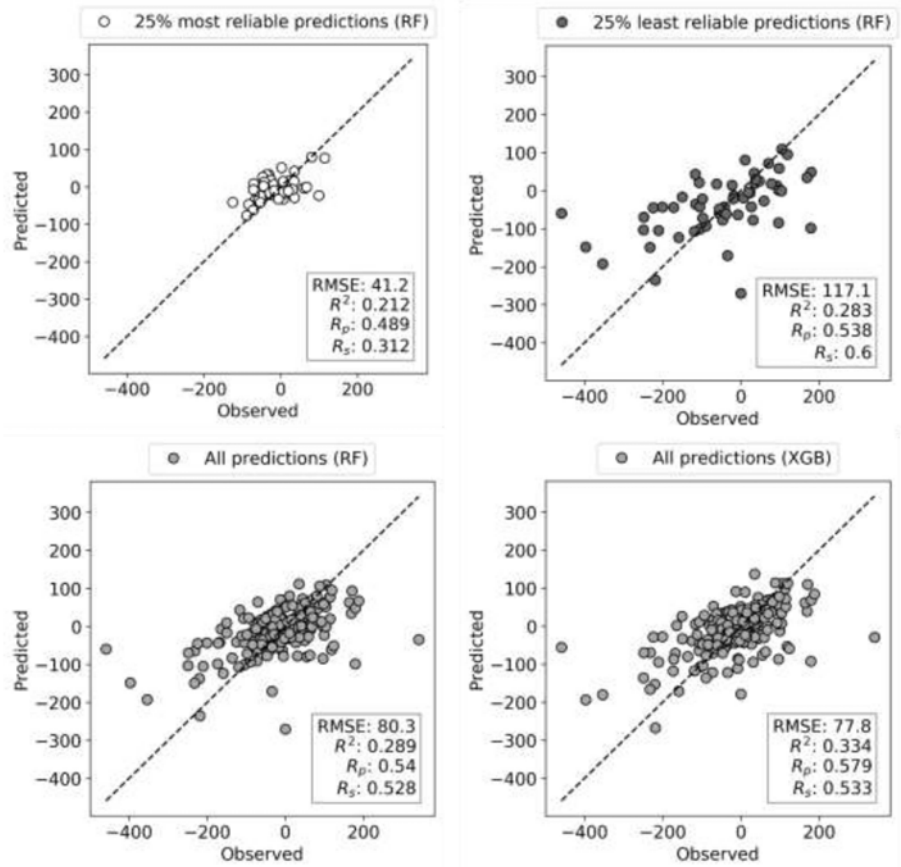


FIGURE 5 | Observed vs. Predicted ComboScore plots for HL-60 leukemia cell line test set (10% of data). Models are built on 90% of FG dataset corresponding to this cell line using RF and XGB methods, both tuned. Each circle is now a drug combination from the entire test set, with its shade of gray indicating one of the three scenarios (as in **Figure 4**). All performance metrics are displayed in each plot. The subset with the most reliable ComboScore RF predictions (top left plot) achieves half the RMSE of the entire test set (bottom right). Importantly, this is a much larger reduction in RMSE than that introduced by XGB (bottom right) with respect to RF (bottom left). Furthermore, the most reliable predicted ComboScores (top right) obtain a third of the RMSE of the least reliable predictions (top left).

cross-validation experiments for each cell line. In LODO cross-validation, every combination containing the considered left-out drug is placed in its test set, and the model is built on the remaining combinations tested on that cell line. Thus, there are as many folds as drugs in the dataset. In this way, the LODO cross-validation simulates the model's behavior when presented with a new chemical entity outside of the model's scope, as if it was not included in the dataset.

Figure 6 shows the outcome of LODO cross-validation for XGB per cell line. We henceforth use XGB with the recommended values for hyperparameters, as tuning them for each LODO cross-validation fold and cell line is prohibitive and would only provide marginal gains (see **Figure 2**). LODO results show that combinations associated with 75% of the left-out drugs can be predicted with an accuracy of at least $R_p = 0.3$ against any cell line. This accuracy raises to at least $R_p = 0.5$ for 50% of the left-out drugs. The latter is not much worse than the median R_p across cell lines when using 90/10 data partitions ($R_p = 0.641$ as shown in **Figure 2**'s boxplot VIII). k-fold cross-validation results are available for comparison in **Supplementary Figure 3**.

Figure 7 shows the analysis for LODO cross-validations in terms of RMSE. About 75% of models demonstrate at least moderate accuracy (RMSE < 50). The exceptions are mostly leukemia cell line models, which obtain higher RMSE due to having the highest variances in ComboScores among cancer types. An important result is that using RF models restricted to the most reliable predictions allows us to reduce the error of prediction further in every cell line (RMSE < 40), in full agreement with the findings from random 90/10 partitions (see **Figure 4**) and also outperforming the best models without reliability prediction.

Analyzing LODO results per left-out drug instead of per cell line reveals that synergy prediction is much worse for certain left-out drugs in each cancer type. LODO performance of each drug across cell lines is shown in **Supplementary Figure 4**. This figure shows that models for arsenic trioxide, highly dissimilar to other drugs, have the lowest performance across cell lines and partner drugs (median R_p of models concerning this drug is -0.28). Conversely, partners of tyrosine kinase inhibitors, well-represented in these datasets (e.g., Imatinib, Nilotinib or

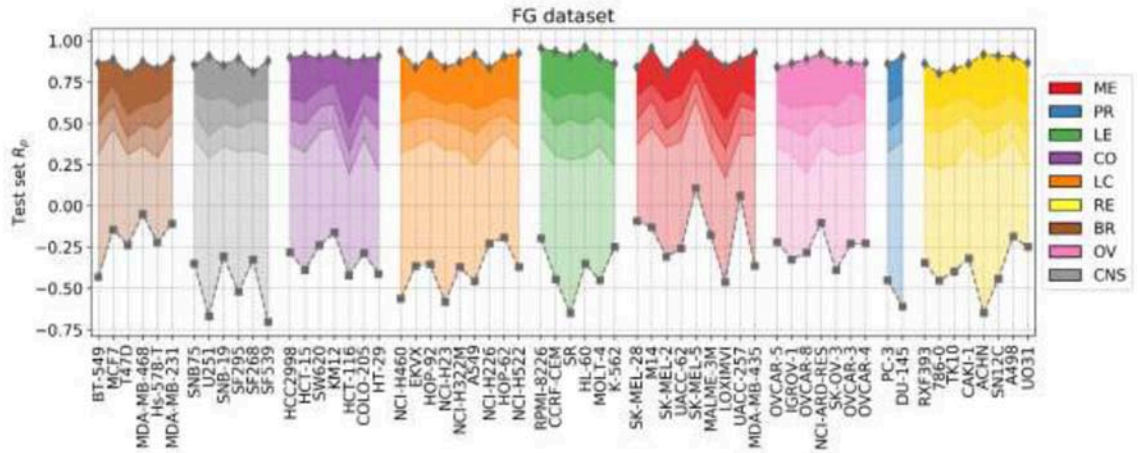


FIGURE 6 | LODO cross-validation results using XGB with the recommended values for their hyperparameters on the FG dataset. Distribution of models' performances is shown by cancer type (color code). Each colored zone represents 25% of models per cell line: from dense zone—top performing 25%; to light zone—bottom quartile. The method performs with at least moderate accuracy ($R_p > 0.30$) in 75% of left-out drugs (the top 3 quartiles) across cell lines. Left-out drugs within the top quartile, darkest shade among the four employed per cell line, are predicted with a R_p ranging from 0.471 (HCT-116) to 0.986 (SK-MEL-5). Although there are no large differences in how well different cancer types are predicted, left-out drugs on melanoma (ME, in red) and leukemia (LE, in green) cell lines obtain slightly higher average performance (median R_p of drug-out models for corresponding cell lines are 0.554 and 0.524, respectively).

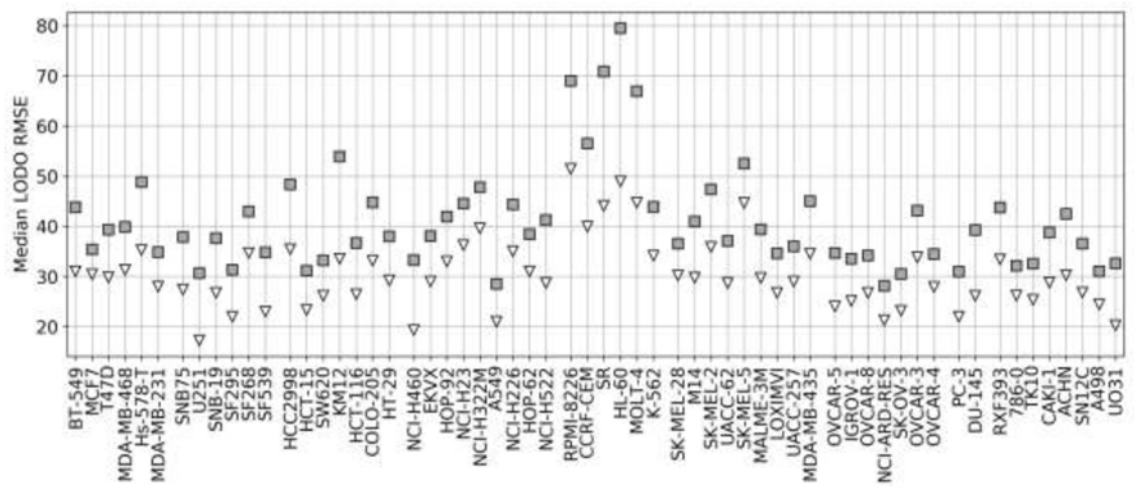


FIGURE 7 | Median RMSE in LODO cross-validation for XGB with the recommended values for their hyperparameters (gray squares) and RF top 25% most reliable predictions (white inverted triangles) for each cell line (grouped by cancer type). As the plot shows, combinations with one left-out drug can be predicted with at least moderate accuracy across cell lines (RMSE < 50 for XGB, RMSE < 40 for RF with reliability estimation; both being approximate thresholds).

Lapatinib), are predicted with high accuracy (e.g., models for imatinib have median $R_p = 0.82$). Topoisomerase inhibitors (Teniposide and Etoposide) are also among the best-predicted left-out drugs. These *in silico* models could be used to anticipate how the synergies of a drug in combination with its partner drugs would vary across NCI-60 cell lines. However, since high accuracy is only obtained on those left-out drugs well-represented in NCI-ALMANAC, such selectivity predictions should only be accurate for drugs with similar chemical structure to those in NCI-ALMANAC. As models predicting drug-induced cell line response have been shown to improve by integrating drug features with multi-omics cell features (Menden et al., 2013; Xia et al., 2018), we expect that predicting drug synergy across cell lines will also improve by following such multi-task learning approach on this closely related problem.

Comparing Predictive Models Built With Data From Different Screening Centers

So far we have exclusively employed data from the FG screening center, which represents about half of NCI-ALMANAC data. Practically all the remaining ComboScores come from the FF screening center and are also determined with a 3 × 3 grid of

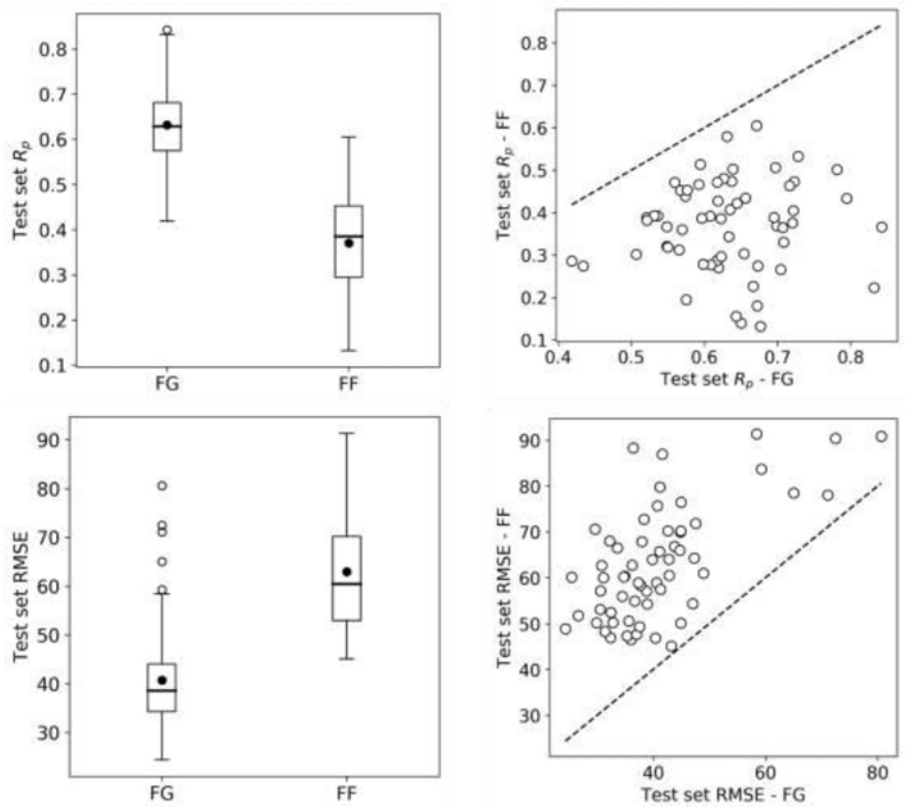


FIGURE 8 | RF model performance comparison for FG and FF datasets. Models are built following the final setup in the exploratory analysis (a 90/10 data partition is employed for each cell line); MFPC with physico-chemical features as well as data augmentation are also used). On the left, boxplots for cell line models test set R_p's (top row) and RMSE (bottom row) for both centers data. On the right, R_p (top row) and RMSE (bottom row) of models trained on FG dataset against models trained on FF dataset, each point shows the two model performances for the cell line. FF models obtain consistently lower performance than FG models. As the same modeling workflow was used, this strongly suggests that FF data is less predictive than FG data.

non-zero concentrations. Thus, we evaluate here the predictive potential of FF datasets. We start by building RF models from FF data using the same 90/10 partitions as with FG. Surprisingly, FF-based models obtained worse performance in every cell line (**Figure 8**) and thus were objectively worse at predicting ComboScores.

In trying to understand this unexpected result, we started by investigating whether this was due to modeling differences, but this was not the case. First, FF training sets are slightly larger than FG datasets (see **Supplementary Table 1**), which theoretically favors better performance on FF. Furthermore, using tuned XGB models led to essentially the same result (median R_p of 0.641 for FG vs. 0.368 for FF) as shown in **Figure 8** with RF. In addition to these non-linear methods, we also used Elastic Net (EN), but FF models were still substantially less predictive than FG models (median R_p of 0.37 for FG vs. 0.23 for FF). When we carried out LODO cross-validations instead of 90/10 partitions, the same trend was observed (**Supplementary Figures 5, 6** also show worse performance of FF-based LODO than that of FG-based LODO in **Figure 6**).

To shed light into this issue, we looked at the only factor that we can compare between these screening centers: the relative growth inhibition (PERCENTGROWTH) induced by a given concentration of a drug tested individually. Interestingly, by counting the different test dates, we observed that FG had on average tested a non-combined drug 3.77 times per cell line, whereas FF almost doubled this number (7.13 times per cell line). A higher number of tests is not in itself worrisome if the growth inhibition of the drug-concentration-cell line tuple is similar between dates. However, if the measurements from these tests are substantially different, this is a problem because the set of ComboScores determined with variable measurements from the same tuple will be inconsistent as well. Consequently, synergy differences between such combinations will not only come from their intrinsic properties, but also from unrelated experimental variability.

To show that higher growth inhibition variability in FF data results in less predictive models, we analyzed five drugs (Thioguanine, Chlorambucil, Altretamine, Fluorouracil, and Melphalan) with a high number of different test dates in both centers. We first consider the drugs on a cell line were only FG models obtain high average accuracy in predicting synergy (NCI/ADR-RES) and subsequently on another where both FF and FG models are on average predictive (NCI-H322M). On each

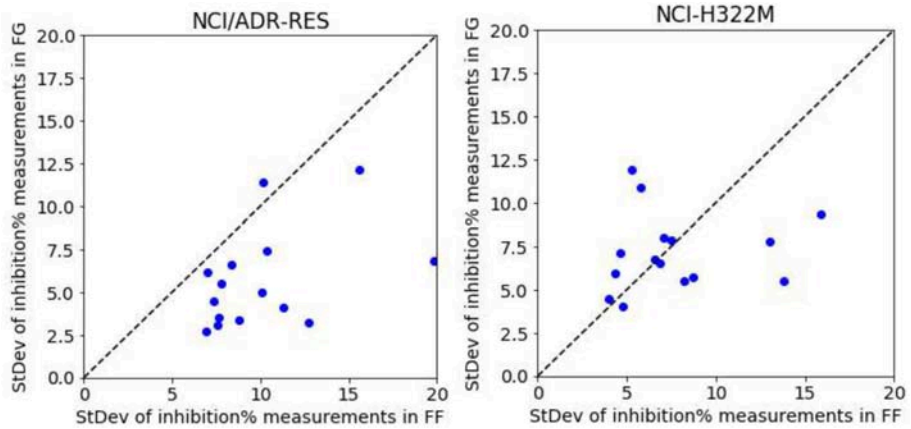


FIGURE 9 | RF model performance comparison for FG and FF datasets. Models are built following the final setup in the exploratory analysis (i.e., a 90/10 data partition is employed for each cell line; MFPC with physico-chemical features as well as data augmentation are also used). We analyzed five drugs (Thioguanine, Chlorambucil, Altretamine, Fluorouracil, and Melphalan) with a high number of different test dates in both centers. On the left, the results with the cell line that is worst predicted by RF with FF data (NCI/ADR-RES with $R_p = 0.14$ in 90/10 partition), which is much better predicted with FG ($R_p = 0.65$, using the same partition). This plot shows the standard deviation of the values of each set from FG against those from FF.

cell line, each drug has a set of growth inhibition replicates per concentration and screening center (i.e., 15 sets per screening center). The performance on NCI/ADR-RES using FF data is indeed poor ($R_p = 0.14$ in 90/10 partition by RF), but it is much better predicted using FG data ($R_p = 0.65$, using the same partition and method). Fourteen of the fifteen sets have higher standard deviation of growth inhibition with FF data (**Figure 9**), which is consistent with the lower accuracy in predicting synergy obtained with this dataset. Conversely, we repeated this operation with NCI-H322M where synergy is well-predicted by RF with both FF ($R_p = 0.61$ in 90/10 partition) and FG data ($R_p = 0.66$, on the same partition). The standard deviations from both screening centers are now similar (**Figure 9**). Taken together, these experiments suggest that the reason why FF data results in less predictive models is the noise introduced in ComboScore determination by larger variability of growth inhibition measurements.

DISCUSSION

NCI-ALMANAC is an extremely valuable resource for the discovery of novel synergistic drug combinations on NCI-60 cell lines. First, it is by far the largest-to-date HTS of drug combinations, therefore allowing *in silico* models with much higher accuracy and broader domain of applicability in predicting the synergy of other combinations. Second, some of the synergistic drug combinations discovered *in vitro* by NCI-ALMANAC were subsequently tested on human tumor mouse xenografts of the same cell line. 48% of them were also synergistic in at least one of these *in vivo* models (Holbeck et al., 2017), which led to the launch of two clinical trials so far (NCT02211755 and NCT02379416).

In this study, we have found that it is possible to predict the synergy of unseen drug combinations against NCI-60 panel cell lines with high accuracy by exploiting NCI-ALMANAC data. We have established a general ML workflow (types of structural features, data preprocessing strategy, ML method) to generate such models. When trained on FG data, predicted synergies from these models match observable synergies with R_p correlations comprised between 0.43 and 0.86 depending on the considered cell line. Incidentally, these regression problems must be highly non-linear, as EN leads to substantially less predictive models than XGB or RF.

Some cell lines and drug combinations can be predicted with higher accuracy than others. For example, models for the SK-MEL-5 cell line perform best with any method (**Figure 6**). However, if we use RMSE instead of R_p to reduce the influence of the ComboScore range, models for the NCI-ARD-RES are now best (gray squares in **Figure 7**). Another explanatory factor for this variability is the adequacy of the employed ML technique to the problem instance to solve (each cell line constitutes here a different problem instance). Even if training set size, features and classifier are the same, the modeled relationship between drug synergy and features depend on training set composition and cell line properties (implicitly). It is well-established that the performance of supervised learning algorithms varies depending on the problem instance in ways that cannot be anticipated without doing the actual numerical experiments (Fernández-Delgado et al., 2014). LODO cross-validation also revealed both best and worst partner drugs. These differences are mainly due to the number of similar partner drugs. For example, it is difficult to predict synergy of combinations containing arsenic trioxide because its 103 partner drugs are highly dissimilar in terms of chemical structure and physico-chemical properties. Indeed, machine learning from dissimilar data instances tend to be less accurate, although here the dissimilarity can be partial as arsenic trioxide's partner can be similar to other NCI-ALMANAC drugs. On the other hand, combinations containing some other drugs are better represented in NCI-ALMANAC and hence tend to be predicted with higher accuracy. This is the case of various

alkylating agents, tyrosine kinase inhibitors and topoisomerase inhibitors (**Supplementary Figure 4**).

Recent QSAR and drug combination modeling studies have evaluated the application of the latest machine learning algorithms (e.g., XGBoost, Deep Neural Network). These studies have found that these algorithms provide better performance on average across targets than RF. However, these gains are small and hence do not always justify the much greater resources required for hyperparameter tuning (Sheridan et al., 2016; Preuer et al., 2018). Performance gains have also been found small here with NCI-ALMANAC data, as the average test set R_p of XGBoost across the 60 cell lines is just +0.007 larger than with RF. An important result is that restricting to the most reliable RF predictions provides much greater predictive accuracy than that introduced by a more suitable learning algorithm (e.g., XGBoost). It is surprising that this powerful technique is so uncommonly used, as has already been pointed out (Sheridan, 2013; Mathea et al., 2016). In fact, we are not aware of any other previous study applying reliability estimation to the prediction of drug synergy on cancer cell lines. Here reliability prediction permitted to reduce the RMSE by up to 50% depending on the cell line. This is particularly exciting for virtual screening problems, where only a small subset of the predictions can be tested *in vitro*. In this scenario, it is useful to identify those combinations that are not only predicted to be synergistic, but also reliable because this should provide higher hit rates. Lastly, highly synergistic combinations predicted with low reliability should also be tested, as the corresponding measurements would be those broadening the applicability domain of future models the most.

We have also found that using FG datasets leads to substantially more predictive models than FF datasets. This result is robust in that it is observed with various types of models (XGB, RF, EN). Moreover, it occurs in spite of the availability of slightly more training data. Further investigation revealed that there are many more measurements of growth inhibition and with greater variability in FF than in FG. This in turn introduces more noise into ComboScore determinations in FF, thus impairing its modeling. Inconsistencies between centers measuring the response of cancer cell lines to drugs have been observed before (Haibe-Kains et al., 2013). There has been intense controversy about the extent, sources and impact of these inconsistencies (Stransky et al., 2015; Geeleher et al., 2016; Safikhani et al., 2016, 2017). In any case, it is clear that data permits the development of predictive models regardless of the screening center (Ammad-ud-din et al., 2014; Covell, 2015; Fang et al., 2015; Naulaerts et al., 2017), as it has also been the case here with NCI-ALMANAC. Owing to this controversy on datasets from multiple screening centers, a better understanding of their limitations and the identification of protocols to generate them with improved consistency has emerged (Haverty et al., 2016). These protocols will ultimately permit that merging datasets from different screening centers result in further predictive accuracy.

CONCLUSION

While NCI-ALMANAC measured the synergies of over 5,000 combinations per cell line, this still represents a minuscule part of all conceivable combinations. Even if we restricted ourselves to the set of 12,000 drug molecules estimated to have reached clinical development or undergone significant preclinical profiling (Janes et al., 2018), almost 72 million combinations per cell line would have to be tested *in vitro* to identify the most synergistic among them. Therefore, the developed *in silico* models are of great importance because these can drastically reduce the number of required *in vitro* tests by predicting which of the considered combinations are likely to be synergistic.

DATA AVAILABILITY

Data available at: https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-ALMANAC.

Code available at: http://ballester.marseille.inserm.fr/NCI-Alm-Predictors.zip.

AUTHOR CONTRIBUTIONS

PB conceived the study and wrote the manuscript. PS implemented the software and carried out the numerical experiments with the assistance of SN. All authors commented and proposed improvements to the manuscript.

FUNDING

ANR Tremplin-ERC grant n° N° ANR-17-ERC2-0003-01 (PB).

ACKNOWLEDGMENTS

The initially submitted version of this manuscript was previously released as a Pre-Print at https://www.biorxiv.org/content/10.1101/504076v1 (Sidorov et al., 2018).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00509/full#supplementary-material

REFERENCES

Ali, M., and Aittokallio, T. (2018). Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys. Rev.* 11, 31–39. doi: 10.1007/s12551-018-0446-z

Ammad-ud-din, M., Georgii, E., Gönen, M., Laitinen, T., Kallioniemi, O., Wennerberg, K., et al. (2014). Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. *J. Chem. Inf. Model.* 54, 2347–2359. doi: 10.1021/ci500152b

Bayat Mokhtari, R., Homayouni, T. S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., et al. (2017). Combination therapy in combating cancer. *Oncotarget* 8, 38022–38043. doi: 10.18632/oncotarget.16723

Bliss, C. I. (1939). The toxicity of poisons applied jointly I. *Ann. Appl. Biol.* 26, 585–615. doi: 10.1111/j.1744-7348.1939.tb06990.x

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Budman, D. R., Calabro, A., Rosen, L., and Lesser, M. (2012). Identification of unique synergistic drug combinations associated with downexpression of survivin in a preclinical breast cancer model system. *Anticancer Drugs* 23, 272–279. doi: 10.1097/cad.0b013e32834ebda4

Bulusu, K. C., Guha, R., Mason, D. J., Lewis, R. P., Muratov, E., Kalantar Motamedi, Y., et al. (2016). Modelling of compound combination effects and applications to efficacy and toxicity: state-of-the-art, challenges and perspectives. *Drug Discov. Today* 21, 225–238. doi: 10.1016/j.drudis.2015.09.003

Chen, J., Shen, Y., Liao, S., Chen, L., and Zheng, K. (2007). DFT-based QSAR study and molecular design of AHMA derivatives as potent anticancer agents. *Int. J. Quantum Chem.* 107, 1468–1478. doi: 10.1002/qua.21285

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '16* (New York, NY: ACM Press), 785–794. Available online at: https://dl.acm.org/citation.cfm?id=2939672&picked=prox

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., et al. (2014). QSAR modeling: where have you been? where are you going to? *J. Med. Chem.* 57, 4977–5010. doi: 10.1021/jm4004285

Chou, T.-C. (2010). Drug combination studies and their synergy quantification using the chou-talalay method. *Cancer Res.* 70, 440–446. doi: 10.1158/0008-5472.CAN-09-1947

Chou, T.-C., and Talalay, P. (1984). Quantitative analysis of dose-effect relationships: the combined effects of multiple drugs or enzyme inhibitors. *Adv. Enzyme Regul.* 22, 27–55. doi: 10.1016/0065-2571(84)90007-4

Covell, D. G. (2015). Data mining approaches for genomic biomarker development: applications using drug screening data from the cancer genome project and the cancer cell line encyclopedia. *PLoS ONE* 10:e0127433. doi: 10.1371/journal.pone.0127433

Crystal, A. S., Shaw, A. T., Sequist, L. V., Friboulet, L., Niederst, M. J., Lockerman, E. L., et al. (2014). Patient-derived models of acquired resistance can identify effective drug combinations for cancer. *Science* 346, 1480–1486. doi: 10.1126/science.1254721

de Luca, A., Horvath, D., Marcou, G., Solov'ev, V., and Varnek, A. (2012). Mining chemical reactions using neighborhood behavior and condensed graphs of reactions approaches. *J. Chem.Inf. Model.* 52, 2325–2338. doi: 10.1021/ci300149n

Di Veroli, G. Y., Fornari, C., Wang, D., Mollard, S., Bramhall, J. L., Richards, F. M., et al. (2016). Combenefit: an interactive platform for the analysis and visualization of drug combinations. *Bioinformatics* 32, 2866–2868. doi: 10.1093/bioinformatics/btw230

Doucet, J.-P., Barbault, F., Xia, H., Panaye, A., and Fan, B. (2007). Nonlinear SVM approaches to QSPR/QSAR studies and drug design. *Curr. Comput. Aided-Drug Des.* 3, 263–289. doi: 10.2174/157340907782799372

Fang, Y., Qin, Y., Zhang, N., Wang, J., Wang, H., and Zheng, X. (2015). DISIS: prediction of drug response through an iterative sure independence screening. *PLoS ONE* 10:e0120408. doi: 10.1371/journal.pone.0120408

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* 15, 3133–3181. Available online at: http://jmlr.org/papers/v15/delgado14a.html

Foucquier, J., and Guedj, M. (2015). Analysis of drug combinations: current methodological landscape. *Pharmacol. Res. Perspect.* 3:e00149. doi: 10.1002/prp2.149

Geeleher, P., Gamazon, E. R., Seoighe, C., Cox, N. J., and Huang, R. S. (2016). Consistency in large pharmacogenomic studies. *Nature* 540, E1–E2. doi: 10.1038/nature19838

González-Díaz, H., Bonet, I., Terán, C., De Clercq, E., Bello, R., García, M. M., et al. (2007). ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *Eur. J. Med. Chem.* 42, 580–585. doi: 10.1016/J.EJMECH.2006.11.016

Greco, W. R., Bravo, G., and Parsons, J. C. (1995). The search for synergy: a critical review from a response surface perspective. *Pharmacol. Rev.* 47, 331–385.

Haibe-Kains, B., El-Hachem, N., Birkbak, N. J., Jin, A. C., Beck, A. H., Aerts, H. J., et al. (2013). Inconsistency in large pharmacogenomic studies. *Nature* 504, 389–393. doi: 10.1038/nature12831

Haverty, P. M., Lin, E., Tan, J., Yu, Y., Lam, B., Lianoglou, S., et al. (2016). Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* 533, 333–337. doi: 10.1038/nature

Holbeck, S. L., Camalier, R., Crowell, J. A., Govindharajulu, J. P., Hollingshead, M., Anderson, L. W., et al. (2017). The National Cancer Institute ALMANAC: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Res.* 77, 3564–3576. doi: 10.1158/0008-5472.can-17-0489

Holohan, C., Van Schaeybroeck, S., Longley, D. B., and Johnston, P. G. (2013). Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer* 13, 714–726. doi: 10.1038/nrc3599

Janes, J., Young, M. E., Chen, E., Rogers, N. H., Burgstaller-Muehlbacher, S., Hughes, L. D., et al. (2018). The ReFRAME library as a comprehensive drug repurposing library and its application to the treatment of cryptosporidiosis. *Proc. Natl. Acad. Sci. U.S.A.* 155, 10750–10755. doi: 10.1073/pnas.1810137115

Kashif, M., Andersson, C., Hassan, S., Karlsson, H., Senkowski, W., Fryknäs, M., et al. (2015). *In vitro* discovery of promising anti-cancer drug combinations using iterative maximisation of a therapeutic index. *Sci. Rep.* 5:14118. doi: 10.1038/srep14118

Kischkel, F. C., Eich, J., Meyer, C. I., Weidemüller, P., Krapfl, J., Yassin-Kelepir, R., et al. (2017). New *in vitro* system to predict chemotherapeutic efficacy of drug combinations in fresh tumor samples. *PeerJ* 5:e3030. doi: 10.7717/peerj.3030

Kumar, R., Chaudhary, K., Singla, D., Gautam, A., and Raghava, G. P. (2014). Designing of promiscuous inhibitors against pancreatic cancer cell lines. *Sci. Rep.* 4:4668. doi: 10.1038/srep04668

Kurtz, S. E., Traer, E., Martinez, J., Park, A., Wagner, J., Pandya, R., et al. (2015). Identification of effective targeted drug combinations using functional *ex vivo* screening of primary patient specimens. *Blood* 126:865. Available online at: http://www.bloodjournal.org/content/126/23/865

Kuz'min, V. E., Artemenko, A. G., and Muratov, E. N. (2008). Hierarchical QSAR technology based on the Simplex representation of molecular structure. *J. Comput. Aided. Mol. Des.* 22, 403–421. doi: 10.1007/s10822-008-9179-6

Lamdrum, G. (2015). *RDKit: Open-source Cheminformatics*. Available online at: http://www.rdkit.org/ (accessed April 3, 2015).

Leach, A. R., and Gillet, V. J. (2007). *An Introduction To Chemoinformatics*. Dordrecht: Springer Netherlands.

Li, H., Li, T., Quang, D., and Guan, Y. (2018). Network Propagation Predicts Drug Synergy in Cancers. *Cancer Res.* 78, 5446–5457. doi: 10.1158/0008-5472.CAN-18-0740

Lieu, C. H., Tan, A.-C., Leong, S., Diamond, J. R., and Eckhardt, S. G. (2013). From bench to bedside: lessons learned in translating preclinical studies in cancer drug development. *JNCI J. Natl. Cancer Inst.* 105, 1441–1456. doi: 10.1093/jnci/djt209

Mathea, M., Klingspohn, W., and Baumann, K. (2016). Chemoinformatic classification methods and their applicability domain. *Mol. Inform.* 35, 160–180. doi: 10.1002/minf.201501019

Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., et al. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS ONE* 8:e61318. doi: 10.1371/journal.pone.0061318

Menden, M. P., Wang, D., Mason, M. J., Szalai, B., Bulusu, K. C., Guan, Y., et al. (2019). Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.* 10:2674. doi: 10.1038/s41467-019-09799-2

Mullen, L. M. A., Duchowicz, P. R., and Castro, E. A. (2011). QSAR treatment on a new class of triphenylmethyl-containing compounds as potent anticancer agents. *Chemom. Intell. Lab. Syst.* 107, 269–275. doi: 10.1016/J.CHEMOLAB.2011.04.011

Muratov, E. N., Varlamova, E. V., Artemenko, A. G., Polishchuk, P. G., and Kuz'min, V. E. (2012). Existing and developing approaches for QSAR analysis of mixtures. *Mol. Inform.* 31, 202–221. doi: 10.1002/minf.201100129

Naulaerts, S., Dang, C. C., and Ballester, P. J. (2017). Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours. *Oncotarget* 8:20923. doi: 10.18632/oncotarget.20923

O'Neil, J., Benita, Y., Feldman, I., Chenard, M., Roberts, B., Liu, Y., et al. (2016). An unbiased oncology compound screen to identify novel combination strategies. *Mol. Cancer Ther.* 15, 1155–1162. doi: 10.1158/1535-7163.MCT-15-0843

Pick, A., Müller, H., Mayer, R., Haenisch, B., Pajeva, I. K., Weigt, M., et al. (2011). Structure–activity relationships of flavonoids as inhibitors of breast cancer resistance protein (BCRP). *Bioorg. Med. Chem.* 19, 2090–2102. doi: 10.1016/J.BMC.2010.12.043

Polishchuk, P. G., Madzhidov, T. I., and Varnek, A. (2013). Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput. Aided. Mol. Des.* 27, 675–679. doi: 10.1007/s10822-013-9672-4

Preuer, K., Lewis, R. P. I., Hochreiter, S., Bender, A., Bulusu, K. C., and Klambauer, G. (2018). DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics* 34, 1538–1546. doi: 10.1093/bioinformatics/btx806

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi: 10.1021/ci100050t

Ruggiu, F., Marcou, G., Solov 'ev, V., Horvath, D., and Varnek, A. (2015). *ISIDA Fragmentor2015—User Manual*.

Sabet, R., Mohammadpour, M., Sadeghi, A., and Fassihi, A. (2010). QSAR study of isatin analogues as *in vitro* anti-cancer agents. *Eur. J. Med. Chem.* 45, 1113–1118. doi: 10.1016/J.EJMECH.2009.12.010

Safikhani, Z., El-Hachem, N., Smirnov, P., Freeman, M., Goldenberg, A., Birkbak, N. J., et al. (2016). Safikhani et al. reply. *Nature* 540, E2–E4. doi: 10.1038/nature19839

Safikhani, Z., Smirnov, P., Freeman, M., El-Hachem, N., She, A., Rene, Q., et al. (2017). Revisiting inconsistency in large pharmacogenomic studies. *F1000Res.* 5:2333. doi: 10.12688/f1000research.9611.3

Sheridan, R. P. (2013). Using random forest to model the domain applicability of another random forest model. *J. Chem. Inf. Model.* 53, 2837–2850. doi: 10.1021/ci400482e

Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J., and Gifford, E. M. (2016). Extreme gradient boosting as a method for quantitative structure–activity relationships. *J. Chem. Inf. Model.* 56, 2353–2360. doi: 10.1021/acs.jcim.6b00591

Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* 6, 813–823. doi: 10.1038/nrc1951

Sidorov, P., Naulaerts, S., Ariey-Bonnet, J., Pasquier, E., and Ballester, P. (2018). Predicting synergism of cancer drug combinations using NCI-ALMANAC data. *bioRxiv* 10:504076. doi: 10.1101/504076

Singh, H., Singh, S., Singla, D., Agarwal, S. M., and Raghava, G. P. (2015). QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. *Biol. Direct* 10:10. doi: 10.1186/s13062-015-0046-9

Speck-Planche, A., Kleandrova, V. V., Luan, F., and Cordeiro, M. N. (2011). Multi-target drug discovery in anti-cancer therapy: fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. *Bioorg. Med. Chem.* 19, 6239–6244. doi: 10.1016/J.BMC.2011.09.015

Speck-Planche, A., Kleandrova, V. V., Luan, F., and Cordeiro, M. N. (2012). Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the *in silico* discovery of anti-breast cancer agents. *Eur. J. Pharm. Sci.* 47, 273–279. doi: 10.1016/J.EJPS.2012.04.012

Stransky, N., Ghandi, M., Kryukov, G. V., Garraway, L. A., Lehár, J., Liu, M., et al. (2015). Pharmacogenomic agreement between two cancer cell line data sets. *Nature* 528, 84–87. doi: 10.1038/nature15736

Sugahara, K. N., Teesalu, T., Karmali, P. P., Kotamraju, V. R., Agemy, L., Greenwald, D. R., et al. (2010). Coadministration of a tumor-penetrating peptide enhances the efficacy of cancer drugs. *Science* 328, 1031–1035. doi: 10.1126/science.1183057

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 43, 1947–1958. doi: 10.1021/ci034160g

Todeschini, R., and Consonni, V. (2000). *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH Verlag GmbH.

Tropsha, A., Gramatica, P., and Gombar, V. (2003). The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* 22, 69–77. doi: 10.1002/qsar.200390007

Varnek, A., Fourches, D., Hoonakker, F., and Solov'ev, V. P. (2005). Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput. Aided. Mol. Des.* 19, 693–703. doi: 10.1007/s10822-005-9008-0

Xia, F., Shukla, M., Brettin, T., Garcia-Cardona, C., Cohn, J., Allen, J. E., et al. (2018). Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics* 19:3. doi: 10.1186/s12859-018-2509-3

Yu, D., Kahen, E., Cubitt, C. L., McGuire, J., Kreahling, J., Lee, J., et al. (2015). Identification of synergistic, clinically achievable, combination therapies for osteosarcoma. *Sci. Rep.* 5:16991. doi: 10.1038/srep16991

Zoli, W., Ricotti, L., Tesei, A., Barzanti, F., and Amadori, D. (2001). *In vitro* preclinical models for a rational design of chemotherapy combinations in human tumors. *Crit. Rev. Oncol. Hematol.* 37, 69–82. doi: 10.1016/S1040-8428(00)00110-4

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Sidorov, Naulaerts, Ariey-Bonnet, Pasquier and Ballester. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 16 July 2019
doi: 10.3389/fchem.2019.00510

Finding Constellations in Chemical Space Through Core Analysis

J. Jesús Naveja[1,2]* *and José L. Medina-Franco*[2]*

[1] PECEM, School of Medicine, Universidad Nacional Autónoma de México, Mexico City, Mexico, [2] Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, Mexico

Herein we introduce the constellation plots as a general approach that merges different and complementary molecular representations to enhance the information contained in a visual representation and analysis of chemical space. The method is based on a combination of a sub-structure based representation and classification of compounds with a "classical" coordinate-based representation of chemical space. A distinctive outcome of the method is that organizing the compounds in analog series leads to the formation of groups of molecules, aka "constellations" in chemical space. The novel approach is general and can be used to rapidly identify, for instance, insightful and "bright" Structure-Activity Relationships (StARs) in chemical space that are easy to interpret. This kind of analysis is expected to be especially useful for lead identification in large datasets of unannotated molecules, such as those obtained through high-throughput screening. We demonstrate the application of the method using two datasets of focused inhibitors designed against DNMTs and AKT1.

Keywords: analog series, data visualization, descriptor, scaffold, structure-property relationships

OPEN ACCESS

Edited by:
Simona Rapposelli, University of Pisa, Italy

Reviewed by:
Chanin Nantasenamat, Mahidol University, Thailand
Oscar Mendez Lucio, Bayer, France

***Correspondence:**
J. Jesús Naveja
naveja@comunidad.unam.mx
José L. Medina-Franco
medinajl@unam.mx

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *14 May 2019*
Accepted: *03 July 2019*
Published: *16 July 2019*

Citation:
Naveja JJ and Medina-Franco JL (2019) Finding Constellations in Chemical Space Through Core Analysis. Front. Chem. 7:510. doi: 10.3389/fchem.2019.00510

INTRODUCTION

The concept of chemical space is broadly used in drug discovery because of its multiple potential applications; for instance, in library design, compound or dataset classification, compound selection, exploration of structure-activity relationships (SAR), and navigation though structure-property relationships (SPR) in general. However, a precise unique definition of chemical space is not simple. An even more challenging task is the visual representation of this subjective concept.

Chemical space is usually defined as the set of all possible organic compounds (Lipinski and Hopkins, 2004). It is widely recognized that the virtual chemical space is more than astronomically large, as not even all atoms in the universe would suffice to synthesize a single molecule from each of all the 10^{63} possible organic compounds of a size up to 30 atoms (Clayden et al., 2012). Nevertheless, massive efforts have been undertaken to enumerate billions of hypothetical organic compounds, thus allowing large virtual screening campaigns to take place (Reymond, 2015; Lyu et al., 2019).

Along with the increasing size of the mapped chemical space, the interest of applying cartographic methods to visualize the space has expanded (Oprea and Gottfries, 2001). As a result, numerous visualization and conceptualization approaches into chemical space have emerged (Larsson et al., 2007; Osolodkin et al., 2015; Naveja and Medina-Franco, 2017). A cornerstone and key aspect of all proposed methods is the molecular representation and parameters used to define the space where the compounds will reside. Chemical space visualizations have to reduce the dimensionality of the problem of comparing molecular structures, which can be done through algorithms such as principal components analysis and t-distributed stochastic neighbor embedding (see Osolodkin et al., 2015).

In most chemical space approaches, it is desirable that chemical analogs are closer to each other than unrelated and dissimilar molecules since this allows machine learning methods to identify clusters of structurally-related molecules (Medina-Franco et al., 2008; Naveja and Medina-Franco, 2015; Naveja et al., 2016, 2018a). In addition, clustering analog series would allow, at least in principle, to map SAR/SPR into that space. However, due to the vast amplitude of the chemical space and the inevitable loss of information with an initially large space projected into lower dimensions, it is expected that non-analog compounds will end up in the same cluster. Also, when many points in the chemical space are considered at once, visualizations become harder to interpret. To address this issue, approaches such as virtual reality have emerged (Probst and Reymond, 2018).

In parallel to such chemical space approaches based on coordinates, scaffold analysis is a more consistent and chemically-intuitive approach for exploring and identifying collections of analogs (Hu et al., 2011). Ever since the pioneering work by Bemis and Murcko (1996), computational identification of chemical scaffolds has been refined. In this line, Stumpfe et al. (2016) recently introduced the analog series-based scaffold (ASBS), a revolutionary scaffold concept that is more flexible and chemically sound than its predecessors. In fact, the ASBS has proven to yield more biologically meaningful structure-activity/property relationships (SA/PR) than other scaffold definitions (Dimova et al., 2016; Kunimoto et al., 2017; Bajorath, 2018; Dimova and Bajorath, 2018).

Although the chemical space of single analog series can be effectively explored and used, for instance, to guide lead optimization programmes (Vogt et al., 2018), methods for analyzing the relationship among scaffolds of different analog series remain to be explored. Of note, a difficulty in this regard emerges as analog-series based scaffolds tend not to be as consistent as Bemis-Murcko scaffolds, since they result from the retrospective analysis of analog series (Bajorath, 2018). Accordingly, a core framework inspired in the design of the ASBS avoids the shortcoming of inconsistency by allowing molecules to be annotated with more than one putative core (Naveja et al., Submitted). Hence, large libraries containing analogs can be condensed into fewer cores. In this way, SA/PR can be preferentially analyzed for the most explored regions of the chemical space: analog series.

Herein, we present a general methodology for applying the putative core framework to produce more concise and meaningful representations of the chemical space. To our understanding, this is the first method designed for charting multiple analog series into a coordinate-based chemical space, thus combining in a single plot two general and useful approaches of molecular representation and mapping. Of note, since within this framework cores may share analogs (i.e., analog series are allowed to share compounds), such cores can be connected, thus resembling constellations in the chemical space. Therefore, we termed the resulting graphics "constellation plots." As it will be discussed, activity data (or any property of interest) can be mapped into the constellation plot allowing to explore SA/PRs in the space and quickly identify interesting regions in the space. The rest of this methodological paper is organized as follows: first, the concept scheme is presented and the formalism explained through a toy example; thereafter, two case studies using exemplary datasets are presented; finally, we discuss the conclusions and perspectives of this novel approach for combining the scaffold and the chemical space concepts.

METHODS

Datasets Used in the Examples

For illustrating the application of constellation plots in two different context of analysis, we used two benchmark datasets that have been previously explored with other analysis approaches. One set was a group of 827 AKT1 inhibitors extracted and curated from ChEMBL (Gaulton et al., 2017; Naveja et al., 2018b). The second dataset was a collection of 286 compounds tested as inhibitors of DNMT (DNA methyltransferases). This second data set was integrated from multiple sources of information as described in Naveja and Medina-Franco (2018). Since this dataset integrates qualitative (such as those containing crystallographic data) and quantitative databases (such as those containing experimental determination of inhibition curves), for this dataset, we use a categorical classification of activity in "active" or "inactive." The files of the two datasets are included as **Supplementary Information**.

Chemical Space and Analog Series

As mentioned above, constellation plots fuse two ligand-based concepts: chemical space and core analysis. Standard chemical space analysis is carried out by computing descriptors for a collection of molecules (e.g., physicochemical properties and/or structural features) and then applying dimensionality reduction approaches (Rosén et al., 2009; Osolodkin et al., 2015; González-Medina et al., 2016; Prieto-Martínez et al., 2016; Naveja and Medina-Franco, 2017; Borrel et al., 2018). As a result, every data point represents a single molecule (see **Figure 1**). This can render many visualizations hard to read and analyze by the naked eye. Furthermore, the numerous descriptors used are combined, such that every axis in the visualization turns out to have a quite abstract meaning. Herein, for the purpose of charting chemical space, *t*-distributed stochastic neighbor embedding (*t*-SNE) is used. This methodology reduces the number of data points in the center of the map as compared to other approaches and has been used successfully in chemical space charting (Maaten and Hinton, 2008; Lewis et al., 2015). However, other coordinate-based representations of chemical space can be used in this general approach.

In contrast to chemical space, standard scaffold and analog series analysis aims toward a clear and consistent picture of the relationships among compounds. For instance, a scaffold is a substructure shared by all compounds annotated with it. A state-of-the-art approach for defining analog series-based scaffolds was proposed by Stumpfe et al. (2016). They have reasoned that for a scaffold to be relevant in medicinal chemistry, it should not only be a substructure of a molecule, but it also has to comply with three additional criteria: (i) be a major component of the whole molecule, (ii) be derived from the molecule through retrosynthetic rules, and (iii) summarize an

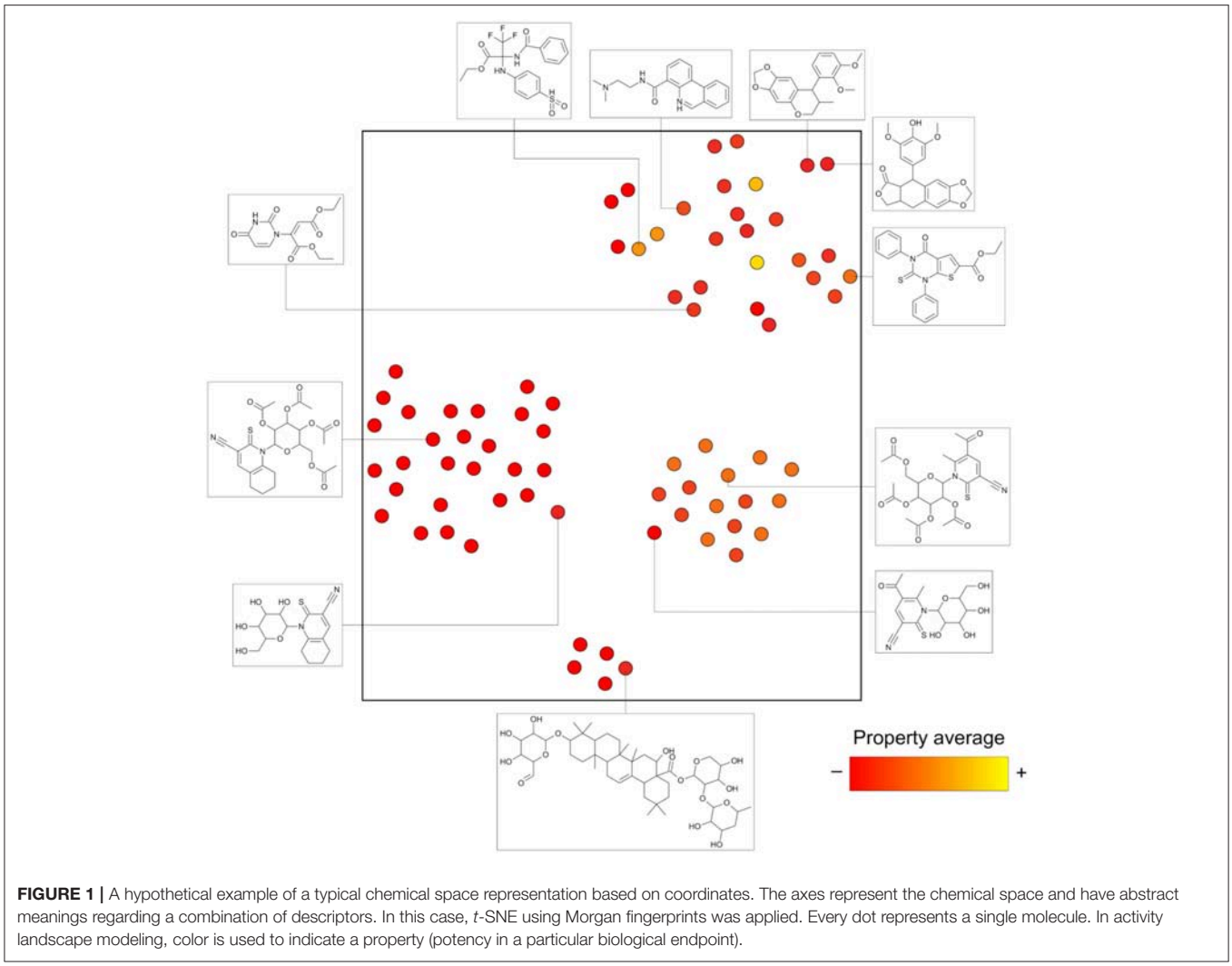


FIGURE 1 | A hypothetical example of a typical chemical space representation based on coordinates. The axes represent the chemical space and have abstract meanings regarding a combination of descriptors. In this case, *t*-SNE using Morgan fingerprints was applied. Every dot represents a single molecule. In activity landscape modeling, color is used to indicate a property (potency in a particular biological endpoint).

analog series in a particular dataset. A number of computational approaches for obtaining ASBS have been proposed (Dimova et al., 2016; Stumpfe et al., 2016; Bajorath, 2018; Naveja et al., 2019). Within these approaches, an analog series is defined as a subnetwork connected by matched molecular pairs (MMPs) (Griffen et al., 2011).

Chemical space analysis of individual analog series has been carried out to measure progression in lead optimization and saturation of analog series (Kunimoto et al., 2018; Vogt et al., 2018; Yonchev et al., 2018). Nevertheless, the fact that assumption (iii) makes analog series inconsistent in as much as the scaffold definition is dependent on the dataset used (Bajorath, 2018) is a limitation for the exploration of chemical space of multiple analog series at once. In a recent study (Naveja et al., Submitted), we discussed that by removing assumption (iii) two effects take place: first, every molecule is allowed to be annotated to more than a single core (equivalent to the term "scaffold"); and second, complete consistency is achieved as no core annotations are ever omitted for any molecule (see **Figure 2**). It is within this general core framework that we propose using constellation plots.

Summarizing Analog Series Information in a Dataset Within the General Core Framework

Since the general core framework can assign multiple cores to single molecules, a useful step prior to mapping cores in the chemical space would be summarizing analog series in the smallest number of cores possible. As illustrated in **Figure 3**, in some instances it is possible to summarize a whole analog series in a single core structure, while in other cases this cannot be done without loss of information. Hence, for avoiding such situations, we did not discard cores unless only one compound mapped to it. Furthermore, if two or more cores mapped to exactly the same compounds, then only the largest core was kept and the others were disregarded from the analysis.

Constellation Plots

After processing a collection of compounds under the general core framework, information is obtained in multiple regards, namely: (a) the chemical structure of every core; (b) the sets of

molecules mapping to each core; (c) the molecules annotated to multiple cores; and (d) the analog series to which each compound and core are annotated. We propose a visualization methodology summarizing these four dimensions in a single graphic: the constellation plot that is schematically illustrated in **Figure 4**.

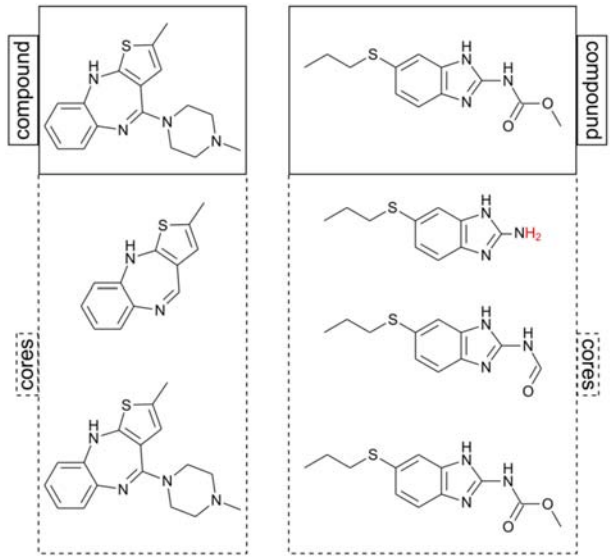


FIGURE 2 | Two examples of putative cores computed for two molecules. Note that in this approach the same chemical structure can be its own core (structures at the bottom). After RECAP fragmentation, hydrogens are added to the core structure to avoid invalid valence (marked in red).

Essentially, in a constellation plot, the chemical structure of representative cores in a database (for example, those annotated with a predefined minimum number of compounds) is used to find descriptors and map them into a chemical space as if they were single molecules. The size of the circles is used to represent the relative number of compounds annotated to each core. Cores sharing compounds are connected by lines forming "constellations" in the chemical space. Every circle is labeled with an identifier for the analog series to which each core belongs. Additionally, a color scale can be used to represent an average of a given property or activity of the compounds annotated with each core, thereby turning constellation plots useful for activity landscape modeling (Waddell and Medina-Franco, 2012). Of note, the activity can be, for instance, measured for a single molecular target. However, the property could also be a representative measure of the selectivity or promiscuity profile of all the compounds sharing a core across multiple biological endpoints (see section Conclusions and Perspectives).

Figure 4, as opposed to **Figure 1**, is able to summarize a larger number of compounds than points depicted and contains information about actual analogs. For instance, analog series I, J, and L form separate clusters, but the cluster top right has multiple chemotypes of distinct analog series. This could not be inferred from clustering algorithms applied to the chemical space information only.

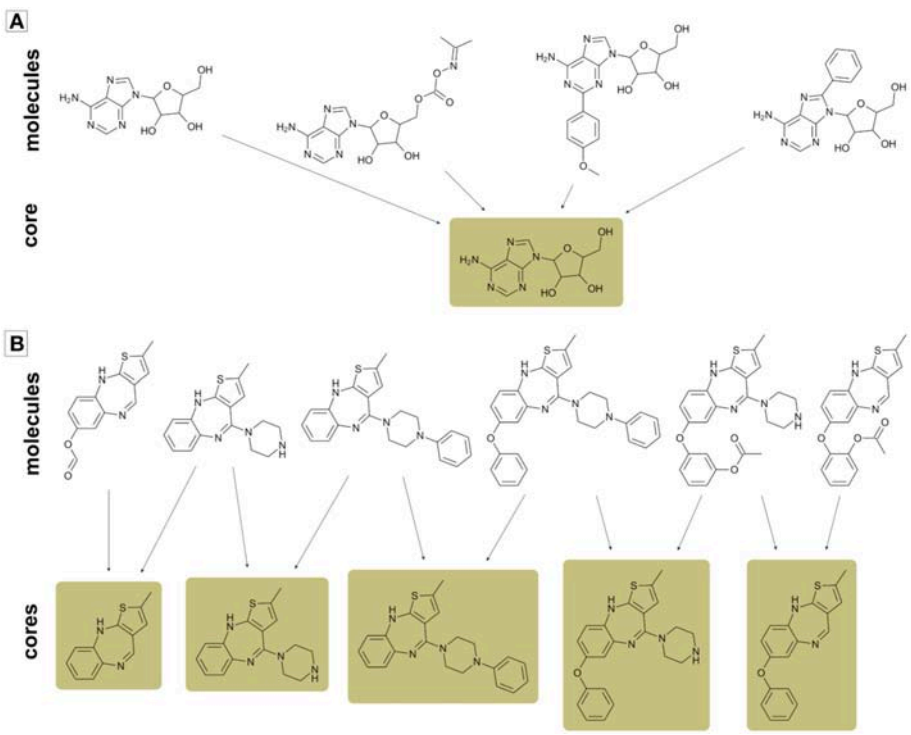


FIGURE 3 | Examples of two analog series with multiple compounds and cores. **(A)** Analog series that can be summarized in a single core; **(B)** Analog series formed by multiple cores. In case **(B)** a single core is not enough for summarizing all information in an analog series.

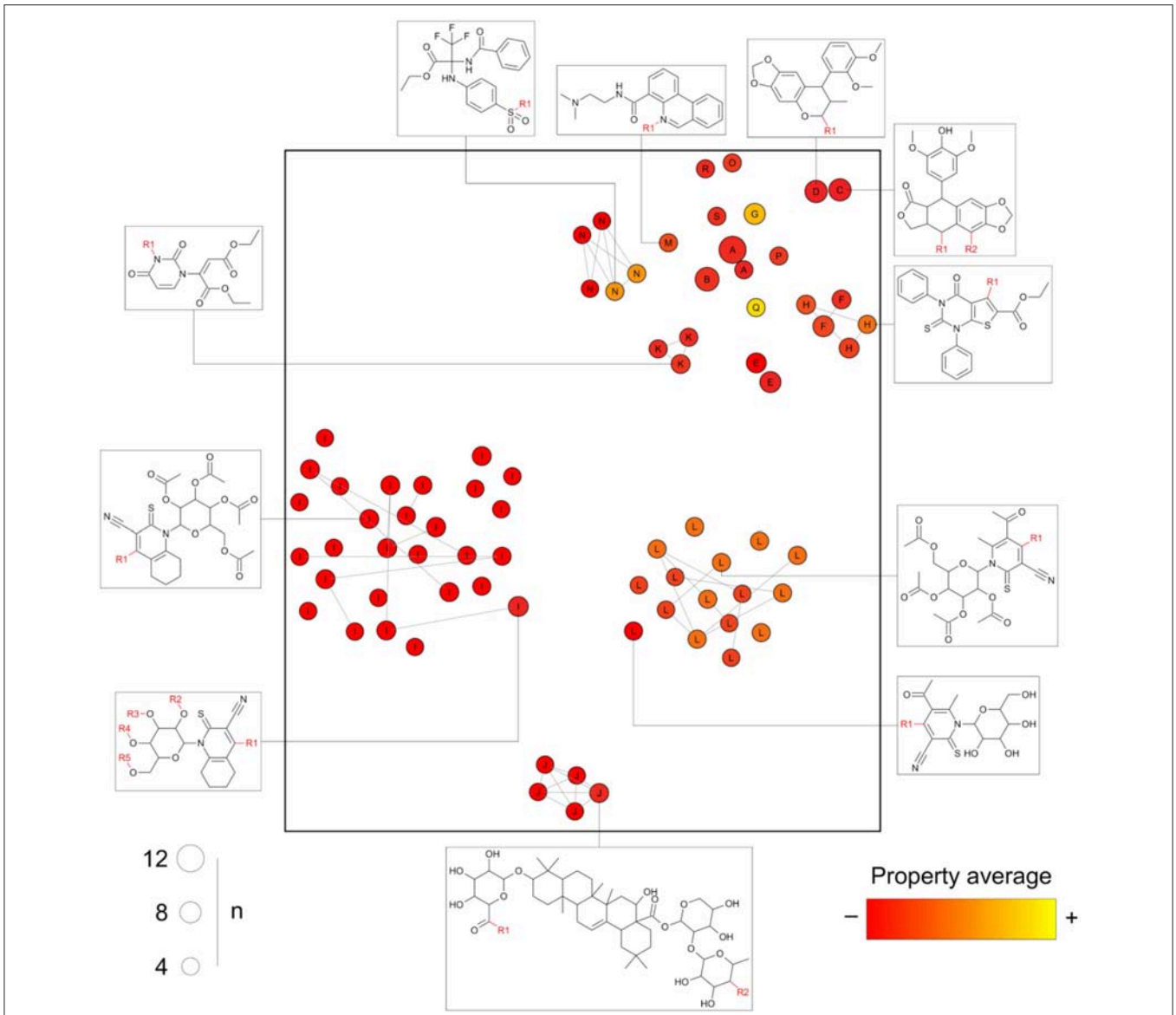


FIGURE 4 | Schematic representation of a general form of a constellation plot. Every circle in the plot represents a core; the axes comprise the coordinates of the chemical structure of the core projected into a 2D representation of the chemical space as computed by any of the standard approaches (e.g., generated using continuous properties or molecular fingerprints and applying *t*-SNE or principal components analysis); the size of the circles indicates the "n" number of compounds annotated to a given core; connected circles are cores sharing compounds; the labels indicate the analog series every point belongs to; the color scale represents the average of a property/activity of the compounds mapping to the core.

Implementation

All scripts required for producing the data herein reported use free Python code and are made freely available in **Supplementary Information**. RDkit was used for computing fingerprints and manipulation of chemical structures (http://www.rdkit.org). Scikit-learn was used for computing *t*-SNE (Pedregosa et al., 2011).

RESULTS AND DISCUSSION

The construction of constellations plots and exemplary applications are illustrated with two case studies of general interest in drug discovery. As mentioned in the section Methods, the first example consists of a dataset of 827 AKT1 inhibitors obtained from ChEMBL (Gaulton et al., 2017) and cheminformatically described in Naveja et al. (2018b). The second example employs a data set of 286 DNA methyltransferase (DNMT) inhibitors obtained from the integration of several databases as described in Naveja and Medina-Franco (2018).

Case Study 1: AKT1 Inhibitors

Analogs in this library could be summarized in 144 cores as discussed in the section Methods. The cores were organized in 79 analog series and contained 440 compounds (about half of the

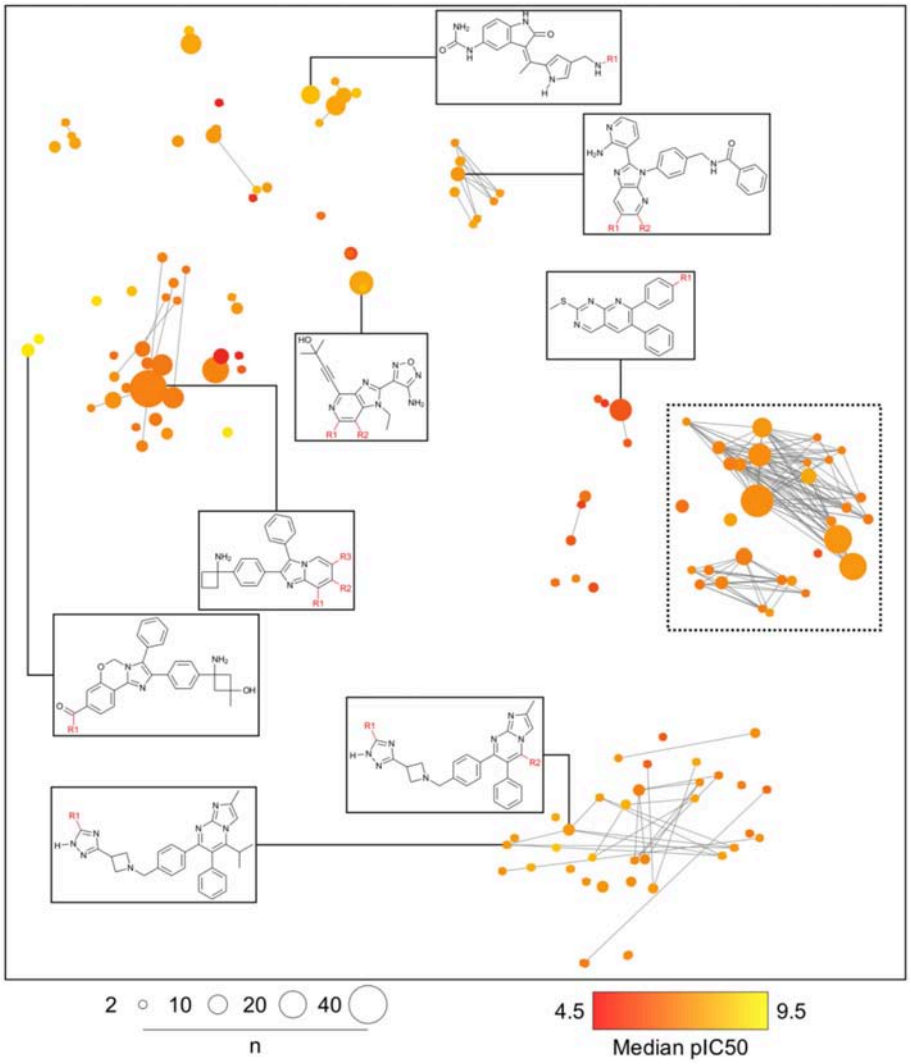


FIGURE 5 | General constellation plot for a dataset of AKT1 inhibitors. It is possible to navigate this map, as observed in **Figure 6**, where the constellation framed within dashed lines is further explored.

initial dataset). **Figure 5** is the constellation plot for these data, where it becomes apparent that chemical space and chemical substructure information play simultaneous roles in describing the SAR. For instance, although some inactive cores are close to active cores in chemical space, they are not usually contained in the same analog series. Therefore, these could be categorized as "scaffold cliffs" rather than simple activity cliffs conceptualized as two small molecules with similar structures and very different activities (Maggiora, 2006). In this case, collections of molecules, rather than single molecules, are being compared.

Figure 6 is a zoomed-in picture into a single "bright" (or predominantly active) constellation comprising five analog series and 55 compounds. As it is readily observed, analog series close in the chemical space have only slight dissimilarities within their scaffolds; in this case, they all share a naphthyridine or naphthyridinone scaffold. Constellation plots allow for a more precise visual SAR analysis and generation of hypotheses. For instance, the core associated to analog series 62 has only a different position for the nitrogens in the rings as well as where substitutions occur. Structural studies could then be conducted to elucidate which are the most relevant features for this kind of scaffolds to be active against AKT1. In this regard, a recent publication co-crystallizing 1,6-naphthyridinone derivatives similar to those in analog series 20 has shown that this scaffold is relevant in forming a π-π stacking interaction with the side chain of Trp80 of the PH-domain (Uhlenbrock et al., 2019). Nonetheless, variation of the position of nitrogen atoms in the scaffold were not considered in the cited study. Indeed, previous SAR studies of these analogs have found the position of the nitrogen atoms in these scaffolds to be critical for the activity against AKT (Zhao et al., 2005; Bilodeau et al., 2008).

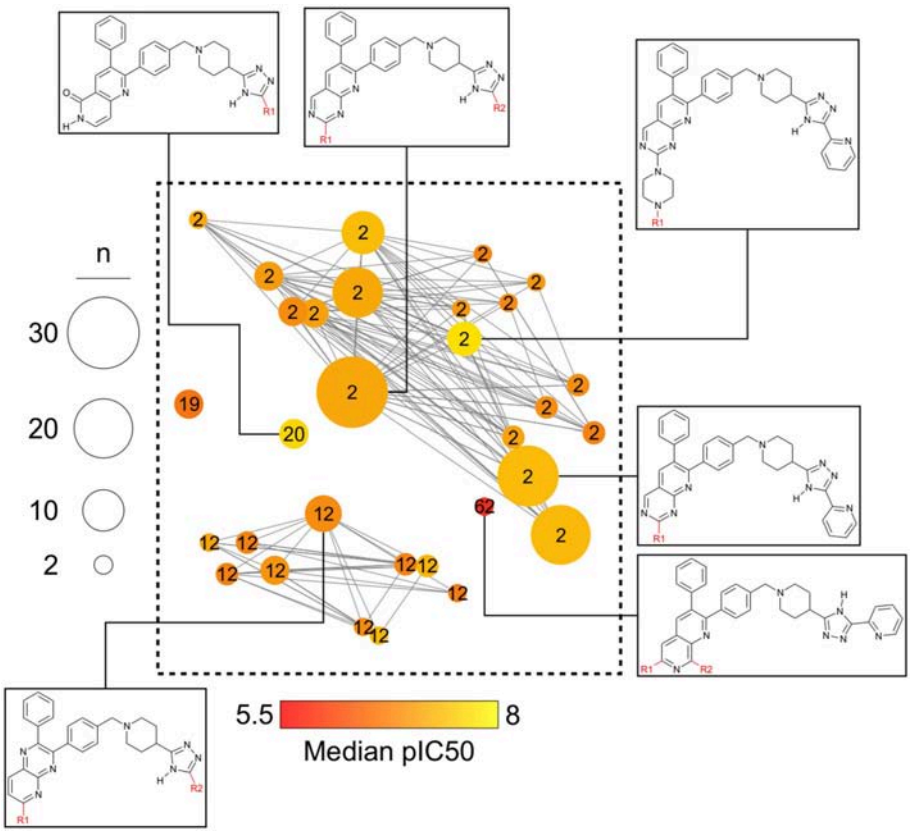


FIGURE 6 | "Zoom-in" into the constellation plot for AKT1 inhibitors selected from **Figure 5**. The core in analog series 62 is not as active as the nearby cores in analog series 2 and 12. Few structural differences can be noted for the compounds in this constellation.

Case Study 2: DNMT Inhibitors

Analogs in this library could be summarized in 23 cores following the procedure discussed in the section Methods. The cores were organized in 13 analog series and contained 46 compounds (about 16% of the initial dataset). Compounds in this library have annotated activity with DNMT1, DNMT3A, and/or DNMT3B. **Figure 7** shows three constellations plots, where chemical space is the same and colors change to represent the activities against each DNMT. As elaborated on the section Methods, each circle in the plot represents a core in which coordinates in the 2D graph is given by similarity measurements computed from Morgan fingerprints using *t*-SNE for dimensionality reduction. Labels indicate the analog series to which cores belong. The color represents the percentage of active compounds sharing that core using a continuous color scale from red (less active cores) to yellow (more active cores). For this example of use of the constellation plots, the definition of "active" was determined from integrating qualitative and quantitative data sources as described in Naveja and Medina-Franco (2018). Circles in gray indicate cores with no reported activity for that particular DNMT. The size of the circle indicates the number of compounds sharing the core. Connected circles are cores sharing compounds. **Figure 7** also shows the chemical structures of representative cores.

The constellation plots for DNMT inhibitors in **Figure 7** allow for rapidly getting several interesting insights of the SAR. For instance, cores at the top left part of the plot from analog series "A" are a bright constellation against DNMT1, i.e., a region in chemical space with active analogs. However, these analogs have not been tested against the other DNMT isoenzymes, which would help determine whether these inhibitors are selective.

Of note, there is a "dark" (or predominantly inactive) constellation in the chemical space of DNMT1 formed by six cores from analog series "D." This dark constellation, however, is more active overall against DNMT3A and appears to be active against DNMT3B. Furthermore, not all cores in this constellation have been tested against DNMT3A and DNMT3B, where they have greater chances of being active.

The plot also reveals a constellation of nucleoside analogs from series "B" at the bottom-right region of the plot that is, overall, selective toward DNMT3B vs. DNMT1. This series has not been tested against DNMT3A yet. Moreover, most of the cores have been tested in DNMT1 only, thus hindering discussions on selectivity. In this regard, analysis of constellation plots is visually helpful in guiding multitarget drug discovery campaigns and in finding opportunities for selectivity.

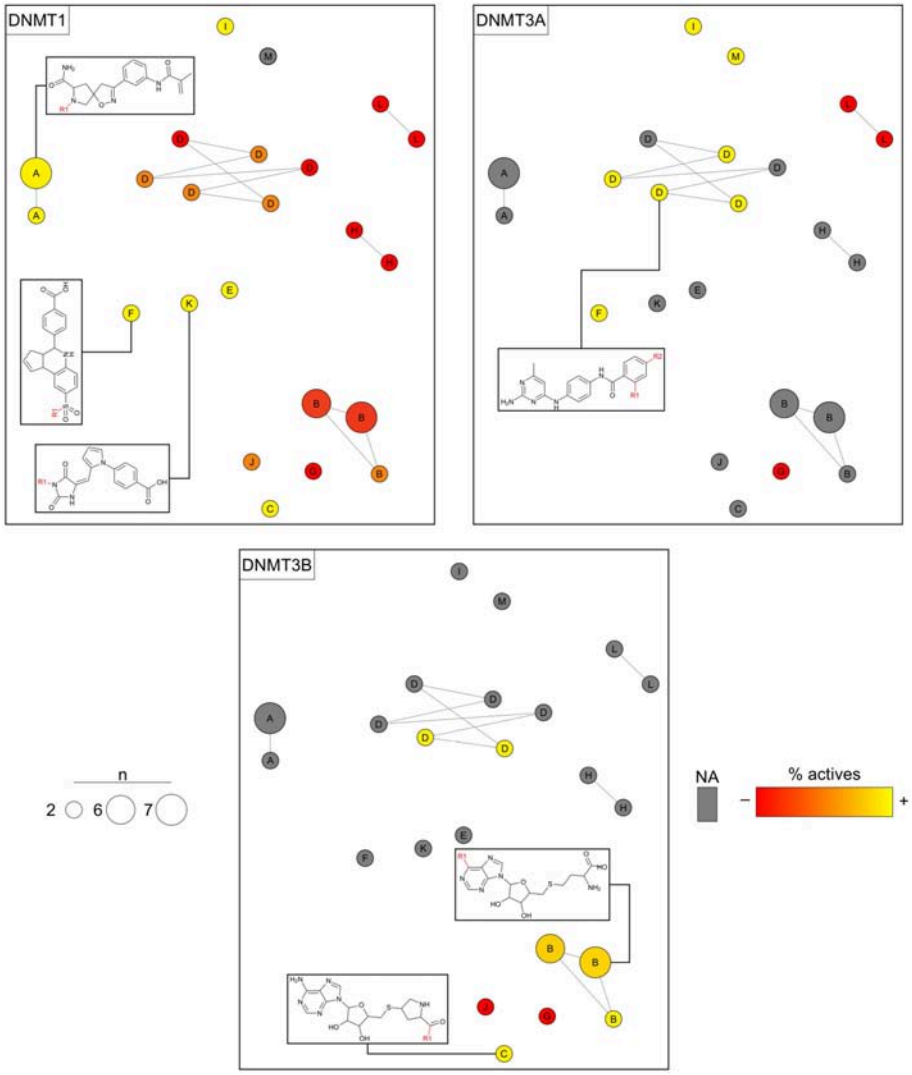


FIGURE 7 | Constellation plots for a dataset of DNA methyltransferase (DNMT) inhibitors tested with DNMT1, DNMT3A, and DNMT3B.

CONCLUSIONS AND PERSPECTIVES

We introduced a novel approach for combining chemical space and analog series methodologies into a single descriptive analysis that can be summarized in a constellation plot. Adding the analog series concept into the chemical space facilitates discussions of regions in the space, as every point summarizes a collection of analogs. A so-called "constellation in chemical space" can be conceptualized as those regions in chemical space formed by core scaffolds with similar structure (as defined by a coordinate-based projection). Mapping activity on the plot readily uncovers active and inactive zones, e.g., bright or dark regions, in chemical space. Of note, constellation plots would be useful for exploring virtually any chemical property, such as biological activity (as demonstrated with two case studies), but also physicochemical properties, complexity or selectivity statistics. In this regard, constellation plots are a flexible approach with multiple potential applications in academia and industry, aiding in the quest of finding potential leads from large collections of screening data (e.g., such as that produced by high-throughput screening campaigns). One of the next steps of this work is the application of the constellations plots to navigate through cell selectivity data of a comprehensive screening dataset. Results will be disclosed in due course.

DATA AVAILABILITY

The datasets generated for this study can be found in **Supplementary Information**.

AUTHOR CONTRIBUTIONS

JN participated in the conceptualization, data gathering, data analysis, and drafted the first

version of the manuscript. JM-F participated in the conceptualization, data analysis, and revision of the manuscript.

ACKNOWLEDGMENTS

JN was thankful to CONACyT for the granted scholarship number 622969. We also thank Consejo Nacional de Ciencia y Tecnología (CONACyT) for grant 282785.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00510/full#supplementary-material

REFERENCES

Bajorath, J. (2018). Improving the utility of molecular scaffolds for medicinal and computational chemistry. *Future Med. Chem.* 10, 1645–1648. doi: 10.4155/fmc-2018-0106

Bemis, G. W., and Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* 39, 2887–2893. doi: 10.1021/jm9602928

Bilodeau, M. T., Balitza, A. E., Hoffman, J. M., Manley, P. J., Barnett, S. F., Defeo-Jones, D., et al. (2008). Allosteric inhibitors of Akt1 and Akt2: a naphthyridinone with efficacy in an A2780 tumor xenograft model. *Bioorg. Med. Chem. Lett.* 18, 3178–3182. doi: 10.1016/j.bmcl.2008.04.074

Borrel, A., Kleinstreuer, N. C., and Fourches, D. (2018). Exploring drug space with ChemMaps.com *Bioinformatics* 34, 3773–3775. doi: 10.1093/bioinformatics/bty412

Clayden, J., Greeves, N., and Warren, S. (2012). *Organic Chemistry*. Oxford, UK: Oxford University Press. Available online at: http://global.oup.com/ukhe/product/organic-chemistry-9780199270293?cc=mx&lang=en&

Dimova, D., and Bajorath, J. (2018). Collection of analog series-based scaffolds from public compound sources. *Future Sci. OA* 4:FSO287. doi: 10.4155/fsoa-2017-0135

Dimova, D., Stumpfe, D., Hu, Y., and Bajorath, J. (2016). Analog series-based scaffolds: computational design and exploration of a new type of molecular scaffolds for medicinal chemistry. *Future Sci. OA* 2:FSO149. doi: 10.4155/fsoa-2016-0058

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., et al. (2017). The ChEMBL database in 2017. *Nucleic Acids Res.* 45, D945–D954. doi: 10.1093/nar/gkw1074

González-Medina, M., Prieto-Martínez, F. D., Naveja, J. J., Méndez-Lucio, O., El-Elimat, T., Pearce, C. J., et al. (2016). Chemoinformatic expedition of the chemical space of fungal products. *Future Med. Chem.* 8, 1399–1412. doi: 10.4155/fmc-2016-0079

Griffen, E., Leach, A. G., Robb, G. R., and Warner, D. J. (2011). Matched molecular pairs as a medicinal chemistry tool. *J. Med. Chem.* 54, 7739–7750. doi: 10.1021/jm200452d

Hu, Y., Stumpfe, D., and Bajorath, J. (2011). Lessons learned from molecular scaffold analysis. *J. Chem. Inf. Model.* 51, 1742–1753. doi: 10.1021/ci200179y

Kunimoto, R., Dimova, D., and Bajorath, J. (2017). Application of a new scaffold concept for computational target deconvolution of chemical cancer cell line screens. *ACS Omega* 2, 1463–1468. doi: 10.1021/acsomega.7b00215

Kunimoto, R., Miyao, T., and Bajorath, J. (2018). Computational method for estimating progression saturation of analog series. *RSC Adv.* 8, 5484–5492. doi: 10.1039/C7RA13748F

Larsson, J., Gottfries, J., Muresan, S., and Backlund, A. (2007). ChemGPS-NP: tuned for navigation in biologically relevant chemical space. *J. Nat. Prod.* 70, 789–794. doi: 10.1021/np070002y

Lewis, R., Guha, R., Korcsmaros, T., and Bender, A. (2015). Synergy maps: exploring compound combinations using network-based visualization. *J. Cheminform.* 7, 36. doi: 10.1186/s13321-015-0090-6

Lipinski, C., and Hopkins, A. (2004). Navigating chemical space for biology and medicine. *Nature* 432, 855–861. doi: 10.1038/nature03193

Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., et al. (2019). Ultra-large library docking for discovering new chemotypes. *Nature* 566, 224–229. doi: 10.1038/s41586-019-0917-9

Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605. Available online at: http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

Maggiora, G. M. (2006). On outliers and activity cliffs–why QSAR often disappoints. *J. Chem. Inf. Model.* 46, 1535. doi: 10.1021/ci060117s

Medina-Franco, J., Martinez-Mayorga, K., Giulianotti, M., Houghten, R., and Pinilla, C. (2008). Visualization of the chemical space in drug discovery. *CAD* 4, 322–333. doi: 10.2174/157340908786786010

Naveja, J. J., Cortés-Benítez, F., Bratoeff, E., and Medina-Franco, J. L. (2016). Activity landscape analysis of novel 5α-reductase inhibitors. *Mol. Divers.* 20, 771–780. doi: 10.1007/s11030-016-9659-x

Naveja, J. J., and Medina-Franco, J. L. (2015). Activity landscape sweeping: insights into the mechanism of inhibition and optimization of DNMT1 inhibitors. *RSC Adv.* 5, 63882–63895. doi: 10.1039/C5RA12339A

Naveja, J. J., and Medina-Franco, J. L. (2017). ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds. *F1000Res.* 6:Chem Inf Sci-1134. doi: 10.12688/f1000research.12095.2

Naveja, J. J., and Medina-Franco, J. L. (2018). Insights from pharmacological similarity of epigenetic targets in epipolypharmacology. *Drug Discov. Today* 23, 141–150. doi: 10.1016/j.drudis.2017.10.006

Naveja, J. J., Norinder, U., Mucs, D., López-López, E., and Medina-Franco, J. L. (2018a). Chemical space, diversity and activity landscape analysis of estrogen receptor binders. *RSC Adv.* 8, 38229–38237. doi: 10.1039/C8RA07604A

Naveja, J. J., Oviedo-Osornio, C. I., Trujillo-Minero, N. N., and Medina-Franco, J. L. (2018b). Chemoinformatics: a perspective from an academic setting in Latin America. *Mol. Divers.* 22, 247–258. doi: 10.1007/s11030-017-9802-3

Naveja, J. J., Vogt, M., Stumpfe, D., Medina-Franco, J. L., and Bajorath, J. (2019). Systematic extraction of analogue series from large compound collections using a new computational compound–core relationship method. *ACS Omega* 4, 1027–1032. doi: 10.1021/acsomega.8b03390

Oprea, T. I., and Gottfries, J. (2001). Chemography: the art of navigating in chemical space. *J. Comb. Chem.* 3, 157–166. doi: 10.1021/cc0000388

Osolodkin, D. I., Radchenko, E. V., Orlov, A. A., Voronkov, A. E., Palyulin, V. A., and Zefirov, N. S. (2015). Progress in visual representations of chemical space. *Expert Opin. Drug Discov.* 10, 959–973. doi: 10.1517/17460441.2015.1060216

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Prieto-Martínez, F. D., Gortari, E. F., Méndez-Lucio, O., and Medina-Franco, J. L. (2016). A chemical space odyssey of inhibitors of histone deacetylases and bromodomains. *RSC Adv.* 6, 56225–56239. doi: 10.1039/C6RA07224K

Probst, D., and Reymond, J.-L. (2018). Exploring drugbank in virtual reality chemical space. *J. Chem. Inf. Model.* 58, 1731–1735. doi: 10.1021/acs.jcim.8b00402

Reymond, J.-L. (2015). The chemical space project. *Acc. Chem. Res.* 48, 722–730. doi: 10.1021/ar500432k

Rosén, J., Lövgren, A., Kogej, T., Muresan, S., Gottfries, J., and Backlund, A. (2009). ChemGPS-NP(Web): chemical space navigation online. *J. Comput. Aided Mol. Des.* 23, 253–259. doi: 10.1007/s10822-008-9255-y

Stumpfe, D., Dimova, D., and Bajorath, J. (2016). Computational method for the systematic identification of analog series and key compounds representing series and their biological activity profiles. *J. Med. Chem.* 59, 7667–7676. doi: 10.1021/acs.jmedchem.6b00906

Uhlenbrock, N., Smith, S., Weisner, J., Landel, I., Lindemann, M., Le, T. A., et al. (2019). Structural and chemical insights into the covalent-allosteric inhibition of the protein kinase Akt. *Chem. Sci.* 10, 3573–3585. doi: 10.1039/c8sc05212c

Vogt, M., Yonchev, D., and Bajorath, J. (2018). Computational method to evaluate progress in lead optimization. *J. Med. Chem.* 61, 10895–10900. doi: 10.1021/acs.jmedchem.8b01626

Waddell, J., and Medina-Franco, J. L. (2012). Bioactivity landscape modeling: chemoinformatic characterization of structure-activity relationships of compounds tested across multiple targets. *Bioorg. Med. Chem.* 20, 5443–5452. doi: 10.1016/j.bmc.2011.11.051

Yonchev, D., Vogt, M., Stumpfe, D., Kunimoto, R., Miyao, T., and Bajorath, J. (2018). Computational assessment of chemical saturation of analogue series under varying conditions. *ACS Omega* 3, 15799–15808. doi: 10.1021/acsomega.8b02087

Zhao, Z., Leister, W. H., Robinson, R. G., Barnett, S. F., Defeo-Jones, D., Jones, R. E., et al. (2005). Discovery of 2,3,5-trisubstituted pyridine derivatives as potent Akt1 and Akt2 dual inhibitors. *Bioorg. Med. Chem. Lett.* 15, 905–909. doi: 10.1016/j.bmcl.2004.12.062

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Naveja and Medina-Franco. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 06 August 2019
doi: 10.3389/fchem.2019.00534

In silico Guided Drug Repurposing: Discovery of New Competitive and Non-competitive Inhibitors of Falcipain-2

Lucas N. Alberca[1], *Sara R. Chuguransky*[1], *Cora L. Álvarez*[2], *Alan Talevi*[1]* *and Emir Salas-Sarduy*[3]

[1] *Laboratory of Bioactive Compounds Research and Development (LIDeB), Department of Biological Sciences, Exact Sciences College, Universidad Nacional de La Plata, La Plata, Argentina,* [2] *Departamento de Biodiversidad y Biología Experimental, Facultad de Farmacia y Bioquímica, Facultad de Ciencias Exactas y Naturales, Consejo Nacional de Investigaciones Científicas y Técnicas, Instituto de Química y Fisico-Química Biológicas (IQUIFIB) "Prof. Alejandro C. Paladini", Universidad de Buenos Aires, Buenos Aires, Argentina,* [3] *Instituto de Investigaciones Biotecnológicas "Dr. Rodolfo Ugalde", Universidad Nacional de San Martín, CONICET, Buenos Aires, Argentina*

OPEN ACCESS

Edited by:
Simone Brogi,
University of Pisa, Italy

Reviewed by:
Marco Tutone,
University of Palermo, Italy
Jaewoo Kang,
Korea University, South Korea

***Correspondence:**
Alan Talevi
alantalevi@gmail.com

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *25 April 2019*
Accepted: *12 July 2019*
Published: *06 August 2019*

Citation:
Alberca LN, Chuguransky SR, Álvarez CL, Talevi A and Salas-Sarduy E (2019) In silico Guided Drug Repurposing: Discovery of New Competitive and Non-competitive Inhibitors of Falcipain-2. Front. Chem. 7:534. doi: 10.3389/fchem.2019.00534

Malaria is among the leading causes of death worldwide. The emergence of *Plasmodium falciparum* resistant strains with reduced sensitivity to the first line combination therapy and suboptimal responses to insecticides used for Anopheles vector management have led to renewed interest in novel therapeutic options. Here, we report the development and validation of an ensemble of ligand-based computational models capable of identifying falcipain-2 inhibitors, and their subsequent application in the virtual screening of DrugBank and Sweetlead libraries. Among four hits submitted to enzymatic assays, two (odanacatib, an abandoned investigational treatment for osteoporosis and bone metastasis, and the antibiotic methacycline) confirmed inhibitory effects on falcipain-2, with Ki of 98.2 nM and 84.4 μM. Interestingly, Methacycline proved to be a non-competitive inhibitor ($\alpha = 1.42$) of falcipain-2. The effects of both hits on falcipain-2 hemoglobinase activity and on the development of *P. falciparum* were also studied.

Keywords: malaria, *Plasmodium falciparum*, falcipain-2, drug repositioning, virtual screening, drug rescue, odanacatib, methacycline

INTRODUCTION

Despite decades of successful interventions aimed at reducing its incidence and mortality, malaria continues being one of global leading causes of death, being the main global cause globally in the 5- to 14-year-old population and the third cause among children below five (World Health Organization (WHO), 2017; Ritchie and Roser, 2018). The most recent estimates from the World Health Organization (WHO) report around 216 million cases and 445,000 related deaths worldwide in 2016 (Ritchie and Roser, 2018). The emergence of *Plasmodium falciparum* drug-resistant strains with reduced sensitivity to the first line artemisinin combination therapy and suboptimal response to insecticides used for vector management pose a threat to control interventions (Satimai et al., 2012; Ajayi and Ukwaja, 2013; Kisinza et al., 2017). Accordingly, novel therapeutic options are urgently required.

Falcipains are *P. falciparum* cysteine proteases involved in different processes of the erythrocytic cycle of the parasite, including hydrolysis of host hemoglobin and erythrocyte invasion and rupture. Four falcipains have so far been identified, with falcipain-2 and falcipain-3 constituting promising targets in the search for novel therapies due to their significant hemoglobinase capacity (Marco and Coterón, 2012; Bekono et al., 2018).

Drug repurposing involves finding novel medical uses for existing drugs, including approved, investigational, discontinued, and shelved therapeutics. Repurposing a drug has several advantages in comparison to *de novo* drug discovery, since the new therapeutic indication is built on already available pharmacokinetic, toxicological, and manufacturing data, thus leading to therapeutic solutions in a expedite manner (Ferreira and Andricopulo, 2016; Corsello et al., 2017). While the best-known examples of successful repurposing have been serendipitous or arose from intelligent exploitation of side effects (Corsello et al., 2017; Talevi, 2018), the drug discovery community has recently focused on systematic, large-scale repurposing efforts, including the use of genomic tools, and *in silico* and *high-throughput* screening (Jin and Wong, 2015; Talevi, 2018; Yella et al., 2018). The Virtual lock-and-key approach (Lauria et al., 2011, 2014; Tutone et al., 2017) and the BIOlogical Target Assignation method (Lauria et al., 2014) can be mentioned among many other interesting examples of the use of computational resources to deepen the rational basis of drug repurposing programs.

Here, we have implemented a computer-aided drug repurposing campaign to discover new inhibitors of falcipain-2. Four hits were acquired and tested against the enzyme, with two of them confirming inhibitory activity. The abandoned drug odanacatib displayed competitive inhibition, while the antibiotic methacycline also showed inhibitory effects through non-competitive inhibition.

MATERIALS AND METHODS

Dataset Collection

P. falciparum falcipain-2 inhibitors were compiled from literature. A total of 515 compounds previously assayed against falcipain-2 were collected from over 20 original articles, conforming the dataset used for model calibration and validation (Domínguez et al., 1997; Chiyanzu et al., 2003; Shenai et al., 2003; Desai et al., 2004, 2006; Greenbaum et al., 2004; Fujii et al., 2005; Goud et al., 2005; Micale et al., 2006; Valente et al., 2006; Biot et al., 2007; Chipeleme et al., 2007; Li et al., 2009; Hans et al., 2010; Praveen Kumar et al., 2011; Shah et al., 2011; Huang et al., 2012; Luo et al., 2012; Conroy et al., 2014; Ettari et al., 2014; Jin et al., 2014; Wang et al., 2014; Weldon et al., 2014; Bertoldo et al., 2015; Mundra and Radhakrishnan, 2015a,b; Sharma et al., 2015; Singh et al., 2015; Schmidt et al., 2016). Such compounds were labeled as ACTIVE or INACTIVE according to their reported inhibitory data. The ACTIVE category included compounds with $IC_{50} \leq 5\,\mu M$, plus compounds with a percentage of inhibition $\geq$ 50% against the enzyme at $10\,\mu M$ or $\geq$ 80% at $20\,\mu M$ (when a single-point inhibition assay was reported). When none of the previous conditions were met, the compound was labeled as INACTIVE. Considering such criteria, the dataset includes 122 active compounds and 393 inactive compounds. Such dataset was curated using the standardization tool available in Instant JCHEM v. 17.2.6.0 (Chemaxon). The molecular diversity of the whole dataset and within each category can be appreciated in the heatmap displayed in **Figure 1**, which shows, for every compound pair, the Tanimoto distance computed using ECFP_4 molecular fingerprints. The heatmap was built using Gitools v. 2.3.1 (Perez-Llamas and Lopez-Bigas, 2011) and Tanimoto distances were calculated using ScreenMD—Molecular Descriptor Screening v. 5.5.0.1 (ChemAxon). The dataset is included as **Supplementary Data Sheet 1**.

Dataset Partition

It has been observed that rational/representative splitting of datasets into training and test sets tends to produce models with better predictivity (Golbraikh et al., 2003; Leonard and Roy, 2006; Martin et al., 2012). In the present study, a representative sampling procedure was thus used to divide the datasets into: (a) a training set, that was used to calibrate the models and; (b) a test set, that was used to independently assess model predictivity. Such representative partition of the dataset resulted from a serial combination of two clustering procedures. First, we have used the hierarchical clustering method included in LibraryMCS software (version 17.2.13.0–ChemAxon), which relies on the Maximum Common Substructure (MCS). A compound from each of the resulting cluster was randomly chosen and used as a seed to perform a non-hierarchical clustering using the k-means algorithm, as implemented in Statistica 10 Cluster Analysis module (Statsoft). Hierarchical clustering allowed deciding on an initial partition of n molecules into k groups, and this

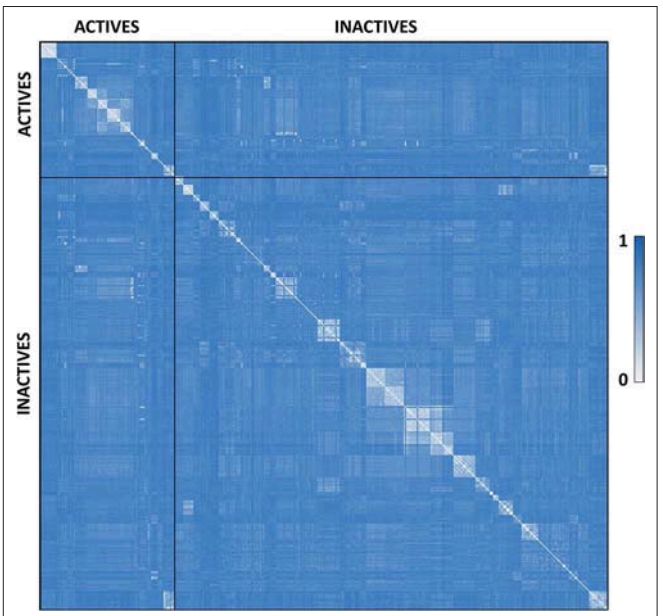


FIGURE 1 | Dissimilarity heatmap of the whole dataset. Light areas indicate high similarity between the compared compounds while dark blue areas indicate low similarity between the compared compounds.

preliminary clustering was then optimized through the non-hierarchical procedure, as suggested by Everitt et al. (2011). We have previously used this combined approach for representative dataset partitioning, with good results (Alberca et al., 2016, 2018; Gantner et al., 2017). The clustering procedure was performed separately for the ACTIVE and INACTIVE categories.

75% of the compounds in each cluster of the ACTIVE category were kept for the training set (making a total of 91 compounds); an equal number of compounds were taken from the INACTIVE category clusters (23% of each INACTIVE cluster). We have under-sampled the INACTIVE category, so that a balanced training sample (comprising an identical proportion of active and inactive examples) was obtained and model bias toward predicting the larger category was avoided. The remaining 31 active and 302 inactive compounds were assigned to test set (333 compounds in total), which was later used for external validation of the models.

Molecular Descriptor Calculation and Modeling Procedure

3,668 conformation-independent descriptors were computed with Dragon 6.0 software. A random subspace-based method was applied to obtain 1,000 descriptor subsets of 200 potential independent variables each. In the random subspace approach, the molecular descriptors are randomly sampled, and each model is trained on one subset of the feature space (Yu et al., 2012; El Habib Daho and Chikh, 2015); as a result, individual models do not over-focus on features that display high explanatory power in the training set.

A dummy variable (class label) was used as dependent variable. It was assigned observed values of 1 for compounds within the ACTIVE category and observed values of 0 for compounds in the INACTIVE one. Using a Forward Stepwise procedure and a semi-correlation approach (Toropova and Toropov, 2017), 1,000 linear classifiers were obtained, one from each of the random subsets of features. In order to avoid overfitting, only one molecular descriptor every 12 training instances was allowed into each model, with no more than 12 independent variables per model. Also, a maximal Variance Inflation Factor of 2 was tolerated. No descriptor with regression coefficient with p-value above 0.05 was allowed into the model. R language and environment was used for all data analysis. The R package data table (https://cran.r-project.org/package=data.table) was used to handle datasets.

The robustness and predictive ability of the models were initially estimated through randomization and Leave-Group-Out cross-validation tests. In the case of randomization, the class label was randomized across the compounds in the training set. The training set with the randomized dependent variable was then used to train new models from the descriptor selection step. Such procedure was repeated 10 times within each descriptor subset and the 95% confidence interval was built around the mean accuracy of the randomized models. It is expected that the randomized models will perform poorly compared to the real ones. Regarding the Leave-Group-Out cross-validation, random stratified subsets comprising 10 active compounds and 10 inactive compounds were removed from the training set in each cross-validation round, and the model was regenerated using the remaining compounds as training examples. The resulting model was used to predict the class label for the 20 removed compounds. The procedure was repeated 10 times, with each of the training set compounds removed at least once. The results were informed as the average percentage of good classifications (accuracy) across the folds, and this was compared to the accuracy of the model for the original training set and also, as advised by Gramatica (2013), to the No-Model error rate or risk (NOMER%), i.e., the error provided in absence of model:

$$NOMER\% = \frac{(n - nm)}{n} \times 100$$

where n is the total number of objects and nm is the number of objects of the most represented class.

Finally, the predictivity of each individual model was assessed through external validation, using the 333-compound test set that was already described in section Dataset Collection. A diversity of statistical parameters commonly used to assess the performance of classificatory models (Roy and Mitra, 2011; Gramatica, 2013) were estimated for both the training and test sets: sensitivity (Se, i.e., true positive rate), specificity (Sp, i.e., true negative rate), accuracy (Acc, i.e., overall percentage of good classifications), positive and negative predictivity and the F-measure, which is defined as follows (Roy and Mitra, 2011):

$$F - measure = \frac{2 \times Se \times (1 - Sp)}{Se + (1 - Sp)}$$

Ensemble Learning

Classifier ensembles are known to provide better generalization and accuracy than single model classifiers (El Habib Daho and Chikh, 2015; Carbonneau et al., 2016; Min, 2016). Here, we have used two retrospective virtual screening campaigns to assess the performance of individual classifiers and classifier ensembles. As described in the next subsection, the first retrospective virtual screen allowed assessing the performance of individual classifiers and provided the basis to decide which individual models would be selectively combined in the model ensemble and how they would be combined. The second retrospective virtual screen served to the sole purpose of assessing the performance of the chosen model ensemble.

The best individual classifiers were selected and combined using the area under the ROC curve metric (AUC ROC) in the first retrospective screen as criterion of performance. To choose the ideal number of models to be included in the ensemble, systematic combinations of the 2 to 100 best performing classifiers were analyzed (the two best-performing models were combined, then the three best-performing models, the four best-performing models, and so on up to a total of 100 models included in the ensemble). Four combination schemes were applied to obtain a combined score: MIN operator (which returns the minimum score among the individual scores of the combined models); Average Score; Average Ranking and; Average Voting. Voting was computed according to the equation previously used

by Zhang and Muegge (2006). AUC ROCs were obtained with the pROC package (Robin et al., 2011); the Delong method was used to statistically compare the AUC ROCs. BEDROC and RIE (1%) were also computed (Truchon and Bayly, 2007). For that purpose, we resorted to the R package enrichvs (enrichment assessment of virtual screening approaches; Yabuuchi et al., 2011) and the online tool ROCKER (Lätti et al., 2016).

Retrospective Screening Campaigns

Through simulated ranking experiments, Truchon and Bayly (2007) demonstrated that the AUC ROC metric is dependent on the ratio of active compounds/inactive compounds, and the standard deviation of the metric converges to a constant value when small yields of actives (Ya) of the screened library are used (Ya below 0.05 seem to provide more robust results). Reasonably small Ya also ensures that the saturation effect is constant or absent. A high number of decoys (around 1,000 or higher) and a small Ya contribute to a controlled statistical behavior (Truchon and Bayly, 2007). Accordingly, to estimate the enrichment performance of our models and model ensembles in a real virtual screening scenario, we have performed retrospective virtual screening experiments. For that purpose, we have seeded known active compounds among a large number of decoys obtained with the help of the Directory of Useful Decoys Enhanced (DUD-E; Mysinger et al., 2012), a widely used benchmarking tool which allows the obtention of putative inactive compounds paired to known active compounds by physicochemical properties (e.g., molecular weight, logP, number of rotatable bonds, among others), but topologically dissimilar to such active compounds. In this way, two chemical libraries for such pilot screens were obtained. The first one, that we will call DUDE-A, was compiled by using the active compounds from the test set as queries in the DUD-E website. Such active compounds were later dispersed among the so-obtained paired decoys (putative inactive compounds). As a result, DUDE-A contained 31 known active compounds dispersed among 1500 DUD-E decoys and displayed a Ya of 0.020. DUDE-A was used to estimate the performance of the individual models in a virtual screening experiment and to choose the best individual models that would be included in the ensemble (i.e., to train the ensembles). It was also applied to choose which score threshold would be applied in prospective virtual screening campaigns. A second library, called DUDE-B, was obtained to validate the ensemble that showed the best performance in the DUDE-A screen. For that purpose, we compiled from literature 33 recently reported active compounds against FP-2 (IC50 $\leq 5\,\mu M$; Nizi et al., 2018; Stoye et al., 2019). The DUDE-B library was generated by merging these 33 active compounds with 4,337 decoys from the DUD-E website. The calculated Ya for DUDE-B is about 0.007.

Building Positivity Predictive Value Surfaces and Choosing an Adequate Score Threshold Value

A practical concern when implementing *in silico* screens involves estimating the actual probability that a predicted hit will confirm its activity when submitted to experimental testing (Positive Predictive Value, PPV). Estimation of such probability is however precluded due to its dependency on the Ya of the screened library, which is not known *a priori*:

$$PPV = \frac{Se\ Ya}{Se\ Ya + \left(1 - Sp\right)\left(1 - Ya\right)}$$

where *Se* represents the sensitivity associated to a given score cutoff value and *Sp* represents the specificity. The former equation was applied to build PPV surfaces. In order to choose an optimal cutoff value to select predicted hits in prospective virtual screening experiments, 3D plots showing the interplay between PPV, the Se/Sp ratio and Ya were built for each individual model and for each model ensemble. This approach has recently been reported by our group (Alberca et al., 2018). Using DUDE-A (described in previous subsection), Se and Sp were computed in all the range of possible cutoff score values. Though there is no guarantee that the Se and Sp associated to each score value for DUDE-A will be the same when applying the classifiers to other libraries, e.g., in the prospective virtual screening campaign, since controlled statistical behavior is observed for database sizes of 1,000 compounds or more and Ya below 0.05, we can reasonably assume that the ROC curve and derived metrics will be similar when applying the models to classify other large chemical databases with low Ya. Taking into consideration that in real virtual screening applications Ya is ignored a priori but invariably low, Ya was varied between 0.001 and 0.010. The R package plotly (https://cran.r-project.org/package=plotly) was used to obtain all the PPV graphs. Visual analysis of the resulting PPV surfaces allowed us to select a score threshold value with a desired range of PPV.

Prospective Virtual Screening

Based on visual inspection of the resulting of PPV graphs, we have applied in a prospective virtual screen an 11-model ensemble using the MIN operator to combine individual classifiers. Based on PPV surface analysis, we chose a score threshold that provides a PPV $\geq$ 20% at Ya = 0.01.

We have used the 11-model ensemble to screen two databases: (a) DrugBank 4.0, an online database containing extensive information about the US Food and Drug Administration (FDA) approved, experimental, illicit and investigational drugs (Law et al., 2014); (b) SWEETLEAD, a curated database of drugs approved by other international regulatory agencies, plus compounds isolated from traditional medicinal herbs and regulated chemicals (Novick et al., 2013). Both databases were curated using Standardizer 16.10.10.0 (ChemAxon). The following actions were applied to obtain homogeneous representations of the molecular structure for the subsequent virtual screen: (1) Strip salts; (2) Remove Solvents; (3) Clear Stereo; (4) Remove Absolute Stereo; (5) Aromatize; (6) Neutralize; (7) Add Explicit Hydrogens; and (8) Clean 2D. Duplicated structures were removed using Instant JCHEM v. 17.2.6.0. Four hits were selected for experimental evaluation, using the following criteria: (a) no previous report of falcipain-2 inhibition; (b) availability through local suppliers; (c) cost. Methacycline, benzthiazide, and bendroflumethiazide were

acquired from Sigma-Aldrich. Odanacatib (99% HPLC) was acquired from AK Scientific (Y0388).

Molecular Docking

To gain insight into the possible mode of action of the active hits, we studied their possible interactions with falcipain-2 by docking simulations. The structure of the enzyme was obtained from the Protein data bank. We retrieved the available experimental structure of the target in complex with an inhibitor, the epoxysuccinate E64, that shows 2.9 Å of resolution (PDB code: 3BPF; Kerr et al., 2009). Among the four chains crystalized, we selected the B chain for the simulations. We used AutoDockTools4 software to remove the inhibitor and the crystallographic water molecules off the pdb file, and to add the hydrogen atoms of the protein.

Autodock4 was used for docking simulations. The docking software and conditions were selected based on the previous investigation by Mugumbate et al. (2013). The authors showed the capacity of this software to replicate the experimental pose of E64 in the re-docking experiment and to identify known inhibitors from non-inhibitors through the docking scores. Our own results of the re-docking E64 into the binding site of FP-2 were similar to such previous investigations, since Autodock4 was able to reproduce the experimental pose with a RMSD value of 1.84 Å.

Odanacatib was docked into the active site of the enzyme. Calculations were conducted with a grid of 40X40X40 grid points, centered on the experimental ligand E64 (coordinates: −36.75, 31.05, −47.07 in x, y, and z, respectively) and with a spacing between grid point of 0.39 Å. We used the default Autodock4 parameters for a population of 150, 50 genetic algorithm runs, 2.5×10^6 evaluations and 27,000 maximum generations.

Regarding non-competitive inhibitor methacycline, we first used DoGSiteScore server (Volkamer et al., 2012) to detect possible binding pockets in the protein. The server proposed two regions of binding, besides the known catalytic binding site of the enzyme. Methacycline was docked in both regions and the best results were achieved in the area delimited by CYS39, SER41, TRP43, GLU67, GLN68, LEU70, VAL71, ASP72, CYS73, SER74, PHE75, ASN77, TYR78, GLY79, CYS80, TYR106, VAL107, SER108, ASP109, ALA110, PRO111, ASN112. The simulation was conducted in the conditions described before for odanacatib and E64, except for the position of the grid, which was centered on the side chain on VAL71, specifically in the carbon atom defined as CG1.

Falcipain-2 Expression and Refolding

Falcipain-2 (MEROPS ID: C01.046) was expressed as inclusion bodies in BL21(DE3) *Escherichia coli* strain, purified by IMAC under denaturing conditions (final purity: 91%) and refolded to active enzyme as previously described (Pradines et al., 2001).

Falcipain-2 Kinetic Assay

Falcipain-2 activity was assayed fluorometrically with Z-LR↓AMC (Bachem) as substrate in 100 mM acetate buffer pH 5.5 containing 5 mM DTT and 0.01% Triton X-100, as this is expected to increase enzyme stability and reduce the number of false positives (Jadhav et al., 2010). Assays (final reaction volume ∼80 μL) were performed at 30°C in solid black 384-well plates (Corning) at fixed enzyme concentration (3.3 nM). Except stated otherwise, fluorogenic substrate was added at final concentration of 5 μM (∼1 x K_M) to match balanced assay conditions (Copeland, 2005). The release of 7-amino-4-methylcoumarin was monitored continuously for 60 min with a FilterMax F5 Multimode Microplate Reader (Molecular Devices) using standard 360 nm excitation and 465 nm emission filter set. Enzyme activity was estimated as the slope of the linear region of the resultant progress curves. Under the described conditions, falcipain-2 activity showed no significant changes in the presence of DMSO (0–8%) and the Selwyn test (Selwyn, 1965) indicated that enzyme remained stable during the assay.

Falcipain-2 Inhibition Assay

1 μL of each compound (2.5 mM in DMSO), N-(trans-epoxysuccinyl)-l-leucine 4-guanidinobutylamide (E-64, Sigma-Aldrich) (10 μM in DMSO) or DMSO were dispensed into each well. Then, 40 μL of activity buffer containing falcipain-2 (6.6 nM) were added to each well, plates were homogenized (30 seg, orbital, medium intensity) and each well subjected to a single autofluorescence read (exc/ems = 360/465 nm). Plates were incubated in darkness for 15 min at 30°C and then 40 μL of Z-LR-AMC (10 μM in assay buffer) were added to each well to start the reaction. After homogenization (30 seg, orbital, medium intensity), the fluorescence of AMC (exc/ems = 360/465 nm) was acquired kinetically for each well (12 read cycles, one cycle every 300 s). Fluorescence measurements were used to determine the slope (dF/dt) of progress curves by linear regression and inhibition percentage (%Inh) was calculated for each compound according to:

$$\%Inh = 100 \cdot [1 - (dF/dt^{WELL} - \mu^{C-})/(\mu^{C+} - \mu^{C-})]$$

where dF/dt^{WELL} represents the slope of each compound well and μ^{C+} and μ^{C-} the average of falcipain-2 + DMSO (no-inhibition) and substrate (no-enzyme) controls, respectively.

Compounds were re-tested in a dose-response manner (final concentration ranging from 375 μM to 44.7 pM) using identical assay conditions. 6 μL of compounds stock (10 mM in DMSO), E-64 (10 μM in DMSO), and DMSO were added to the first wells (column 1), followed by addition of 34 μL of activity buffer. After addition of 20 μL of buffer to subsequent wells, 24 serial 2-fold dilutions were made horizontally. Then, 40 μL of activity buffer containing falcipain-2 (6.6 nM) were added to each well, except for those corresponding to C-; completed with 40 μL of activity buffer. After homogenization, incubation, and autofluorescence measurement, 20 μL of Z-LR-AMC substrate (20 μM in activity buffer) were added. Data collection and processing were performed exactly as described above. At the concentration tested, no significant autofluorescence (360/465 nm) was apparent for the investigated compounds.

Percentage of falcipain-2 residual activity was calculated for each condition according to:

$$\%Res.Act = 100 \cdot \left[\frac{(dF/dt^{WELL} - \mu^{C-})}{(\mu^{C+} - \mu^{C-})} \right]$$

Half-maximal inhibitory concentration (IC_{50}) and Hill slope parameters were estimated by fitting experimental data from dose-response curves to the four-parameter Hill equation by using GraphPad Prism program (version 5.03).

Determining Reversibility, Mode of Inhibition, and Ki

Reversibility and time dependence of falcipain-2 inhibition by investigational compounds was assayed as previously described (Morrison, 1969). In brief, odanacatib (15 μM) and falcipain-2 (330 nM) were incubated at 30 °C for 60 min in activity buffer. Two microliters of the mix were rapidly added to 200 μL of Z-LR-AMC (5 μM in activity buffer) pre-incubated at the same temperature. Immediately after mixing, AMC fluorescence (λexc/ems = 355/460 nm, sensitivity = 550 V) was continuously monitored every second using a thermostated (30°C) Aminco Bowman Series 2 spectrofluorometer (Thermo Spectronic). In the case of the methacycline, the final inhibitor concentration in the mixture with falcipain-2 was 330 μM. For falcipain-2 control, the equivalent volume of DMSO vehicle was pre-incubated with the enzyme. To determine the kinetics of inhibition onset, falcipain-2 (3.3 nM final concentration) was added to 200 μL of reaction mix (previously tempered at 30°C) containing activity buffer, odanacatib (0.15 μM) and Z-LR-AMC (5 μM). Immediately after mixing, AMC release was monitored as indicated above. For methacycline (33 μM final concentration in the reaction mix), the experiment was exactly the same.

The identification of the mode of inhibition was performed as indicated previously. For odanacatib, falcipain-2 activity was determined for at least six different substrate concentrations (ranging from 62.5 to 1.95 μM) in the absence and presence of three fixed doses of inhibitor: 0.15, 0.5, and 2.5 μM. Data were re-arranged to estimate percentage of falcipain-2 residual activity for each condition and the values for IC_{50} and Hill slope were estimated by fitting experimental data to the four-parameter Hill equation by using GraphPad Prism. To estimate Ki, kinetic data were arranged in the form of Michaelis curves (dF/dt vs. $[Z\text{-}LR\text{-}AMC]_0$) and globally fitted to the competitive inhibition equation present in GraphPad Prism (version 5.03). Finally, to estimate Ki by using the tight-binding inhibition approach (Morrison, 1969), data were transformed to fractional velocity vs. inhibitor concentration and re-analyzed by global fitting to the Morrison equation by using GraphPad Prism (version 5.03).

To identify the mode of inhibition of methacycline, dose-response curves (0–625 μM) were performed as described above at six different substrate concentrations (ranging from 1 to 50 μM) and fitted as indicated above to estimate the values of IC_{50} and Hill slope. Finally, kinetic data were arranged in the form of Michaelis curves (dF/dt vs. $[Z\text{-}LR\text{-}AMC]_0$) and globally fitted to the mixed inhibition equation present in GraphPad Prism (version 5.03) for the simultaneous estimation of α and Ki.

Determining the Sensitivity of Methacycline Inhibition to RedOx Potential

The inhibitory activity of decreasing concentrations (375 μM−91.6 pM) of methacycline were determined in activity buffer containing DTT (0.1–10 mM) or L-cysteine (0.1–10 mM) as indicated above. Resultant dose-response curves were fitted as previously indicated to estimate the values of IC_{50} and Hill slope.

Densitometric Estimation of the Inhibition of Falcipain-2 Hemoglobinase Activity

Increasing concentrations of methacycline (200 μM, 500 μM, and 1 mM) and odanacatib (0.5, 5, and 50 μM) were preincubated with falcipain-2 (132 nM) for 30 min at 37°C in buffer 100 mM NaAc, 10 mM DTT pH 5.5. Then, human hemoglobin (H7379, Sigma-Aldrich) was added to a final concentration of 100 μg/mL to initiate reaction (final assay volume = 50 μL). E64 (10 μM) and DMSO were used as negative and positive controls, respectively. Also, a blank (no falcipain-2) control was included. In all cases, the final concentration of DMSO was 10%. Mixes were incubated without agitation for 3 h at 37°C to allow the enzymatic reaction to proceed. Then, reactions were stopped by addition of 15 μL of 5xSDS–PAGE sample buffer + 7.5 μL of DTT (1 M) and boiled for 5 min. Samples (22.5 μL, equivalent to 1.5 μg of hHb) were electrophoretically resolved by SDS-PAGE on a 15% acrylamide gel and Coomassie stained. The amount of undegraded hHb, observed as a doublet of around 15 kDa, was estimated densitometrically by using ImageJ 1.38d software (Nation al Institutes of Health, USA).

Evaluation of Antiparasitic Activity

Human erythrocytes were obtained from volunteer donors with a procedure approved by CEIC (Committee for Ethics on Clinical Investigation, Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires EXP-UBA: 0048676/2017). Human erythrocytes infected with the NF54 strain of *P. falciparum* were cultivated in RPMI 1640 medium supplemented with 0.5% albumax II (Invitrogen), 22 mM glucose, 25 mM HEPES, 0.65 mM hypoxanthine, and 50 mg/mL gentamicin. Cultures were maintained at 37°C by routine passage at 5% hematocrit with a maximum parasitemia of 5% in a 90% N_2/5% O_2/5% CO_2 atmosphere as previously described (Alvarez et al., 2014).

When needed, ring-stage parasites were synchronized by using sorbitol treatment (Aley et al., 1986). After 24 h synchronization, cultures of infected erythrocytes at trofozoite-stage were treated with various concentrations of odanacatib, methacycline and E-64 for 48 h. Briefly, 100 μL of synchronous trophozoite-stage infected erythrocytes cultures were plated in 96-well at 4% hematocrit and 1% parasitemia. 100 μL of odanacatib (200, 20, or 2 μM), methacycline (1,000, 100, or 20 μM), E-64 (50, 10, or 2 μM), DMSO (vehicle control), or RPMI 1640 medium (control) were dispensed into each well to achieve final hematocrit of 2%, 0.5% parasitemia and the final concentration of each compound tested in a final volume of 200 μL. Parasitemia was evaluated by light microscopy counting infected forms of the parasite (ring, trophozoite, and

schizont-stages) in a thick blood smear stained with Giemsa. A total of ∼1,500 erythrocytes distributed in at least 15 random microscopic fields were evaluated of each smear and the parasitemia was calculated as (infected erythrocytes / total erythrocytes)*100. Then, each treatment was normalized to control parasitemia and expressed as percentage.

Significance was determined using one-way Analysis of Variance followed by a Tukey Multiple Comparison Test. Computations were carried out using PRISM statistical software (GraphPad Software, Inc., version 6). A p-value < 0.05 was considered significant. The number of determinations (n) for independent preparations (N) are indicated.

RESULTS

A ligand-based virtual screening approach was used to discover falcipain-2 inhibitors. 1,000 individual linear classifiers were obtained by applying a random subspace approximation on a pool of more than 3,000 Dragon molecular descriptors. The individual models were internally and externally validated.

Results of the internal validation are shown in **Table 1**. Regarding the Leave-Group-Out, results for each individual classifier are informed as the average accuracy across the folds, which is in all cases above 80% and close to the correspondent accuracy on the training set, suggesting the models are robust. Since the proportions of the active and inactive compounds in the training set are identical (as are in each of the Leave-Group-Out folds) the correspondent NOMER% (associated to random classification) is 50%, well below the behavior of the models in the cross-validation.

Regarding the randomization results, **Table 1** shows the 95% confidence interval around the mean accuracy of the randomized models. As expected, the accuracy of the randomized models is in all cases much below the accuracy of the true (non-randomized) models, and very close to the NOMER%, suggesting a low probability of chance correlations for the true models.

External validation was performed using the 333-compound independent test set. The results are summarized in **Table 2**.

TABLE 1 | Results of the internal validation procedures for the best 11 individual classifiers.

Model	Training set	LGO	Randomization	
	Acc (%)	Average Acc (%)	Confidence interval 95%	
594	84.615	83.740	53.882	61.283
516	88.462	86.712	56.623	63.267
477	84.615	83.835	56.158	64.501
975	87.363	86.374	53.309	62.844
244	90.110	88.712	55.312	62.160
504	85.714	83.765	55.978	61.824
870	86.264	82.753	56.962	65.126
154	86.813	85.168	54.677	62.246
764	84.615	83.080	53.273	62.112
564	85.714	82.636	54.814	62.000
80	86.264	85.378	56.980	61.921

In general, the individual classifiers show an acceptable performance. Due to the unbalanced nature of the test set (31 active compounds and 302 inactive ones) in comparison to the test set (91 active and 91 inactive compounds) some of the differences in the statistical parameters of the training and test sets are to be expected (i.e., decreased sensitivity in the test set, sharp drop in the positive predictivity and concomitant increase in the negative predictivity).

The best individual model included the following features:

Model 594

Class = −0.48333 + 0.38415*SM08_AEA(bo) - 6.50601*SpPosA_A - 0.12786*C-005 + 0.35800*B05[N-N] + 0.13459*nR=Cs + 0.21576*CATS2D_02_DD + 0.25881*nS(=O)2 - 0.33510*B03[O-S] - 0.07816*N-072 + 0.16832*B06[C-S]

Wilks' Lambda:.45705 approx. $F_{(10,\ 171)} = 20.314$ $p < 0.0000$.

Dragon's nomenclature for the molecular descriptors has been kept in the previous expression. SM08_AEA(bo) corresponds to the spectral moment of order 8 from augmented edge adjacency matrix weighted by bond order; SpPosA_A is the

TABLE 2 | Statistical parameters of the best individual classifiers, for both the training and test sets.

Model	Se	Sp	Acc	F-measure	Positive predictivity	Negative predictivity
Training set						
Model 594	0.88	0.81	0.85	0.31	0.82	0.87
Model 516	0.89	0.88	0.88	0.21	0.88	0.89
Model 477	0.86	0.84	0.85	0.28	0.84	0.85
Model 975	0.89	0.86	0.87	0.25	0.86	0.89
Model 244	0.92	0.88	0.90	0.21	0.88	0.92
Model 504	0.87	0.85	0.86	0.26	0.85	0.87
Model 870	0.89	0.84	0.86	0.28	0.84	0.88
Model 154	0.88	0.86	0.87	0.25	0.86	0.88
Model 764	0.84	0.86	0.85	0.24	0.85	0.84
Model 564	0.85	0.87	0.86	0.23	0.87	0.85
Model 80	0.89	0.84	0.86	0.28	0.84	0.88
Test set						
Model 594	0.84	0.80	0.80	0.32	0.30	0.98
Model 516	0.77	0.82	0.82	0.29	0.31	0.97
Model 477	0.71	0.84	0.83	0.26	0.31	0.97
Model 975	0.87	0.80	0.81	0.32	0.31	0.98
Model 244	0.84	0.77	0.78	0.36	0.28	0.98
Model 504	0.81	0.81	0.81	0.30	0.31	0.98
Model 870	0.81	0.81	0.81	0.31	0.30	0.98
Model 154	0.71	0.82	0.81	0.29	0.29	0.96
Model 764	0.74	0.83	0.83	0.27	0.32	0.97
Model 564	0.77	0.84	0.83	0.27	0.33	0.97
Model 80	0.81	0.75	0.76	0.38	0.25	0.97

The default score cutoff value (0.5) was used to discriminate between active and inactive compounds and estimate the parameters. Note that this score has been later optimized to obtain improved Se/Sp relationships.

normalized spectral positive sum from adjacency matrix; C-005 refers to the frequency of CH3X groups where X indicates an electronegative atom (O, N, S, P, Se, halogens); B05[N-N] indicates the presence/absence of the N – N pair at topological distance 5; nR=Cs refers to the number of aliphatic secondary C(sp2); CATS2D_02_DD is the CATS2D Donor-Donor at lag 02; nS(=O)2 symbolizes the number of sulfones; B03[O-S] indicates the presence/absence of O – S pair at topological distance 3; N-072 refers to frequency of the atom-centered fragment RCO-N< / >N-X=X; B06[C-S] denotes the presence/absence of a C – S pair at topological distance 6. The molecular descriptors associated to presence/absence or frequency of a given feature indicate differences in the frequencies at which such features appear in the active and inactive class of the training set. Those descriptors associated in the model to a positive weighting coefficient (B05[N-N], nR=Cs, nS(=O)2, and B06[C-S]) show that such feature is more frequent in the active compounds than in the inactive ones. In contrast, the descriptors associated to a negative coefficient (SpPosA_A, C-005, and N-072) indicate that such features appear more frequently in the compounds of the inactive class than in the ones of the active class. CATS2D_02_DD is a two-dimensional Chemically Advanced Template Search descriptor similar to a pharmacophore pair (Reutlinger et al., 2013), but considering topological distances between the pharmacophore points instead of geometrical distances. Here, the descriptor suggests that two H-bond donors at a topological distance of two are a desirable feature in falcipain-2 inhibitors.

The physicochemical interpretation of the two descriptors associated to spectral moments of a topological matrix (SM08_AEA(bo) and SpPosA_A) is less immediate.

An augmented edge adjacency matrix aE(w) is a symmetric square matrix that can be derived from an edge-weighted molecular graph, for any weighting scheme w. The elements from such matrix [aEw]ij take values of 1 if i and j are adjacent edges/bonds, values wi if i equals j (that is, for the elements in the diagonal) and values of 0 otherwise (i.e., for non-diagonal elements corresponding to non-adjacent edges; Liu et al., 2018). The *k*th spectral order μ_k of a topological matrix M can be defined as:

$$\mu_k = \mathrm{tr}(M^k)$$

where k is the power of the matrix and tr is its trace, i.e., the sum of the diagonal elements (Estrada, 1996). The kth spectral moment of the edge adjacency matrix has a simple graph theoretical interpretation (Estrada, 1996): it is the sum of all self-returning walks of length k in the line graph of the molecular graph, beginning and ending with the same vertex. It may then be appreciated that the value of such descriptor would be highly influenced by the presence of ring systems and the nature of such cycles (e.g., fused rings). Since the considered augmented edge matrix is weighted by the bond order, the presence of double and triple bonds and aromatic systems will tend to increase the value of the descriptor. Generally speaking, active examples in our training set tend to display higher values of SM08_AEA(bo).

Regarding SpPosA_A, it denotes the normalized sum of positive eigenvalues of the adjacency matrix. Its value diminishes with increase branching, with greater emphasis in terminal rather than in central branching (Balaban et al., 1991).

The 11 best individual models and a brief description of the descriptors included in them have been listed as **Supplementary Data Sheet 2**.

For a more challenging and realistic simulation, the enrichment behavior of the individual models was studied through a retrospective virtual screen on DUDE-A library, where a small proportion of active compounds (31) was dispersed among a high number (1500) of putative decoys. Initially, we compute the area under the Receiver Operating Characteristic curve (AUC ROC) to assess the classificatory performance of the models. 100, 93.4 and 3.1% of the individual classifiers displayed AUC ROCs above 0.8 for the training set, the test set and the DUD-A library, in that order. 85.8% of the individual models achieved an AUC ROC > 0.90 for the training set, whereas only one of the models (named model 975) got an AUC ROC above 0.9 for the test set, none of the models achieved an AUC ROC above 0.9 for the DUDE-A library. This suggests that our random subspace approach has been successful in finding individual classifiers with good explanatory and predictive power, but also that some degree of overfitting may also be present. Results also suggests that the retrospective screening experiment on DUDE-A is the more challenging tasks for the classifiers. **Table 3** shows the 11 individual classifiers that showed the best performance on the DUDE-A library, along with their AUC ROC, BEDROC, and RIE values.

Whereas, the performance of the best individual classifiers was quite satisfactory, we explored ensemble learning approaches to obtain meta-classifiers with improved accuracy and a more robust behavior. **Figure 2** shows the AUC ROC values (DUDE-A) obtained when systematically combining between the 2 and

TABLE 3 | Values of the AUC ROC, BEDROC, and RIE metrics for the 11 individual classifiers that displayed the best performance on the DUDE-A library.

Model	AUC ROC			BEDROC ($\alpha = 20$)	RIE 1%
	Training set	Test set	DUDE-A	DUDE-A	DUDE-A
594	0.9162	0.8819	0.8529	0.2565	0.0000
516	0.9396	0.8819	0.8508	0.3150	0.0000
477	0.9123	0.8910	0.8419	0.2647	3.2258
975	0.9475	0.9105	0.8355	0.2191	3.2258
244	0.9492	0.8599	0.8343	0.2597	3.2258
504	0.9343	0.8812	0.8314	0.2355	0.0000
870	0.9291	0.8776	0.8294	0.2195	0.0000
154	0.9275	0.8730	0.8283	0.2769	3.2258
764	0.9195	0.8810	0.8268	0.2361	0.0000
564	0.9271	0.8992	0.8205	0.1673	0.0000
80	0.9233	0.8478	0.8198	0.1969	0.0000

100 individual models that displayed the best performance on the DUDE-A library, using four combination schemes: Minimum score (MIN), average score (AVE), average ranking (RANK), and average voting (VOT). The expectations on the ensembles were confirmed statistically: two combination schemes (minimum of the best 11 models and average ranking of the 4 best models) statistically outperformed the individual models in the DUDE-A database ($p = 0.0003$ and $p = 0.0132$, in that order). The MIN operator consistently outperformed the other combination schemes. When considering the influence of the number of models combined by the MIN operator on the AUC ROC metric, it was observed that above 11 models the AUC ROC did not improve substantially but poorer statistical behavior in terms of the standard deviation of the mean estimation was observed. The enrichment metrics for the best ensembles are shown in **Table 4**. Note that when applied in the screening of the DUDE-B database, the enrichment metrics for the 11-model combination based on the MIN operator (MIN-11) are similar (or in some cases, even better) than when screening DUDE-A database, validating the enrichment power of the best model combination. It may also be observed that best ensemble achieved good to excellent enrichment metrics. For instance, in the DUDE-A library, the RIE metric indicates that among the top 15 ranked compounds, 10 are known active ones.

Based on the previous results, we chose to move to the prospective (real) virtual screening campaign with the combination scheme based on the MIN operator (MIN-11). In our experience, this model combination scheme leads to high-specific model combinations (i.e., small false positive rate), which is a particularly convenient approach in our context (a small academic group from a low- to mid-income country, with limited resources to invest in hit validation); we thus often prefer to reduce the false positive rate even if this means losing sensitivity and sacrificing some active scaffolds. We have chosen to refine the former criteria (prioritizing Sp) by resorting to PPV surface analysis (Alberca et al., 2018). With the help of PPV surfaces, the evolution of the most relevant metric for our purposes, the PPV, i.e., the actual probability that a predicted hit will confirm activity when submitted to experimental testing, can be visually optimized as a function of the (Se/Sp) ratio across a range of Ya values. For this analysis, we have considered that the association between the Se/Sp and the score values of the MIN-11 model ensemble (observed in the retrospective screening campaign on DUDE-A) will hold when performing screens on other libraries (e.g., in a prospective virtual screening application). This is a strong assumption that of course is not necessarily true. However, since the AUC ROC values obtained for the DUDE-A library are unmistakably high (above 0.9 for the best model ensemble) while on the other hand the DUDE-A database Ya ratio (0.02) and size (>1,000 compounds) speak of a controlled statistical behavior (Truchon and Bayly, 2007), we believe it is a reasonable assumption in the present setting.

Using PPV surfaces (**Figure 3**), we chose 0.58 as score threshold to be used in our prospective virtual screening campaign; such score is associated to a Se/Sp ratio of 0.561 for MIN-11, and to a PPV value $\geq$ 20% for a Ya of 0.01. This means that if Ya in the real virtual screen was 0.01, we would have to submit about five predicted hits to experimental testing in order to find one confirmed hit. The virtual screen using the previous score cutoff value resulted in 157 hits, with 72 of them corresponding to approved drugs. Based on the previous analysis and our funding availability, we acquired and submitted four hits (**Figure 4**) to experimental testing: the antibiotic methacycline, the antihypertensives benzthiazide and bendroflumethiazide, and the abandoned drug odanacatib (an inhibitor of the cysteine protease cathepsin K that was pursued as a treatment for osteoporosis and bone metastasis but whose development was abandoned at Phase III long-term clinical trials due to safety issues; Drake et al., 2017).

To evaluate the ability of the selected hits to inhibit falcipain-2, we performed a two-round screening strategy. First, compounds were assayed in single dose (31.25 μM) to discard inactive molecules. Given that all of them were able to reduce to some extent (6–85%) the activity of falcipain-2 in comparison with the DMSO vehicle; we decided to evaluate the four hits in a dose-response manner (375 μM−0.45 pM) under balanced assay conditions to equalize the chances to detect competitive, non-competitive and uncompetitive inhibitors (Copeland, 2003; Yang et al., 2009). At the excitation/emission wavelengths used for AMC recording, compounds showed no significant autofluorescence in the concentration range tested. Prior to the analysis of the complete data, we explored the correlations between inhibition percentages in the primary (31.25 μM) and secondary (23.4 μM) screenings.

Compounds showed consistent results in both screenings (correlation coefficient $r^2 = 0.98$; slope = 0.9732; **Supplementary Data Sheet 3**), with odanacatib and methacycline being the most active. These compounds showed typical progression (**Supplementary Data Sheet 3**) and dose-response curves (**Figure 5**), with measurable IC_{50} and Hill slope values of 0.186 μM and −1.079 for odanacatib, and 106.4 μM and −0.9294 for methacycline. In the same range of concentrations, benzthiazide and bendroflumethiazide showed no dose-dependent inhibition.

We further characterized odanacatib and methacycline in terms of the reversibility and time-dependence of falcipain-2 inhibition. Reversible interaction with falcipain-2 was verified for both compounds by the recovery of enzyme activity after rapid addition of substrate (100-fold dilution) to the pre-incubated mix of enzyme and inhibitor (**Figures 6A,B**). In this experiment, methacycline displayed a linear progress curve (**Figure 6A**) with a stable inhibition value, indicative of rapid onset of steady state (i.e., rapid dissociation of EI complex). In the presence of odanacatib, however, the enzyme took several minutes to recover full activity and to show a permanent inhibition value (concave progress curve), suggesting slow dissociation of inhibitor from the complex with falcipain-2 (**Figure 6B**). Similarly, both inhibitors displayed different kinetic behavior when enzyme was added to a reaction mix previously containing inhibitor and substrate (**Figures 6C,D**). Methacycline displayed a typical linear progress curve (**Figure 6C**), showing a defined (stable) value of inhibition during the whole assay. In contrast, odanacatib showed non-linear kinetics

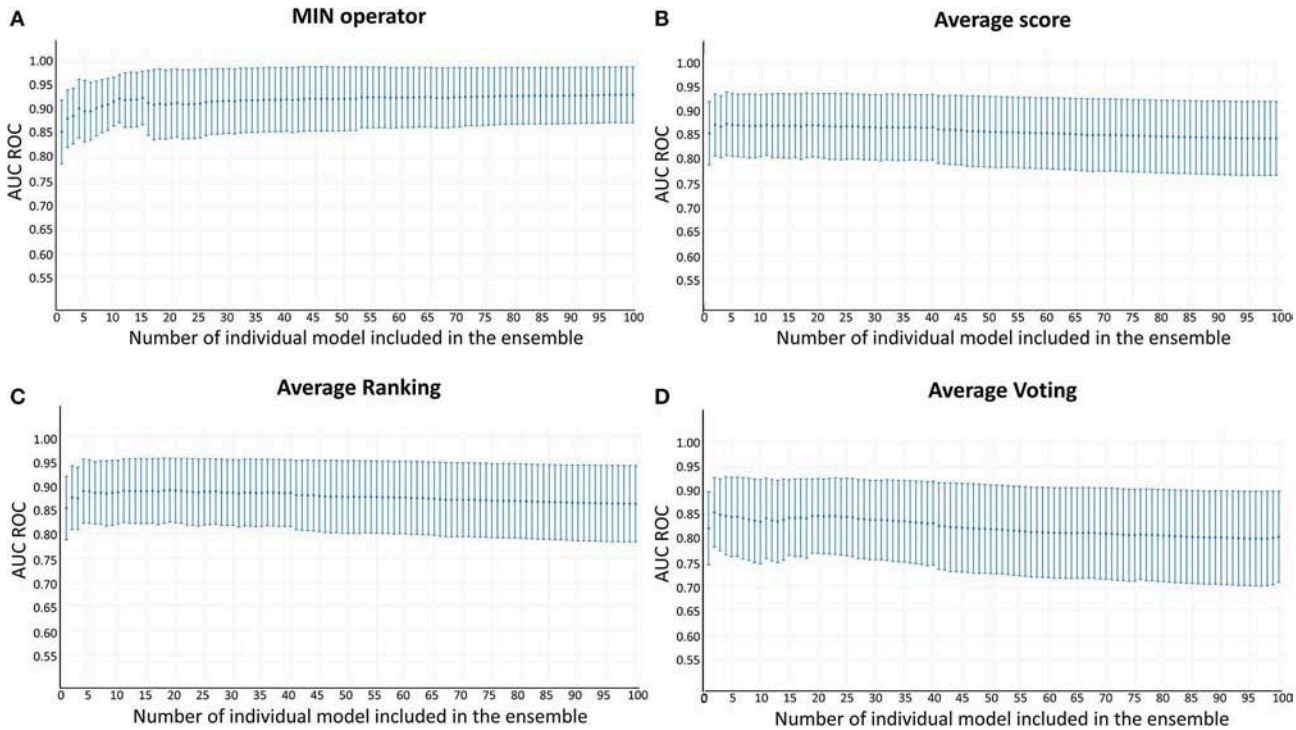


FIGURE 2 | AUC ROC metric vs. the number of combined models in the DUD-E database. **(A)** Minimum score; **(B)** Average score; **(C)** Average ranking; **(D)** Average voting.

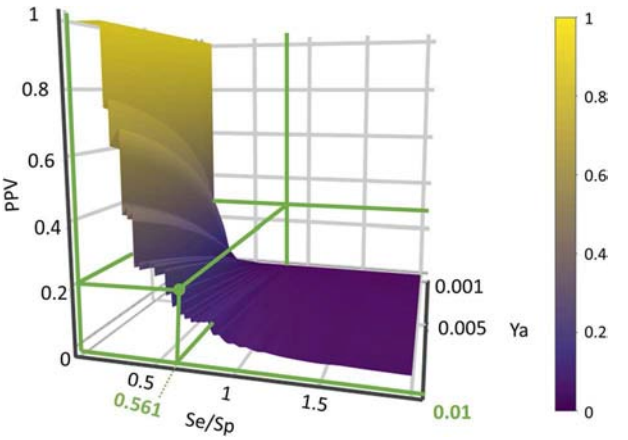


FIGURE 3 | PPV surface for the best 11-model ensemble on the DUDE-A library. The Se/Sp ratio correspondent to the chosen score cut-off value and the associated PPVs within the Ya 0.001–0.010 range are signaled.

TABLE 4 | Values of the AUC ROC, BEDROC, and RIE metrics for the best model combination (DUDE-A library).

	AUC ROC		BEDROC ($\alpha = 20$)		RIE (1%)	
Model	**DUDE A**	**DUDE B**	**DUDE A**	**DUDE B**	**DUDE A**	**DUDE B**
MIN-11	0.9214**	0.8991**	0.6414	0.4252	29.6322	9.0289
M-594	0.8529	0.7415	0.2565	0.2710	0.0000	0.0000

*For comparison purposes, the results of the best individual model (M-594) are also presented, as well as the enrichment metrics for the MIN-11 ensemble on the DUDE-B library. **Statistically significant differences in comparison with the best individual model ($p < 0.01$).*

(**Figure 6D**) with inhibition progressively increasing over time (time-dependent inhibition). As stable inhibition was observed only after ∼15 min, all subsequent kinetic experiments for odanacatib included preincubation (≥ 30 min at 30°C) with the enzyme.

To investigate the mode of inhibition of odanacatib, we first evaluated the impact of substrate concentration on the apparent IC_{50} value over a wide range (0.4–13.2xK_M) of substrate saturation levels. For this, we used a reduced set of three odanacatib concentrations selected to: (i) include IC_{50} value at each substrate condition and (ii) cover the wider inhibition range (∼15–80%) in the central stretch of the dose-response curves. As observed in **Figures 7A,B**, apparent IC_{50} values increased linearly with the increment of substrate concentration, indicating a competitive mode of inhibition for odanacatib on the activity of falcipain-2. The global fitting of all the Michaelis curves to the equation of competitive inhibition (**Figure 7C**) allowed us to estimate a Ki value of 98.2 ± 10.2 nM. As this estimation is in the limit of tight-binding inhibition ($Ki \leq 10^{-7}$ M), kinetic data were transformed to fractional velocity vs. inhibitor concentration and re-analyzed by global fitting to the Morrison equation[32]. Ki value for odanacatib determined from this approach (**Figure 7D**)

was 99.88 ± 8.28 nM, very similar to our previous (more approximate) estimation.

For methacycline, which initially showed potency in the high micromolar range, we constructed complete dose-response curves at six fixed substrate concentrations, ranging from 0.2 to 10xK_M. Although substrate concentration was increased 50-fold, only a slight increase (1.42-fold) was observed in the apparent IC_{50} value, suggesting no competition between methacycline and the small peptidic substrate Z-LR-AMC. To directly estimate Ki and α values, Michaelis plots were globally fitted to the model for mixed inhibition. This approach confirmed that methacycline inhibits falcipain-2 activity non-competitively with a Ki value of 84.4 ± 6.5 μM and α =1.42 ± 0.15 (**Figure 8**).

Given that cysteine peptidases require that its catalytic sulfhydryl group be in reduced state to show their maximal enzymatic activity, they are prone to undergo RedOx interferences caused by several classes of thiol-reactive compounds able to simulate genuine inhibition (Thorne et al., 2010). In many cases, this artifactual inhibition can be significantly relieved by simply changing the reduction potential of the activity buffer, thus providing a diagnostic test to detect false-positive RedOx compounds. To establish if this could be the case for methacycline, we further investigated the effect of the strength and concentration of reducing agents (DTT, a strong reducing agent, and cysteine, a weak reducing agent) on the inhibition of falcipain-2 by this molecule. Dose-response curves were very similar regardless of the final concentration of the reducing agent (100-fold range) present in the assay buffer (**Supplementary Data Sheet 3**). These results rule out common types of RedOx interference and suggest that methacycline genuinely inhibits falcipain-2.

Once established that odanacatib and methacycline inhibit the peptidolytic activity of falcipain-2, we assayed whether these molecules could also modulate falcipain-2 proteolytic activity on its natural substrate, human hemoglobin (hHb). To this end, we pre-incubated the enzyme with increasing concentrations of both inhibitors and then added the hHb substrate. E-64, a specific and highly potent irreversible inhibitor of C1A cysteine peptidases, was used as a positive control. After incubation at 37°C for 3 h, reaction mixes were resolved electrophoretically by SDS-PAGE on a 15% acrylamide gel and Coomassie stained. The amount of undegraded hHb, observed as a doublet of around 15 kDa, was estimated densitometrically. As shown in **Supplementary Data Sheet 3**, odanacatib inhibited the hemoglobinase activity of FP2 in a dose-response manner. For methacycline, however, no inhibition was observed in this assay, even at the highest concentration tested (1 mM). Of note, the effective inhibitory concentrations of odanacatib in this assay were in the low-to-middle micromolar range, a significant shift in comparison to the sub-micromolar potency previously observed in the inhibition of Z-LR-AMC hydrolysis. Overall, these observations suggest that (i) the existing differences between the surrogate (peptidic) and the natural (macromolecular) falcipain-2 substrates are somehow important for the inhibitory efficiency of both inhibitory molecules and that, at least, (ii) low-to-middle micromolar compound concentrations would be required to assess their efficacy in a more physiological context (i.e., cellular culture).

To analyze the influence of odanacatib and methacycline in the intraerythrocytic cycle of *P. falciparum*, a synchronized culture of RBCs infected (trophozoite stage) was treated with increasing concentrations of odanacatib (1, 10, or 100 μM) and methacycline (10, 50, or 500 μM). E-64 (1, 5, or 25 μM) was again used as a positive control. After 48 h, the number of infected erythrocytes was evaluated by light microscopy in stained blood smears. Odanacatib (100 μM) significantly reduced the parasitemia (**Figure 9**), with no apparent reduction in the other two concentrations tested. Methacycline significantly reduced the parasitemia at 500 and 50 μM, but not at 10 μM. As expected, E-64 significantly reduced the parasitemia at 5 and 25 μM, but no at 1 μM. Almost no erythrocytes infected at schizont-stage were observed in the treatments (**Supplementary Data Sheet 3**). It is important to mention that at the highest concentrations assayed, methacycline (500 μM) and odanacatib (100 μM) induced cytotoxic effects on RBC, as observed in the hemolysis assay (**Supplementary Data Sheet 3**).

Molecular docking results were in good agreement with the experimental observations. **Figure 10** shows the best result of the docking simulation for odanacatib in the catalytic binding site, that is, the pose that showed lower value of the scoring function. Two hydrogen bonding interactions were found between the compound and GLN36 and ASN173. The docking score was −6.94 kcal/mol, which is lower than the score achieved for E64 in the same conditions (−4.91 kcal/mol). Regarding methacycline, we detected hydrogen bonding interactions between methacycline and the residues of the proposed (non-catalytic) binding pocket. Residues like ASP72, ASN112, PRO111, and ALA110 could be implicated in the stabilization of the complex (docking score of−5.65 kcal/mol). More studies will be performed to evaluate these predictions.

DISCUSSION AND CONCLUSIONS

Using an ensemble learning approximation, we have performed a ligand-based virtual screening campaign to identify new falcipain-2 inhibitors as potential new treatments against malaria. There are some previous reports on the development of ligand-based models to predict the activity of *P. falciparum* cysteine proteases. The approaches used in such studies show considerable differences with the one reported here: they used conformation-dependent descriptors (3D QSAR) to infer regression models; in almost all cases, congeneric series of comparatively narrow chemical diversity have been used to train the models, thus limiting their applicability domain and; the reported models have mostly been used for explanatory rather than predictive purposes. Xue and coworkers realized CoMFA and CoMSIA 3D-QSAR studies on a series of 93 alkoxylated and hydroxylated chalcones (Xue et al., 2004). Potshangbam and coworkers also carried out CoMSIA and CoMFA studies on a series of 54 2-pyrimidinecarbonitrile

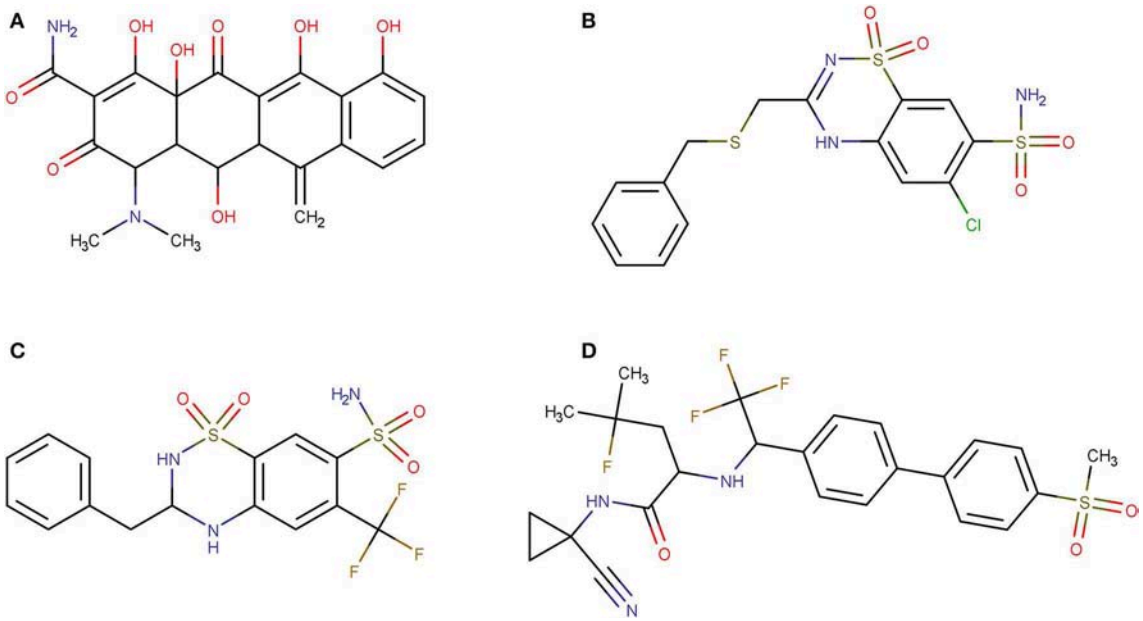


FIGURE 4 | Molecular structures of the hits selected in the prospective virtual screening campaign that were submitted to experimental confirmation. **(A)** methacycline; **(B)** benzthiazide; **(C)** bendroflumethiazide; **(D)** odanacatib.

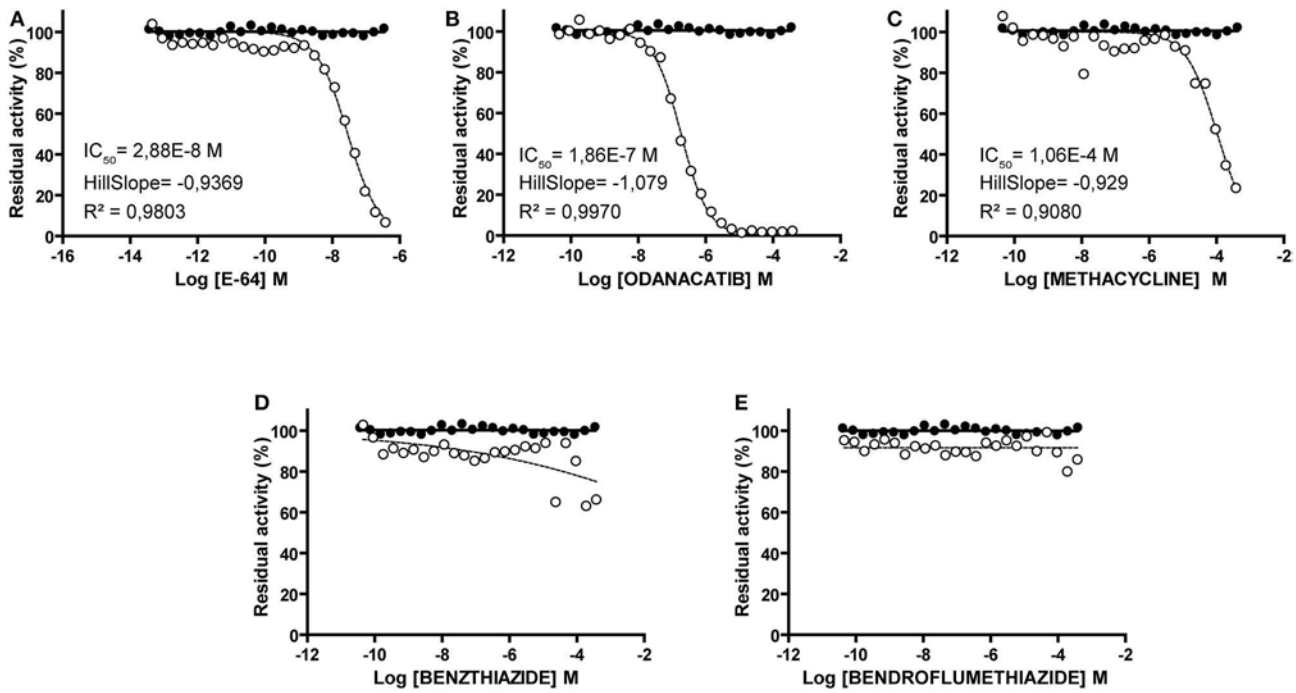


FIGURE 5 | Dose-response curves of identified falcipain-2 inhibitors (open circles). For each compound, dotted line represents the best fit of experimental data to the four-parameter Hill equation. **(A)** E-64. **(B)** Odanacatib. **(C)** Methacycline. **(D)** Benzthiazide. **(E)** Bendroflumethiazide. For those compounds achieving data convergence, the resultant values for the parameters IC50, Hillslope and R2 are indicated. In all cases, equivalent volumes of DMSO vehicle were assayed in parallel (closed circles).

analog inhibitors of falcipain-3 (Potshangbam et al., 2011). Using the same approximations, Wang et al. performed a 3D QSAR study of 247 2-pyrimidinecarbonitrile analog inhibitors of falcipain-3 (Wang et al., 2013). Teixeira and colleagues did a CoMFA and CoMSIA analysis of a series of 39 peptidyl vinyl sulfone derivatives as potential cysteine protease

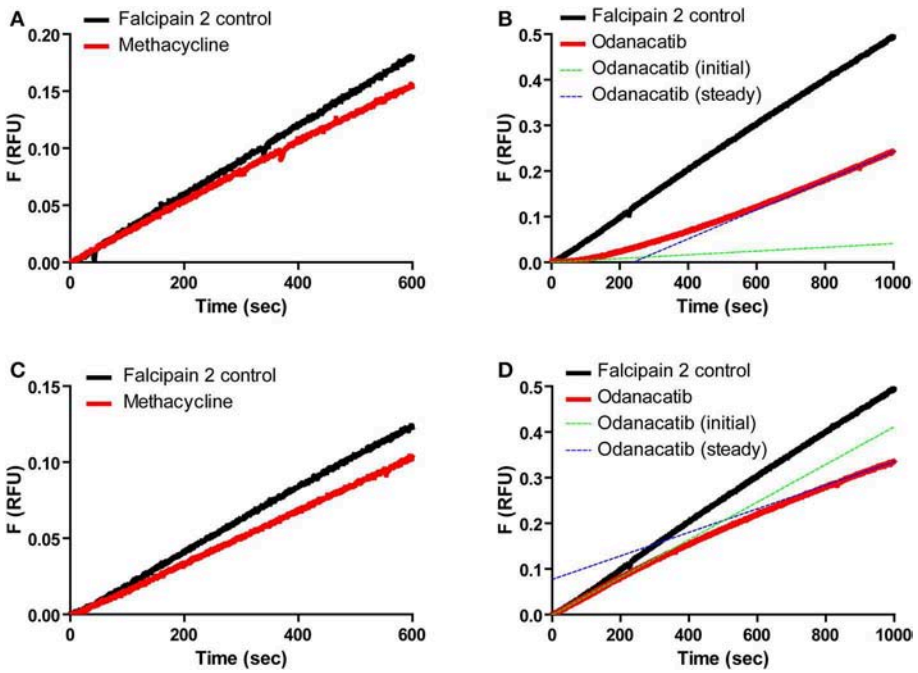


FIGURE 6 | Reversibility and time dependence of the inhibition of falcipain-2 by methacycline and odanacatib. Top panel: Product progress curves for the dissociation of E-I complex by rapid dilution (100-fold) of enzyme-inhibitor mix into substrate solution. **(A)** Methacycline. **(B)** Odanacatib. Bottom Panel: Product progress curves for the formation of E-I complex by rapid addition of enzyme to a substrate-inhibitor mix. **(C)** Methacycline. **(D)** Odanacatib.

inhibitors (Teixeira et al., 2011). Very recently, Allangba et al. derived complexation QSAR models and pharmacophores from a training set of 15 lactone–chalcone and isatin–chalcone hybrid inhibitors with falcipain-2 inhibitory activity (Allangba et al., 2019). The most similar study to our own is possibly the one by Mugumbate and coworkers, who as a part of and hybrid ligand- and structure-based approach, obtained ligand-based models based on Pentacle alignment-independent descriptors, using a training set of nine non-peptide inhibitors of falcipain-2 (Mugumbate et al., 2013). They performed a retrospective pilot screen before using their protocol to explore the ZINC database, retrieving falcipain-2 inhibitors in the low μM range. All in all, the enrichment metrics they computed in their retrospective screen are similar to the ones we obtained here.

Our most predictive model combination was chosen (i.e., trained) based on a retrospective virtual screening campaign (DUDE-A library). The enhanced ability of the selected model-ensemble to retrieve falcipain-2 inhibitors in comparison to our best individual models was checked using a second retrospective virtual screening experiment (DUDE-B library). Four of the hits emerging from our prospective virtual screening experiment were acquired and assayed against the enzyme. Two of them, odanacatib (previously investigated as treatment against osteoporosis and bone metastasis) and methacycline (an antibiotic) confirmed our predictions, reducing the peptidolytic activity of the enzyme. Interestingly, our observed PPV (50%, corresponding to two experimentally confirmed hits out of four assayed compounds) exceeded our theoretic expectations based on PPV surfaces analysis, which suggested a PPV of at least 20% for the chosen score threshold, for a hypothetic yielding of active compounds of 1%.

Both hits displayed different inhibition mechanisms. In agreement with previous reports for the interaction with other Papain-like C1A (Clan CA family) human cathepsins (Gauthier et al., 2008), odanacatib inhibits falcipain-2 in a reversible, competitive and tight-binding (sub-micromomlar) mode. For human cathepsin K, odanacatib inhibition occurs throughout the formation of a covalent but yet reversible thioimidate adduct between the -SH in the catalytic Cys residue and the nitrile warhead (Oballa et al., 2007). This covalent association mechanism results in on- and off-rates of 5.3×10^6 M and 0.0008 s^{-1} ($t_{1/2}$ ∼14 min), respectively (Gauthier et al., 2008). These observations are in line with the slow association and dissociation kinetics observed by us for falcipain-2 inhibition, suggesting that a similar chemical inhibition mechanism could be occurring. Other compounds bearing the N-(1-cyanocyclopropyl)-amide inhibitory scaffold present in odanacatib, have been reported as potent (Ki ∼1–2 nM) and selective (>15-fold over human cathepsins) falcipain-2 inhibitors (Ang et al., 2011; Nizi et al., 2018).

On the other hand, methacycline acts as a reversible, non-competitive and sub-milimolar inhibitor of falcipain-2. Based on our observations (reproducible, reversible and dose-dependent reduction of enzyme activity, rapid equilibrium onset, Hill slopes ∼ −1, inhibitory activity insensitive to the RedOx potential

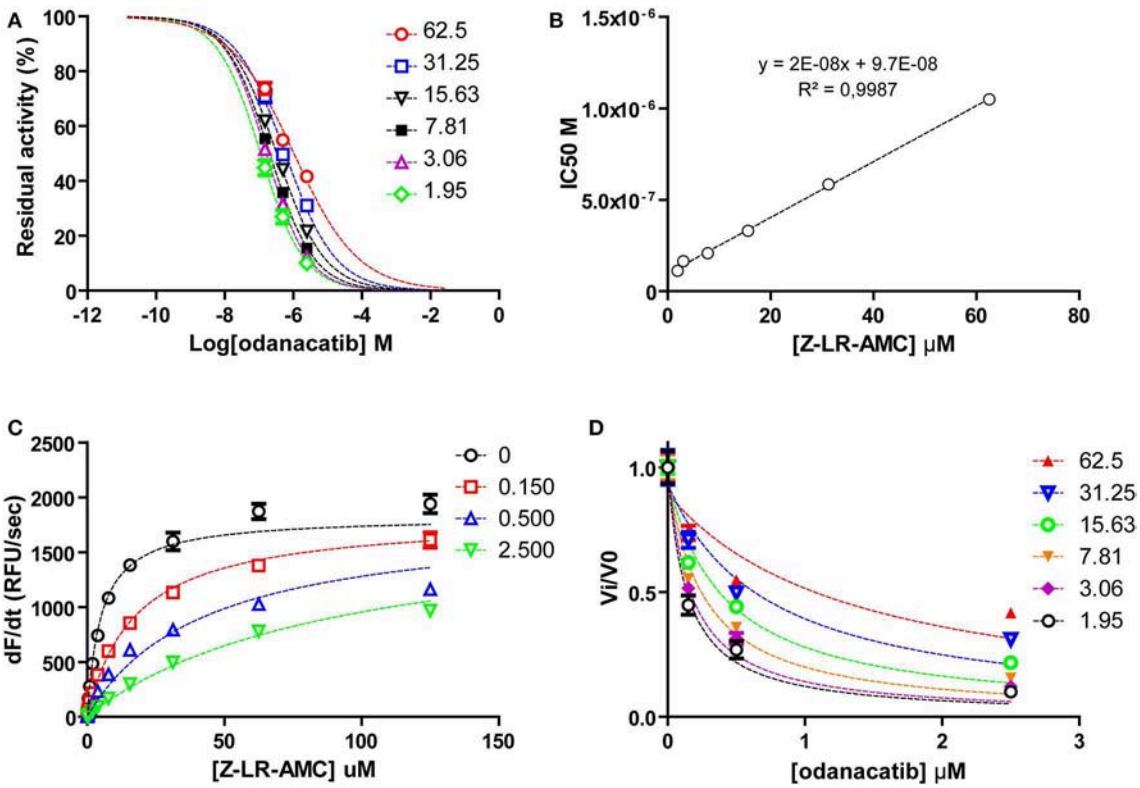


FIGURE 7 | (A) Dose-response curves for odanacatib at fixed substrate concentrations. Dotted lines represent the best fit of experimental data to the four-parameter Hill equation. **(B)** Effect of substrate concentration on the IC50 values of falcipain-2 inhibition by odanacatib. IC50 values increase linearly (>9-fold) with substrate concentration in the range 1.95–62.5 µM. Dotted line represents the best fit of data to linear equation. Y-axis intercept accounts for the Ki value. **(C)** Global fitting of kinetic data to the competitive inhibition model equation. **(D)** Global fit ting of kinetic data to the Morrison equation.

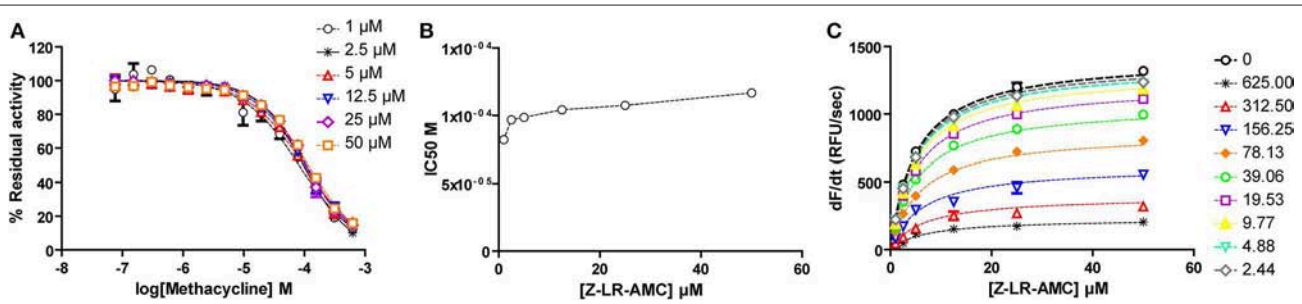


FIGURE 8 | Methacycline is a non-competitive, sub-millimolar inhibitor of falcipain-2. **(A)** Dose-response curves for methacycline at fixed substrate concentrations. Dotted lines represent the best fit of experimental data to the four-parameter Hill equation. **(B)** Half maximal inhibitory concentration of methacycline increases slightly (1.42-fold) with substrate concentration in the range: 1–50 µM. **(C)** Global fitting of kinetic data to the equation of mixed inhibition model.

and no signs of common compound-specific assay interferences such as autofluorescence or aggregate formation), methacycline inhibition seems to occur throughout a genuine mechanism. This may lead to rational optimization efforts to improve affinity. Further studies should be performed to confirm the putative binding pocket suggested by our docking experiments, to move in that direction. To date, only few non-competitive falcipain-2 inhibitors have been reported; including suramin analogs (Marques et al., 2013), heme analogs (Marques et al., 2015), and (E)-chalcones (Bertoldo et al., 2015) reported. Suramin, heme, and their analogs inhibit falcipain-2 with IC_{50} values in the nanomolar range and seem to share a common "non-competitive like" inhibition mechanism that occurs through the formation of a ternary enzyme:inhibitor:substrate complex of stoichiometry 1:1:2. In both cases, the authors argued that the binding of the inhibitor to falcipain-2 reveals a novel regulatory substrate binding site in the enzyme, which allows the subsequent allosteric binding of a second substrate molecule,

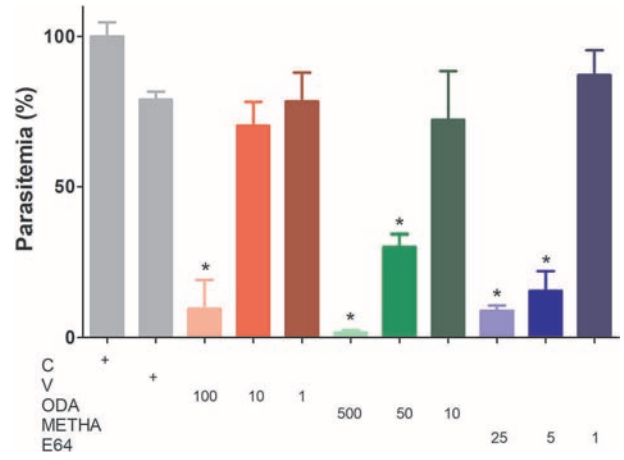


FIGURE 9 | Effect of odanacatib and methacycline on the development of P. falciparum under culture. Cultures of erythrocytes infected (trophozoite stage) at 2% hematocrit and 0.5% parasitemia were incubated with increasing concentrations of odanacatib (ODA; 1, 10, or 100 μM) and methacycline (METHA; 10, 50, or 500 μM). E-64 (1, 5, or 25 μM) was used as positive control of falcipain-2 inhibition. DMSO was used as a vehicle control (V) and RPMI 1640 medium as a control (C). After 48 h, the number of infected erythrocytes (ring, trophozoite, and schizont-stages) was evaluated by light microscopy in stained blood smears. Data are the means ± SD of one experiment performed by triplicate. (*) $p < 0.05$ represents the differences between control and the treatments.

resulting in falcipain-2 inhibition (Marques et al., 2013, 2015). Very similar to what we have found for methacycline, (E)-chalcones 48 (Ki =45 μM, α < 1) and 66 (Ki =7 μM, α = 1) display a classical non-competitive inhibition profile for falcipain-2 with no evidence of substrate inhibition (Bertoldo et al., 2015). None of these inhibitors appear to bind falcipain-2 active site, thus anticipating new routes to overcome the critical issue of selectivity over human cathepsins. In this regard, the identification and targeting of non-active (i.e., allosteric) binding sites within the falcipain-2 molecule seems to be an attractive and effective alternative to traditional active site-directed inhibitors, as recently showed by Pant and coworkers (Pant et al., 2018). Two compounds, rationally designed to target an allosteric site present in the pro-mature domain interface of falcipains-2/-3, were able to bind both pro-enzymes with nanomolar affinities and arrest *P. falciparum* growth, clearly illustrating the potential of this approach.

Of interest, odanacatib inhibits the peptidolytic activity of falcipain-2 much more efficiently than its hemoglobinase activity. Although displaying comparable affinities for the enzyme [K_D^{hHb} =3.3 μM (Hogg et al., 2006); $K_M^{Z-LR-AMC}$ = 4.8 μM], there are important differences between Z-LR-AMC and hHb as falcipain-2 substrates, regarding both their binding modes and their catalytic heterogeneity. In the first case, the small peptidic substrate Z-LR-AMC accommodates completely within falcipain-2 active site and the enzyme-substrate complex relies entirely on active site interactions for its stabilization and catalytic transformation. In addition, Z-LR-AMC substrate comprise a single cleavage site per molecule. These facts seem to make it somehow more vulnerable to the competitive binding of odanacatib to the active site and to the catalytic impairment promoted by the binding of methacycline to a presumptive allosteric site, like the one proposed here. In the case of hHb, it has been shown that interaction with the enzyme depends almost exclusively on a unique falcipain-2 structural motif (the "arm"), located >25 Å away from the active site (Pandey et al., 2005; Wang et al., 2006). In fact, a falcipain-2 mutant lacking most of the arm loop showed no activity or affinity against hHb, although remained fully active against a number of generic peptide and protein substrates (Pandey et al., 2005). Additionally, hHb was able to bind to falcipain-2 molecules with the active site blocked by the irreversible inhibitor E-64, clearly indicating that the recognition of intact hHb is mostly independent of the active site. Considering all these observations, we can hypothesize that the binding of odanacatib to falcipain-2 active site would not be likely to perturb substantially the binding affinity of intact hHb to its exosite and *vice versa*. Thus, with the ability to bind both enzyme forms (falcipain-2 and the falcipain-2-hHb complex) with comparable affinity, odanacatib would probably behave as a non-competitive inhibitor against an exosite-binding substrate such as intact hHb. This has been previously reported for small active site-directed inhibitors of proteases and kinases when acting on their natural macromolecular substrates (Krishnaswamy and Betz, 1997; Pedicord et al., 2004; Blat, 2010). The change in the inhibition modality of odanacatib, however, seems insufficient to explain, *per se,* the magnitude of the drop in its inhibitory potency against the falcipain-2/hHb system. A second line of argument came from the observation that hHb molecule is a catalytically heterogeneous substrate, comprising numerous independent falcipain-2 cleavage sites and whose digestion seems to be a non-ordered process (Subramanian et al., 2009). The occurrence of multiple cleavage events at different sites along the protein sequence leads to the formation of numerous digestion products of lower molecular weight. These digestion products also contain functional falcipain-2 cleavage sites and become new substrates that, after a new round of proteolysis, may generate additional substrate peptides. As the reaction proceeds, this iterative process leads to an increase in the number and the global concentration of peptidyl substrates able to compete for the binding to falcipain-2 active site. This might eventually lead to a partial relief of inhibition by competitive, active site-directed inhibitors, as would be the case of odanacatib.

Considering that falcipain-2 and−3 are the major cysteine proteases required for the intraerythrocytic development of *P. falciparum*, we evaluated the antiparasitic effect of odanacatib and methacycline. The inhibitors of cysteine proteases block the hydrolysis of hemoglobin, causing the development of enlarged, hemoglobin-filled food vacuoles in trophozoites and failure of parasites to complete their development (Marco and Coterón, 2012). The two drugs showed a clear inhibition in a dose-depend manner on the intraerythrocytic cycle of *P. falciparum*. The effective concentrations of odanacatib in *P. falciparum* cultures were in the low-to-middle micromolar range, similar to those observed in the inhibition assay of hemoglobinase activity. This finding is compatible with a hypothetical mode

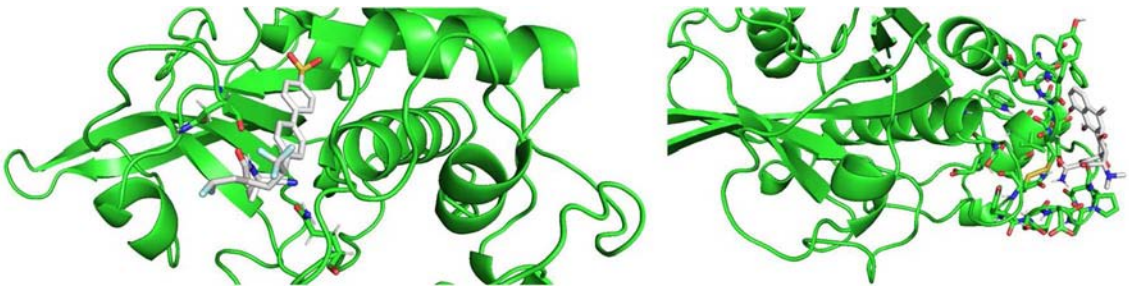


FIGURE 10 | Best poses from the molecular docking experiments for odanacatib **(Left)** and methacycline **(Right)**. Interacting residues are shown in sticks.

of action trough inhibition of falcipain functions within the food vacuole. Although the inhibition of diverse proteases by tetracycline derivatives has previously been reported (Morrison, 1969; Zucker et al., 1985; Sanchez Mejia et al., 2001; Chi et al., 2011), to the best of our knowledge, this is the first report of the inhibition of a C1A cysteine peptidase from a protozoa parasite by an antibiotic of the tetracycline family. The molecular targets for the action of tetracyclines against *Plasmodium* parasites have not been fully elucidated. However, their mode of action seems to include the inhibition of protein synthesis at mitochondrial, plastid and nuclear ribosomes by the association with ribosomal components (Gaillard et al., 2015). Additional mechanisms, such as reduction in *de novo* pyrimidine synthesis and a decrease in the transcription rates of mitochondrial and apicoplast genes, have also been postulated (Briolant et al., 2008; Gaillard et al., 2015). The inclusion of falcipain-2 among the potential targets of tetracycline derivatives adds new possibilities for the development of "two-edged swords" candidate drugs for *P. falciparum*, with potential benefits in terms of potency and delay of resistance appearance (Agarwal et al., 2017). The contribution of this mechanism to the global antimalarial activities of these antibiotics remains to be established in future investigations.

It should also be underlined that odanacatib underwent long-term clinical trials as a treatment of postmenopausal osteoporosis (Bone et al., 2015), which were early stopped due to robust efficacy and a favorable benefit/risk profile. However, its clinical development was dropped due to an increased risk of stroke in the postmenopausal patients on odanacatib vs. a placebo group. Accordingly, our findings on the potential use of the drug against malaria could be considered a drug rescue example, i.e., a proposal on a new medical used of and abandoned or discontinued drug. Do the safety issues of odanacatib pose an inevitable impediment for their potential development as antimalarial treatment? Not necessarily. Although they are indeed a concern, it should be considered that the odanacatib augmented risk of stroke was observed in long-term studies, whereas the drug could possibly be administered in a short-term manner as malaria treatment (for instance, artemisinin-based combination therapies only require a 3-day course to achieve efficacy in cases of uncomplicated *P. falciparum* malaria). Accordingly, the long-term risks of odanacatib use may not have a negative impact on its use as antimalarial. There are well-known examples of drug rescue of discontinued drugs with severe safety issues, that can be re-introduced in a new therapeutic setting with the pertinent precautions. For instance, thalidomide was largely abandoned due to its teratogenic effects, but has been recently relaunched to the market for the treatment of leprosy and multiple myeloma (Teo et al., 2002; Mercurio et al., 2017).

Pharmacokinetics studies reveal that after multiple-dose administration of odanacatib 50 mg (once weekly for 4 weeks), average maximal plasma concentrations of around 400 nM are observed (Chen et al., 2018), although a high fraction of plasma protein bound drug has also been reported (Kassahun et al., 2014). Accordingly, further studies are required to evaluate the dose-compatibility between the previously investigated therapeutic use and the possible antimalarial indication.

DATA AVAILABILITY

All datasets generated for this study are included in the **Supplementary Data Sheet 1**.

ETHICS STATEMENT

Human erythrocytes were obtained from volunteer donors with a procedure approved by CEIC (Committee for Ethics on Clinical Investigation, Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires EXP-UBA: 0048676/2017).

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

The authors thank Agencia Nacional de Promoción Científica y Tencológica (PICT 2013-0520, PICT 2017-0643, PICT 2014-0327), University of La

Plata (Incentivos UNLP) and the International Foundation for Science (Grants F/4081-1 and F/4081-2) for funding.

ACKNOWLEDGMENTS

The authors thank CONICET and UNLP. Silvia Repetto for kindly providing us the laboratory for *Plasmodium falciparum* culture. We would like to thank Prof. Luciana Gavernet for her kind advice on the docking simulations. The Titan Xp used for this research was kindly donated by the NVIDIA Corporation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00534/full#supplementary-material

REFERENCES

Agarwal, D., Gupta, R. D., and Awasthi, S. K. (2017). Are antimalarial hybrid molecules a close reality or a distant dream? *Antimicrob. Agents Chemother.* 61:e00249–e00217. doi: 10.1128/AAC.00249-17

Ajayi, N. A., and Ukwaja, K. N. (2013). Possible artemisinin-based combination therapy-resistant malaria in Nigeria: a report of three cases. *Rev. Soc. Bras. Med. Trop.* 46, 525–527. doi: 10.1590/0037-8682-0098-2013

Alberca, L. N., Sbaraglini, M. L., Balcazar, D., Fraccaroli, L., Carrillo, C., Medeiros, A., et al. (2016). Discovery of novel polyamine analogs with anti-protozoal activity by computer guided drug repositioning. *J. Comput. Aided Mol. Des.* 30, 305–321. doi: 10.1007/s10822-016-9903-6

Alberca, L. N., Sbaraglini, M. L., Morales, J. F., Dietrich, R., Ruiz, M. D., Pino Martínez, A. M., et al. (2018). Cascade ligand- and structure-based virtual screening to identify new trypanocidal compounds inhibiting putrescine uptake. *Front. Cell. Infect. Microbiol.* 8:173. doi: 10.3389/fcimb.2018.00173

Aley, S. B., Sherwood, J. A., Marsh, K., Eidelman, O., and Howard, R. J. (1986). Identification of isolate-specific proteins on sorbitol-enriched *Plasmodium falciparum* infected erythrocytes from Gambian patients. *Parasitology* 92 (Pt 3), 511–525.

Allangba, K. N. P. G., Keita, M., Kre N'Guessan, R., Megnassan, E., Frecer, V., and Miertus, S. (2019). Virtual design of novel *Plasmodium falciparum* cysteine protease falcipain-2 hybrid lactone-chalcone and isatin-chalcone inhibitors probing the S2 active site pocket. *J. Enzyme Inhib. Med. Chem.* 34, 547–561. doi: 10.1080/14756366.2018.1564288

Alvarez, C. L., Schachter, J., de Sa' Pinheiro, A. A., Silva, L., de, S., Verstraeten, S. V., et al. (2014). Regulation of extracellular ATP in human erythrocytes infected with *Plasmodium falciparum*. *PLoS ONE* 9:e96216. doi: 10.1371/journal.pone.0096216

Ang, K. K. H., Ratnam, J., Gut, J., Legac, J., Hansell, E., Mackey, Z. B., et al. (2011). Mining a cathepsin inhibitor library for new antiparasitic drug leads. *PLoS Negl. Trop. Dis.* 5:e1023. doi: 10.1371/journal.pntd.0001023

Balaban, A. T., Ciubotariu, D., and Medeleanu, M. (1991). Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors. *J. Chem. Inf. Comput. Sci.* 31, 517–523. doi: 10.1021/ci00004a014

Bekono, B. D., Ntie-Kang, F., Owono Owono, L. C., and Megnassan, E. (2018). Targeting cysteine proteases from *Plasmodium falciparum*: a general overview, rational drug design and computational approaches for drug discovery. *Curr. Drug Targets.* 19, 501–526. doi: 10.2174/1389450117666161221122432

Bertoldo, J. B., Chiaradia-Delatorre, L. D., Mascarello, A., Leal, P. C., Cordeiro, M. N., Nunes, R. J., et al. (2015). Synthetic compounds from an in house library as inhibitors of falcipain-2 from *Plasmodium falciparum*. *J. Enzyme Inhib. Med. Chem.* 30, 299–307. doi: 10.3109/14756366.2014.920839

Biot, C., Pradines, B., Sergeant, M. H., Gut, J., Rosenthal, P. J., and Chibale, K. (2007). Design, synthesis, and antimalarial activity of structural chimeras of thiosemicarbazone and ferroquine analogues. *Bioorg. Med. Chem. Lett.* 17, 6434–6438. doi: 10.1016/j.bmcl.2007.10.003

Blat, Y. (2010). Non-competitive inhibition by active site binders. *Chem. Biol. Drug. Des.* 75, 535–540. doi: 10.1111/j.1747-0285.2010.00972.x

Bone, H. G., Dempster, D. W., Eisman, J. A., Greenspan, S. L., McClung, M. R., Nakamura, T., et al. (2015). Odanacatib for the treatment of postmenopausal osteoporosis: development history and design and participant characteristics of LOFT, the Long-Term Odanacatib Fracture Trial. *Osteoporos. Int.* 26, 699–712. doi: 10.1007/s00198-014-2944-6

Briolant, S., Fusaï, T., Rogier, C., and Pradines, B. (2008). Tetracycline antibiotics in malaria. *Open Trop. Med. J.* 1, 31–46. doi: 10.2174/1874315300801010031

Carbonneau, M. A., Granger, E., Raymond, A. J., and Gagnon, G. (2016). Robust multiple-instance learning ensembles using random subspace instance selection. *Pattern Recogn.* 58, 83–99. doi: 10.1016/j.patcog.2016.03.035

Chen, X., Jiang, J., Hu, P., Zajic, S., Liu, W., McCrea, J., et al. (2018). Odanacatib pharmacokinetics comparison between Chinese and non-Chinese postmenopausal women. *Clin. Pharmacol. Drug Dev.* 7, 744–750. doi: 10.1002/cpdd.434

Chi, Z., Liu, R., Yang, H., Shen, H., and Wang, J. (2011). Binding of tetracycline and chlortetracycline to the enzyme trypsin: spectroscopic and molecular modeling investigations. *PLoS ONE* 6:e28361. doi: 10.1371/journal.pone.0028361

Chipeleme, A., Grut, J., Rosenthal, P. J., and Chibale, K. (2007). Synthesis and biological evaluation of phenolic Mannich bases of benzaldehyde and (thio)semicarbazone derivatives against the cysteine protease falcipain-2 and a chloroquine resistant strain of *Plasmodium falciparum*. *Bioorg. Med. Chem.* 15, 273–282. doi: 10.1016/j.bmc.2006.09.055

Chiyanzu, I., Hansell, E., Gut, J., Rosenthal, P. J., McKerrow, J. H., and Chibale, K. (2003). Synthesis and evaluation of isatins and thiosemicarbazone derivatives against cruzain, falcipain-2 and rhodesain. *Bioorg. Med. Chem. Lett.* 13, 3527–3530. doi: 10.1016/S0960-894X(03)00756-X

Conroy, T., Guo, J. T., Elias, N., Cergol, K. M., Gut, J., Legac, J., et al. (2014). Synthesis of gallinamide A analogues as potent falcipain inhibitors and antimalarials. *J. Med. Chem.* 57, 10557–10563. doi: 10.1021/jm501439w

Copeland, R. A. (2003). Mechanistic considerations in high-throughput screening. *Anal. Biochem.* 320, 1–12. doi: 10.1016/S0003-2697(03)00346-4

Copeland, R. A. (2005). Evaluation of enzyme inhibitors in drug discovery. A guide for medicinal chemists and pharmacologists. *Methods Biochem. Anal.* 46, 1–265. doi: 10.1002/9781118540398

Corsello, S. M., Bittker, J. A., Liu, Z., Gould, J., McCarren, P., Hirschman, J. E., et al. (2017). The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* 23, 405–408. doi: 10.1038/nm.4306

Desai, P. V., Patny, A., Gut, J., Rosenthal, P. J., Tekwani, B., Srivastava, A., et al. (2006). Identification of novel parasitic cysteine protease inhibitors by use of virtual screening. 2. The available chemical directory. *J. Med. Chem.* 49, 1576–1584. doi: 10.1021/jm0505765

Desai, P. V., Patny, A., Sabnis, Y., Tekwani, B., Gut, J., Rosenthal, P., et al. (2004). Identification of novel parasitic cysteine protease inhibitors using virtual screening. 1. The ChemBridge database. *J. Med. Chem.* 47, 6609–6615. doi: 10.1021/jm0493717

Domínguez, J. N., López, S., Charris, J., Iarruso, L., Lobo, G., Semenov, A., et al. (1997). Synthesis and antimalarial effects of phenothiazine inhibitors of a Plasmodium falciparum cysteine protease. *J. Med. Chem.* 40, 2726–2732. doi: 10.1021/jm970266p

Drake, M. T., Clarke, B. L., Oursler, M. J., and Khosla, S. (2017). Cathepsin K inhibitors for osteoporosis: biology, potential clinical utility, and lessons learned. *Endocr. Rev.* 38, 325–350. doi: 10.1210/er.2015-1114

El Habib Daho, M., and Chikh, M. A. (2015). Combining bootstrapping samples, random subspaces and random forests to build classifiers. *J. Med. Imag. Health Inf.* 5, 539–544. doi: 10.1166/jmihi.2015.1423

Estrada, E. (1996). Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes. *J. Chem. Inf. Comput. Sci.* 36, 844–849.

Ettari, R., Pinto, A., Tamborini, L., Angelo, I. C., Grasso, S., Zappalà, M., et al. (2014). Synthesis and biological evaluation of papain-family cathepsin L-like cysteine protease inhibitors containing a 1,4-benzodiazepine scaffold as antiprotozoal agents. *ChemMedChem* 9, 1817–1825. doi: 10.1002/cmdc.201402079

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis, 5th Edn*. West Sussex: Wiley.

Ferreira, L. G., and Andricopulo, A. D. (2016). Drug repositioning approaches to parasitic diseases: a medicinal chemistry perspective. *Drug Discov. Today* 21, 1699–1710. doi: 10.1016/j.drudis.2016.06.021

Fujii, N., Mallari, J. P., Hansell, E. J., Mackey, Z., Doyle, P., Zhou, Y. M., et al. (2005). Discovery of potent thiosemicarbazone inhibitors of rhodesain and cruzain. *Bioorg. Med. Chem. Lett*. 15, 121–123. doi: 10.1016/j.bmcl.2004.10.023

Gaillard, T., Madamet, M., and Pradines, B. (2015). Tetracyclines in malaria. *Malar. J*. 14:445. doi: 10.1186/s12936-015-0980-0

Gantner, M. E., Peroni, R. N., Morales, J. F., Villalba, M. L., Ruiz, M. E., and Talevi, A. (2017). Development and validation of a computational model ensemble for the early detection of BCRP/ABCG2 substrates during the drug design stage. *J. Chem. Inf. Model*. 57, 1868–1880. doi: 10.1021/acs.jcim.7b00016

Gauthier, J. Y., Chauret, N., Cromlish, W., Desmarais, S., Duong, L. T., Falgueyret, J. P., et al. (2008). The discovery of odanacatib (MK-0822), a selective inhibitor of cathepsin K. *Bioorg. Med. Chem. Lett*. 18, 923–928. doi: 10.1016/j.bmcl.2007.12.047

Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y. D., Lee, K. H., and Tropsha, A. (2003). Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des*. 17, 241–253. doi: 10.1023/A:1025386326946

Goud, P. M., Sheri, A., Desai, P. V., Watkins, E. B., Tekwani, B., Sabnis, Y., et al. (2005). Design, synthesis and evaluation of trisubstituted thiazoles targeting Plasmodium falciparum cysteine proteases. *Med. Chem. Res*. 14, 74–105. doi: 10.1007/s00044-005-0126-y

Gramatica, P. (2013). On the development and validation of QSAR models. *Methods Mol. Biol*. 930, 499–526. doi: 10.1007/978-1-62703-059-5_21

Greenbaum, D. C., Mackey, Z., Hansell, E., Doyle, P., Gut, J., Caffrey, C. R., et al. (2004). Synthesis and structure-activity relationships of parasiticidal thiosemicarbazone cysteine protease inhibitors against *Plasmodium falciparum*, *Trypanosoma brucei*, and *Trypanosoma cruzi*. *J. Med. Chem*. 47, 3212–3219. doi: 10.1021/jm030549j

Hans, R. H., Guantai, E. M., Lategan, C., Smith, P. J., Wan, B., Franzblau, S. G., et al. (2010). Synthesis, antimalarial and antitubercular activity of acetylenic chalcones. *Bioorg. Med. Chem. Lett*. 20, 942–944. doi: 10.1016/j.bmcl.2009.12.062

Hogg, T., Nagarajan, K., Herzberg, S., Chen, L., Shen, X., Jiang, H., et al. (2006). Structural and functional characterization of Falcipain-2, a hemoglobinase from the malarial parasite *Plasmodium falciparum*. *J. Biol. Chem*. 281, 25425–25437. doi: 10.1074/jbc.M603776200

Huang, H., Lu, W., Li, X., Cong, X., Ma, H., Liu, X., et al. (2012). Design and synthesis of small molecular dual inhibitor of falcipain-2 and dihydrofolate reductase as antimalarial agent. *Bioorg. Med. Chem. Lett*. 22, 958–962. doi: 10.1016/j.bmcl.2011.12.011

Jadhav, A., Ferreira, R. S., Kulmpp, C., Mott, B. T., Austin, C. P., Inglese, J., et al. (2010). Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. *J. Med. Chem*. 53, 37–51. doi: 10.1021/jm901070c

Jin, G., and Wong, S. T. C. (2015). Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug Discov. Today* 19, 637–644. doi: 10.1016/j.drudis.2013.11.005

Jin, H., Xu, Z., Cui, K., Zhang, T., Lu, W., and Huang, J. (2014). Dietary flavonoids fisetin and myricetin: dual inhibitors of *Plasmodium falciparum* falcipain-2 and plasmepsin II. *Fitoterapia* 94, 55–61. doi: 10.1016/j.fitote.2014.01.017

Kassahun, K., McIntosh, I., Koeplinger, K., Sun, L., Talaty, J. E., Miller, D. L., et al. (2014). Disposition and metabolism of the cathepsin K inhibitor odanacatib in humans. *Drug. Metab. Dispos*. 42, 818–827. doi: 10.1124/dmd.113.056580

Kerr, I. D., Lee, J. H., Pandey, K. C., Harrison, A., Sajid, M., Rosenthal, P. J., et al. (2009). Structures of falcipain-2 and falcipain-3 bound to small molecule inhibitors: implications for substrate specificity. *J. Med. Chem*. 52, 852–857. doi: 10.1021/jm8013663

Kisinza, W. N., Nkya, T. E., Kabula, B., Overgaard, H. J., Massue, D. J., Mageni, Z., et al. (2017). Multiple insecticide resistance in *Anopheles gambiae* from Tanzania: a major concern for malaria vector control. *Malar. J*. 16, 439. doi: 10.1186/s12936-017-2087-2

Krishnaswamy, S., and Betz, A. (1997). Exosites determine macromolecular substrate recognition by prothrombinase. *Biochemistry* 36, 12080–12086.

Lätti, S., Niinivehmas, S., and Pentikäinen, O. T. (2016). Rocker: Open source, easy-to-use tool for AUC and enrichment calculations and ROC visualization. *J. Cheminform*. 8:45. doi: 10.1186/s13321-016-0158-y

Lauria, A., Tutone, M., and Almerico, A. M. (2011). Virtual lock-and-key approach: the *in silico* revival of Fischer model by means of molecular descriptors. *Eur. J. Med. Chem*. 46, 4274–4280. doi: 10.1016/j.ejmech.2011.06.033

Lauria, A., Tutone, M., Barone, G., and Almerico, A. M. (2014). Multivariate analysis in the identification of biological targets for designed molecular structures: the BIOTA protocol. *Eur. J. Med. Chem*. 75, 106–110. doi: 10.1016/j.ejmech.2014.01.025

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res*. 42, 1091–1097. doi: 10.1093/nar/gkt1068

Leonard, J. T., and Roy, K. (2006). On selection of training and test sets for the development of predictive QSAR models. *QSAR Comb. Sci*. 25, 235–251. doi: 10.1002/qsar.200510161

Li, H., Huang, J., Chen, L., Liu, X., Chen, T., Zhu, J., et al. (2009). Identification of novel falcipain-2 inhibitors as potential antimalarial agents through structure-based virtual screening. *J.Chem. Eng. Data* 63, 4735–4470. doi: 10.1021/jm801622x

Liu, W., Lu, H., Cao, C., Jiao, Y., and Chen, G. (2018). An improved Quantitative Structure Property Relationship model for predicting thermal conductivity of liquid aliphatic alcohols. *J. Chem. Eng. Data* 63, 4735–4740. doi: 10.1021/acs.jced.8b00764

Luo, W., Lu, W., Cui, K., Liu, Y., Wang, J., and Guo, C. (2012). N1-{4-[(10S)-Dihydroartemisinin-10-oxyl]}phenylmethylene-N2-(2-methylquinoline-4-yl)hydrazine derivatives as antiplasmodial falcipain-2 inhibitors. *Med. Chem. Res*. 21, 3073–3079. doi: 10.1007/s00044-011-9854-3

Marco, M., and Coterón, J. M. (2012). Falcipain inhibition as a promising antimalarial target. *Curr. Top. Med. Chem*. 12, 408–444. doi: 10.2174/156802612799362913

Marques, A. F., Esser, D., Rosenthal, P. J., Kassack, M. U., and Lima, L. M. (2013). Falcipain-2 inhibition by suramin and suramin analogues. *Bioorg. Med. Chem*. 21, 3667–3673. doi: 10.1016/j.bmc.2013.04.047

Marques, A. F., Gomes, P. S., Oliveira, P. L., Rosenthal, P. J., Pascutti, P. G., and Lima, L. M. (2015). Allosteric regulation of the *Plasmodium falciparum* cysteine protease falcipain-2 by heme. *Arch. Biochem. Biophys*. 573, 92–99. doi: 10.1016/j.abb.2015.03.007

Martin, T. M., Harten, P., Young, D. M., Muratov, E. N., Golbraikh, A., Zhu, H., et al. (2012). Does rational selection of training and test sets improve the outcome of QSAR modeling? *J. Chem. Inf. Model*. 52, 2570–2578. doi: 10.1021/ci300338w

Mercurio, A., Adriani, G., Catalano, A., Carocci, A., Rao, L., Lentini, G., et al. (2017). A mini-review on thalidomide: chemistry, mechanisms of action, therapeutic potential and anti-angiogenic properties in multiple myeloma. *Curr. Med. Chem*. 24, 2736–2744. doi: 10.2174/0929867324666170601074646

Micale, N., Kozikowski, A. P., Ettari, R., Grasso, S., Zappalà, M., Jeong, J. J., et al. (2006). Novel peptidomimetic cysteine protease inhibitors as potential antimalarial agents. *J. Med. Chem*. 49, 3064–3067. doi: 10.1021/jm060405f

Min, S. H. (2016). A genetic algorithm-based heterogeneous random subspace ensemble model for bankruptcy prediction. *Int. J. Appl. Eng. Res*. 11, 2927–2931.

Morrison, J. F. (1969). Kinetics of the reversible inhibition of enzyme-catalysed reactions by tight-binding inhibitors. *Biochim. Biophys. Acta* 185, 269–286. doi: 10.1016/0005-2744(69)90420-3

Mugumbate, G., Newton, A. S., Rosenthal, P. J., Gut, J., Moreira, R., Chibale, K., et al. (2013). Novel anti-plasmodial hits identified by virtual screening of the ZINC database. *J. Comput. Aided Mol. Des*. 27, 859–871. doi: 10.1007/s10822-013-9685-z

Mundra, S., and Radhakrishnan, M. (2015a). Insignificant *in vitro* falcipain-2 inhibitory activity of novel 2-(4-(substituted benzoyl) piperazine-1-yl)-N-phenylacetamide derivatives. *J. Pharm. Negative Results* 6, 20–26. doi: 10.4103/0976-9234.157381

Mundra, S., and Radhakrishnan, M. (2015b). Evaluation of novel 1-(4-(substituted)piperazin-1-yl)- 2-(phenylamino)ethanone derivatives as Falcipain-2 inhibitors. *J. Young Pharm*. 7, 96–105. doi: 10.5530/jyp.2015.2.7

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem*. 55, 6582–6594. doi: 10.1021/jm300687e

Nizi, E., Sferrazza, A., Fabbrini, D., Nardi, V., Andreini, M., Graziani, R., et al. (2018). Peptidomimetic nitrile inhibitors of malarial protease falcipain-2 with high selectivity against human cathepsins. *Bioorg. Med. Chem. Lett*. 28, 1540–1544. doi: 10.1016/j.bmcl.2018.03.069

Novick, P. A., Ortiz, O. F., Poelman, J., Abdulhay, A. Y., and Pande, V. S. (2013). SWEETLEAD: an *in silico* database of approved drugs, regulated chemicals, and herbal isolates for computer-aided drug discovery. *PLoS ONE* 8:e79568. doi: 10.1371/journal.pone.0079568

Oballa, R. M., Truchon, J. F., Baily, C. I., Chauret, N., Day, S., Crane, S., et al. (2007). A generally applicable method for assessing the electrophilicity and reactivity of diverse nitrile-containing compounds. *Bioorg. Med. Chem. Lett*. 17, 998–1002. doi: 10.1016/j.bmcl.2006.11.044

Pandey, K. C., Wang, S. X., Sijwali, P. S., Lau, A. L., McKerrow, J. H., and Rosenthal, P. J. (2005). The *Plasmodium falciparum* cysteine protease falcipain-2 captures its substrate, hemoglobin, via a unique motif. *Proc. Natl. Acad. Sci. U.S.A*. 102, 9138–9143. doi: 10.1073/pnas.0502368102

Pant, A., Kumar, R., Wani, N. A., Verma, S., Sharma, R., Pande, V., et al. (2018). Allosteric site inhibitor disrupting auto-processing of malarial cysteine proteases. *Sci. Rep*. 8, 16193. doi: 10.1038/s41598-018-34564-8

Pedicord, D. L., Seiffert, D., and Blat, Y. (2004). Substrate-dependent modulation of the mechanism of factor XIa inhibition. *Biochemistry* 43, 11883–11888. doi: 10.1021/bi048964g

Perez-Llamas, C., and Lopez-Bigas, N. (2011). Gitools: analysis and visualisation of genomic data using interactive heat-maps. *PLoS ONE* 6:e19541. doi: 10.1371/journal.pone.0019541

Potshangbam, A. M., Tanneeru, K., Reddy, B. M., and Guruprasad, L. (2011). 3D-QSAR and molecular docking studies of 2-pyrimidinecarbonitrile derivatives as inhibitors against falcipain-3. *Bioorg. Med. Chem. Lett.* 21, 7219–7223. doi: 10.1016/j.bmcl.2011.09.107

Pradines, B., Rogier, C., Fusai, T., Mosnier, J., Daries, W., Barret, E., et al. (2001). *In vitro* activities of antibiotics against *Plasmodium falciparum* are inhibited by iron. *Antimicrob. Agents Chemother.* 2001, 45, 1746–1750. doi: 10.1128/AAC.45.6.1746-1750.2001

Praveen Kumar, S., Gut, J., Guedes, R. C., Rosenthal, P. J., Santos, M. M., and Moreira, R. (2011). Design, synthesis and evaluation of 3-methylene-substituted indolinones as antimalarials. *Eur. J. Med. Chem*. 46, 927–933. doi: 10.1016/j.ejmech.2011.01.008

Reutlinger, M., Koch, C. P., Reker,. D., Todoroff, N., Schneider, P., Rodrigues, T., et al. (2013). Chemically Advanced Template Search (CATS) for scaffold-hopping and prospective target prediction for 'orphan' molecules. *Mol. Inform.* 32, 133–138. doi: 10.1002/minf.201200141

Ritchie, H., and Roser, M. (2018). *Causes of Death*. OurWorldInData.org [Online]. Available online at: https://ourworldindata.org/causes-of-death (accessed August, 2018).

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. doi: 10.1186/1471-2105-12-77

Roy, K., and Mitra, I. (2011). On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Comb Chem. High Throughput Screen* 14, 450–474.

Sanchez Mejia, R. O., Ona, V. O., Li, M., and Friedlander, R. M. (2001). Minocycline reduces traumatic brain injury-mediated caspase-1 activation, Tissue damage, and neurological dysfunction. *Neurosurgery* 48, 1393–1401. doi: 10.1097/00006123-200106000-00051

Satimai, W., Sudathip, P., Vijaykadga, S., Khamsiriwatchara, A., Sawang, S., Potithavoranan, T., et al. (2012). Artemisinin resistance containment project in Thailand. II: responses to mefloquine-artesunate combination therapy among falciparum malaria patients in provinces bordering Cambodia. *Malar. J*. 11:300. doi: 10.1186/1475-2875-11-300

Schmidt, I., Pradel, G., Sologub, L., Golzmann, A., Ngwa, C. J., Kucharski, A., et al. (2016). Bistacrine derivatives as new potent antimalarials. *Bioorg. Med. Chem*. 24, 3636–3642. doi: 10.1016/j.bmc.2016.06.003

Selwyn, M. J. (1965). A simple test for inactivation of an enzyme during assay. *Biochim. Biophys. Acta* 105, 193–195. doi: 10.1016/S0926-6593(65)80190-4

Shah, F., Mukherjee, P., Gut, J., Legac, J., Rosenthal, P. J., Tekwani, B. L., et al. (2011). Identification of novel malarial cysteine protease inhibitors using structure-based virtual screening of a focused cysteine protease inhibitor library. *J. Chem. Inf. Model*. 51, 852–864. doi: 10.1021/ci200029y

Sharma, R. K., Younis, Y., Mugumbate, G., Njoroge, M., Gut, J., Rosenthal, P. J., et al. (2015). Synthesis and structure-activity-relationship studies of thiazolidinediones as antiplasmodial inhibitors of the *Plasmodium falciparum* cysteine protease falcipain-2. *Eur. J. Med. Chem*. 90, 507–518. doi: 10.1016/j.ejmech.2014.11.061

Shenai, B. R., Lee, B. J., Alvarez-Hernandez, A., Chong, P. Y., Emal, C. D., Neitz, R. J., et al. (2003). Structure-activity relationships for inhibition of cysteine protease activity and development of *Plasmodium falciparum* by peptidyl vinyl sulfones. *Antimicrob. Agents Chemother*. 47, 154–160. doi: 10.1128/aac.47.1.154-160.2003

Singh, A. K., Rajendran, V., Pant, A., Ghosh, P. C., Singh, N., Latha, N., et al. (2015). Design, synthesis and biological evaluation of functionalized phthalimides: a new class of antimalarials and inhibitors of falcipain-2, a major hemoglobinase of malaria parasite. *Bioorg. Med. Chem*. 23, 1817–1827. doi: 10.1016/j.bmc.2015.02.029

Stoye, A., Juillard, A., Tang, A. H., Legac, J., Gut, J., White, K. L., et al. (2019). Falcipain inhibitors based on the natural product gallinamide a are potent *in vitro* and *in vivo* antimalarials. *J. Med. Chem*. 62, 5562–5578. doi: 10.1021/acs.jmedchem.9b00504

Subramanian, S., Hardt, M., Choe, Y., Niles, R. K., Johansen, E. B., Legac, J., et al. (2009). Hemoglobin cleavage site-specificity of the *Plasmodium falciparum* cysteine proteases falcipain-2 and falcipain-3. *PLoS ONE* 4:e5156. doi: 10.1371/journal.pone.0005156

Talevi, A. (2018). Drug repositioning: current approaches and their implications in the precision medicine era. *Expert. Rev. Precis. Med. Drug Dev*. 3, 49–61. doi: 10.1080/23808993.2018.1424535

Teixeira, C., Gomes, J. R. B., Couesnon, T., and Gomes, P. (2011). Molecular docking and 3D-quantitative structure activity relationship analyses of peptidyl vinyl sulfones: *Plasmodium falciparum* cysteine proteases inhibitors. *J. Comput. Aided Mol. Des*. 25, 763–775. doi: 10.1007/s10822-011-9459-4

Teo, S., Resztak, K. E., Scheffler, M. A., Kook, K. A., Zeldis, J. B., Stirling, D. I., et al. (2002). Thalidomide in the treatment of leprosy. *Microbes Infect.* 4, 1193–1202. doi: 10.1016/S1286-4579(02)01645-3

Thorne, N., Auld, D. S., and Inglese, J. (2010). Apparent activity in high-throughput screening: origins of compound-dependent assay interference. *Curr. Opin. Chem. Biol.* 14, 315–324. doi: 10.1016/j.cbpa.2010.03.020

Toropova, A. P., and Toropov, A. A. (2017). CORAL: binary classifications (active/inactive) for drug-induced liver injury. *Toxicol. Lett*. 268, 51–57. doi: 10.1016/j.toxlet.2017.01.011

Truchon, J. F., and Bayly, C. L. (2007). Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model*. 47, 488–508. doi: 10.1021/ci600426e

Tutone, M., Perricone, U., and Almerico, A. M. (2017). Conf-VLKA: a structure-based revisitation of the Virtual Lock-and-key Approach. *J. Mol. Graph. Model.* 71, 50–57. doi: 10.1016/j.jmgm.2016.11.006.

Valente, C., Guedes, R. C., Moreira, R., Iley, J., Gut, J., and Rosenthal, P. J. (2006). Dipeptide vinyl sultams: synthesis via the Wittig-Horner reaction and activity against papain, falcipain-2 and *Plasmodium falciparum*. *Bioorg. Med. Chem. Lett*. 16, 4115–4119. doi: 10.1016/j.bmcl.2006.04.079

Volkamer, A., Kuhn, D., Rippmann, F., and Rarey, M. (2012). DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics* 28, 2074–2075. doi: 10.1093/bioinformatics/bts310

Wang, J., Li, F., Li, Y., Yang, Y., Zhang, S., and Yang, L. (2013). Structural features of falcipain-3 inhibitors: an *in silico* study. *Mol. Biosyst*. 9, 2296–2310. doi: 10.1039/c3mb70105k

Wang, L., Zhang, S., Zhu, J., Zhu, L., Liu, X., Shan, L., et al. (2014). Identification of diverse natural products as falcipain-2 inhibitors through

structure-based virtual screening. *Bioorg. Med. Chem. Lett*. 24, 1261–1264. doi: 10.1016/j.bmcl.2014.01.074

Wang, S. X., Pandey, K. C., Somoza, J. R., Sijwali, P. S., Kortemme, T., Brinen, L. S., et al. (2006). Structural basis for unique mechanisms of folding and hemoglobin binding by a malarial protease. *Proc. Natl. Acad. Sci. U.S.A.* 103, 11503–11508. doi: 10.1073/pnas.0600489103

Weldon, D. J., Shah, F., Chittiboyina, A. G., Sheri, A., Chada, R. R., Gut, J., et al. (2014). Synthesis, biological evaluation, hydration site thermodynamics, and chemical reactivity analysis of α-keto substituted peptidomimetics for the inhibition of *Plasmodium falciparum*. *Bioorg. Med. Chem. Lett*. 24, 1274–1279. doi: 10.1016/j.bmcl.2014.01.062

World Health Organization (WHO) (2017). *World Malaria Report 2017*. Geneva. Available online at: http://www.who.int/malaria/publications/world-malaria-report-2017/en/ (accessed August 2018).

Xue, C. X., Cui, S. Y., Liu, M. C., Hu, Z. D., and Fan, B. T. (2004). 3D QSAR studies on antimalarial alkoxylated and hydroxylated chalcones by CoMFA and CoMSIA. *Eur. J. Med. Chem.* 39, 745–753. doi: 10.1016/j.ejmech.2004.05.009

Yabuuchi, H., Niijima, S., Takematsu, H., Ida, T., Hirokawa, T., Hara, T., et al. (2011). Analysis of multiple compound–protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol.* 7, 472. doi: 10.1038/msb.2011.5

Yang, J., Copeland, R. A., and Lai, Z. (2009). Defining balanced conditions for inhibitor screening assays that target bisubstrate enzymes. *J. Biomol. Screen.* 14, 111–120. doi: 10.1177/1087057108328763

Yella, J. K., Yaddanapudi, S., Wang, Y., and Jegga, A. G. (2018). Changing trends in computational drug repositioning. *Pharmaceuticals* 11:57. doi: 10.3390/ph11020057

Yu, G., Zhang, G., Domeniconi, C., Yu, Z., and You, J. (2012). Semi-supervised classification based on random subspace dimensionality reduction. *Pattern Recogn*. 45, 1119–1135. doi: 10.1016/j.patcog.2011.08.024

Zhang, Q., and Muegge, I. (2006). Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. *J. Med. Chem*. 49, 1536–1548. doi: 10.1021/jm050468i

Zucker, S., Lysik, R. M., Ramamurthy, N. S., Golub, L. M., Wieman, J. M., and Wilkie, D. P. (1985). Diversity of melanoma plasma membrane proteinases: inhibition of collagenolytic and cytolytic activities by minocycline. *J. Natl. Cancer Inst*. 75, 517–525. doi: 10.1093/jnci/75.3.517

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Alberca, Chuguransky, Álvarez, Talevi and Salas-Sarduy. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 13 August 2019
doi: 10.3389/fchem.2019.00574

Identification of Novel 3-Hydroxy-pyran-4-One Derivatives as Potent HIV-1 Integrase Inhibitors Using *in silico* Structure-Based Combinatorial Library Design Approach

Hajar Sirous[1†], Giulia Chemi[2†], Sandra Gemma[2], Stefania Butini[2], Zeger Debyser[3], Frauke Christ[3], Lotfollah Saghaie[4], Simone Brogi[5], Afshin Fassihi[4*], Giuseppe Campiani[2*] and Margherita Brindisi[6]*

[1] Bioinformatics Research Center, School of Pharmacy and Pharmaceutical Sciences, Isfahan University of Medical Sciences, Isfahan, Iran, [2] Department of Biotechnology, Chemistry and Pharmacy, Department of Excellence 2018-2022, University of Siena, Siena, Italy, [3] Molecular Medicine, K.U. Leuven and IRC KULAK, Leuven, Belgium, [4] Department of Medicinal Chemistry, Faculty of Pharmacy, Isfahan University of Medical Sciences, Isfahan, Iran, [5] Department of Pharmacy, University of Pisa, Pisa, Italy, [6] Department of Pharmacy, Department of Excellence 2018-2022, University of Naples Federico II, Naples, Italy

We describe herein the development and experimental validation of a computational protocol for optimizing a series of 3-hydroxy-pyran-4-one derivatives as HIV integrase inhibitors (HIV INIs). Starting from a previously developed micromolar inhibitors of HIV integrase (HIV IN), we performed an in-depth investigation based on an *in silico* structure-based combinatorial library designing approach. This method allowed us to combine a combinatorial library design and side chain hopping with Quantum Polarized Ligand Docking (QPLD) studies and Molecular Dynamics (MD) simulation. The combinatorial library design allowed the identification of the best decorations for our promising scaffold. The resulting compounds were assessed by the mentioned QPLD methodology using a homology model of full-length binary HIV IN/DNA for retrieving the best performing compounds acting as HIV INIs. Along with the prediction of physico-chemical properties, we were able to select a limited number of drug-like compounds potentially displaying potent HIV IN inhibition. From this final set, based on the synthetic accessibility, we further shortlisted three representative compounds for the synthesis. The compounds were experimentally assessed *in vitro* for evaluating overall HIV-1 IN inhibition, HIV-1 IN strand transfer activity inhibition, HIV-1 activity inhibition and cellular toxicity. Gratifyingly, all of them showed relevant inhibitory activity in the *in vitro* tests along with no toxicity. Among them **HPCAR-28** represents the most promising compound as potential anti-HIV agent, showing inhibitory activity against HIV IN in the

OPEN ACCESS

Edited by:
Alfonso Carotenuto, University of Naples Federico II, Italy

Reviewed by:
John Joseph Deadman, Independent Researcher, Melbourne, Australia
Simona Collina, University of Pavia, Italy

****Correspondence:***
Simone Brogi
simone.brogi@unipi.it
Afshin Fassihi
fassihi@pharm.mui.ac.ir
Giuseppe Campiani
campiani@unisi.it

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *26 June 2019*
Accepted: *29 July 2019*
Published: *13 August 2019*

Citation:
Sirous H, Chemi G, Gemma S, Butini S, Debyser Z, Christ F, Saghaie L, Brogi S, Fassihi A, Campiani G and Brindisi M (2019) Identification of Novel 3-Hydroxy-pyran-4-One Derivatives as Potent HIV-1 Integrase Inhibitors Using in silico Structure-Based Combinatorial Library Design Approach. Front. Chem. 7:574. doi: 10.3389/fchem.2019.00574

low nanomolar range, comparable to that found for Raltegravir, and relevant potency in inhibiting HIV-1 replication and HIV-1 IN strand transfer activity. In summary, our results outline **HPCAR-28** as a useful optimized hit for the potential treatment of HIV-1 infection by targeting HIV IN.

Keywords: 3-hydroxy-pyran-4-one, HIV-1 integrase inhibitors (HIV-1 INIs), *in silico* combinatorial library design, side chain hopping, hit compounds optimization

INTRODUCTION

HIV-1 integrase (IN) represents an attractive target in anti-HIV drug design mainly due to its specificity. Accordingly, HIV-1 IN does not have a functional equivalent in humans and plays a unique role in establishing irreversible and productive viral infections (Debyser et al., 2002; Delelis et al., 2008). This viral key enzyme catalyzes the insertion of proviral DNA, derived from reverse transcription of HIV-1 RNA, into the genome of the host-infected cells. The insertion is achieved through a two-step enzymatic process which starts with endonucleolytic cleavage of a terminal dinucleotide (GT) from each 3′-end of the proviral DNA (termed "3′-processing"), followed by a second reaction, known as "strand transfer" (ST), involving a concerted nucleophilic attack, by the reactive 3′-OH ends of the viral processed DNA to the host chromosomal DNA. As a result, a the covalent joining of the two DNA strands is observed (Chiu and Davies, 2004; Pommier et al., 2005). Both reactions are accomplished by the catalytic core domain of HIV-1 IN which contains two divalent metal ion cofactors (Mg^{2+}). These metal ions are coordinated by three catalytic carboxylate residues: Asp_{64}, Asp_{116}, and Glu_{152} (DDE triad) within the enzyme active site (Dyda et al., 1994; Neamati et al., 2002).

Targeting the metal cofactors within the active site of a viral metal-activated enzyme like HIV-1 IN has emerged as an attractive and validated strategy for the development of novel anti-HIV agents (Rogolino et al., 2012). With this aim, a metal binding pharmacophore model has been exploited for the design of diverse HIV-1 integrase inhibitors (HIV-1 INIs) as depicted in **Figure 1A**. This model is represented by two distinctive structural features: (1) a planar metal binding group (MBG), able to interact with the metal centers within the IN active site, and (2) a pendent aromatic or hetero-aromatic hydrophobic moiety located in close proximity of the MBG (Kawasuji et al., 2006a,b; Johns and Svolto, 2008). Continuous efforts in exploiting this pharmacophore model have culminated in the design and subsequent FDA approval of three INIs for clinical use as effective anti-HIV drugs: Raltegravir (**RLT**), Elvitegravir (**EVG**), and Dolutegravir (**DTG**) in 2007, 2012, and 2013, respectively (**Figure 1B**; Rowley, 2008; Sato et al., 2009; Katlama and Murphy, 2012).

A variety of MBGs have been extensively studied to design innovative and effective INIs (Liao et al., 2010; Di Santo, 2014). Recently, we were particularly interested in taking advantage of the 3-hydroxy-4-pyranone (HP) scaffold for the development of novel HIV-1 INIs due to its application as MBG in the design of several inhibitors of numerous Zn^{2+}, Mg^{2+}, Mn^{2+}, and Cu^{2+} dependent proteins. Accordingly, HP derivatives represent an impressive class of heterocyclic ligands with strong bidentate chelating capacity toward metal ions (Santos et al., 2012; Rostami et al., 2015; Sirous et al., 2015). As a first example of the potential of this structural template in HIV-IN inhibition, a series of HP compounds featuring a unique C-2 carboxamide moiety, namely 3-hydroxyl-pyran-4-one-2-carboxamide derivatives (HPCARs), were rationally designed and recently reported by us (**Figure 2**; Sirous et al., 2019). The proposed chemotypes were characterized by a chelating triad motif effectively coordinating the two metals according to the pharmacophore shared by INIs. Moreover, an aromatic backbone attached to the amide portion through a linker (substituted benzyl and phenylethyl moieties) was considered for providing the essential interactions with the hydrophobic pocket of the enzyme. Most of these HPCAR analogs offered favorable inhibitory potencies in both enzymatic and cell-based antiviral assays with low micromolar IC_{50} values. In particular, the substitution at the *para* position of the aromatic phenyl ring led to the identification of two halo-benzyl derivatives **HPb** and **HPd** (**Figure 2**) as promising lead HIV-1 IN inhibitors with IC_{50} values of 0.37 and 0.7 μM, respectively (Sirous et al., 2019).

In our quest for the search of innovative and effective INIs and considering the above-mentioned findings, we decided to design novel optimized derivatives exploiting the HPCAR chemotype. In this study, we performed the replacement of the pendant aromatic portion with other heterocyclic moieties in order to maintain the strong hydrophobic interactions within the HIV-1 IN binding site, with the possibility to explore additional functional groups for maximizing the contacts that could further stabilize the binding mode of the novel derivatives, leading to compounds with improved activity against HIV-1 IN. Accordingly, in the present study, an *in silico* protocol combining a combinatorial library design procedure coupled to extensive molecular docking studies and physico-chemical properties prediction was developed in a step-filtering approach to identify novel INIs with improved potency with respect to the HPCAR derivatives. The employed screening workflow for designing new HIV-1 INIs with suitable potency and satisfactory physico-chemical properties is illustrated in **Figure 3**. Considering the importance of the aromatic portion of INIs for their binding to both the viral DNA bases, and the hydrophobic pocket within the catalytic core of IN enzyme (Kawasuji et al., 2006b), many efforts were made for replacing the hydrophobic aromatic side chains (substituted benzyl and phenylethyl moieties) to generate a virtual combinatorial library of HP-based core derivatives. Accordingly, using side chain

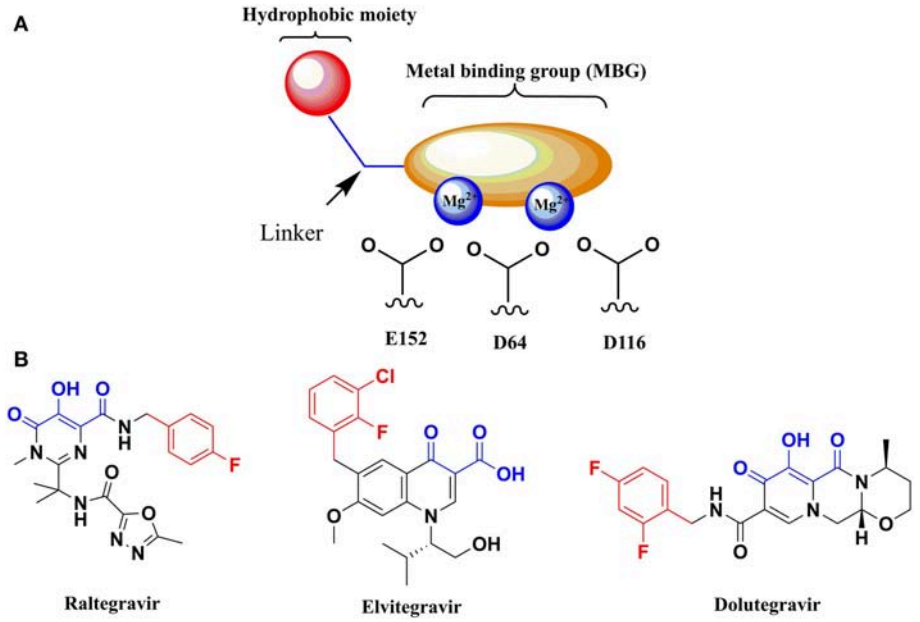


FIGURE 1 | (A) Graphical depiction of the pharmacophore model for HIV-1 INIs. **(B)** Chemical structure of the FDA approved HIV-1 INIs. Atoms in blue are part of MBG of the molecules able to chelate the two metal ions. The hydrophobic aromatic moiety of each compound is highlighted in red.

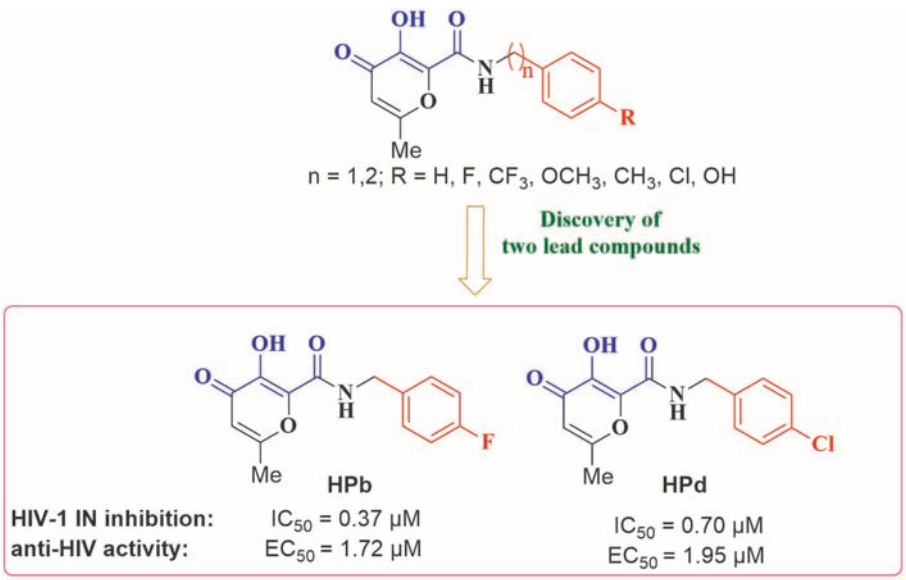


FIGURE 2 | General structure of 3-hydroxy-4-pyranone (HP) based inhibitors previously designed in our laboratory and two halogenated derivatives identified as lead compounds. The proposed chemotype satisfies minimal pharmacophore features for HIV-1 IN inhibition: the metal chelating triad (blue) and terminal hydrophobic benzyl group (red). The most promising HP compounds are shown with the IC_{50} and EC_{50} values for IN inhibitory and anti-HIV activities.

hopping strategy, various cyclic and heterocyclic fragments were attached to the defined position of the HP core in order to find the ideal sidechains with the highest predicted binding affinity for the IN active site. Finally, for validating the computational approach three representative hit candidates identified from this screening workflow were selected, further studied by molecular dynamics (MD) simulations in order to gain additional information about their mechanism of action as INIs, synthesized and submitted to biological evaluation for their HIV-1 IN inhibitory and anti-HIV-1 activities.

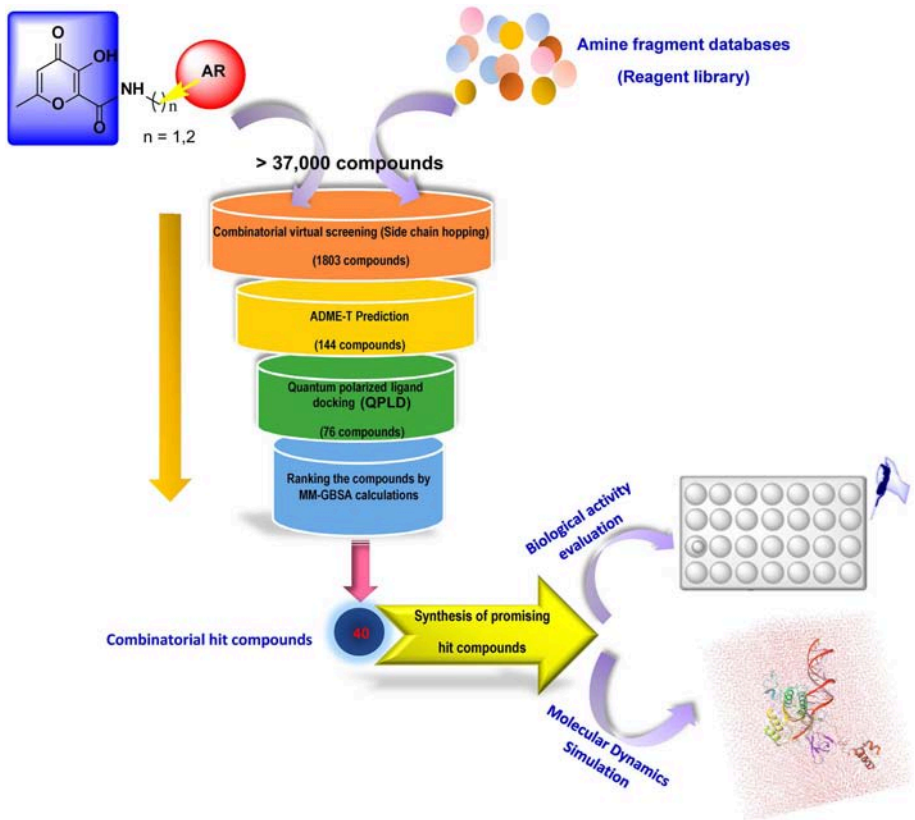


FIGURE 3 | Screening workflow employed in the present study for the *in silico* rational design of new HPCAR derivatives. The core of HPCAR was marked in a blue box. The hydrophobic fragment, AR group, is depicted as a red circle. During combinatorial library screening, diverse sets of fragments were provided from reagent file libraries and attached to the main core at the defined attach positions. The attach position in the scaffold structure is indicated as a yellow vector.

MATERIALS AND METHODS

Computational Details

Ligand Preparation

The 3D structure of the two investigated HPCAR derivatives (substituted with benzyl and phenylethyl moieties) were built by the 3D-sketcher module in Maestro suite (Maestro, version 9.2; Schrödinger, LLC, New York, NY, 2011). Molecular energy minimization of the structures was performed in MacroModel environment using the OPLS-AA 2005 as force field (Jorgensen et al., 1996; Kaminski et al., 2001). GB/SA model was utilized in order to simulate the solvent effects applying "no cut-off" for non-bonded interactions (Still et al., 1990). PRCG method with 1,000 maximum iterations and 0.001 gradient convergence threshold was employed. The same protocol was applied to the novel designed compounds obtained by the combinatorial screening (144 molecules) before submitting them to QPLD procedure. Furthermore, all the compounds were accurately prepared with LigPrep application implemented in Maestro suite (Gasser et al., 2015). Finally the most probable ionization state of the compounds was retrieved by Chemicalize (https://chemicalize.com/) as already reported by us (Brogi et al., 2018).

Protein Preparation

Computational studies were conducted using our recently described theoretical model of full-length HIV-1 IN in complex with viral DNA and Mg^{2+} cofactors (Sirous et al., 2019). The model was subjected to Protein Preparation Wizard protocol implemented in Maestro. This protocol allowed us to obtain a reasonable starting structure of the protein for molecular modeling calculations by a series of computational steps as described (Cappelli et al., 2013; Brogi et al., 2017a,b). Finally, the refined HIV-1 IN model was used for further computational studies.

Generation of Combinatorial Hits

"Combinatorial library enumeration" option available in CombiGlide (CombiGlide, version 2.7; Schrödinger, LLC: New York, 2011), a combinatorial screening software distributed by Schrödinger, was used to carry out structure-based combinatorial library design studies. This software provides the tools for accelerating the lead optimization process, helping in the generation of libraries of optimized derivatives to be selected for the further synthesis. In this direction, two HPCAR derivatives containing methylene and ethylene linkers between

the chelating region and the aromatic moiety, identified from our previous studies (Sirous et al., 2019), were selected as main cores. For each investigated compound, a side chain hopping strategy was successfully applied for replacing the hydrophobic side chain, of the selected main cores, with different aromatic or heteroaromatic fragments as shown in **Figure 3**. This method employs the reagent files, chosen by the operator, as a source of fragments with various structures. The following steps are used to generate a new combinatorial library of ligands.

Reagents Preparation

In this step of CombiGlide workflow, a library of reagents containing diverse sets of fragments was built. The elements of this library can be selected from the available databases or generated by the operator. In fact, in the presented work, in addition to the reagent libraries provided by Maestro software, other reagent libraries with different aromatic groups were downloaded from Zinc fragment database (Irwin and Shoichet, 2005; Irwin et al., 2012) as SDF file format and submitted to the reagent preparation facility in the CombiGlide environment. Then the tasks in the reagent preparation process are: (i) the selection of the source of reagent structures; (ii) the selection of a reagent type (a functional group), and finally (iii) the structural conversion from the 2D structure to the 3D one. The selection of a reagent type was done considering the bond that will be replaced in accordance to the functional group formed when the reagent is added to the core. In this context, primary amine set was selected as reagent type. The detailed description of this reagent type is provided in **Table 1**. Concerning this reagent type, R represents the part of reagent that was kept in the process of combinatorial library generation. The bond that was broken to attach the reagent to the core was marked with a line crossing the bond. After running reagent preparation job, the output structure file in .bld format, containing properly prepared reagents, was used in the combinatorial screening process by CombiGlide.

Defining the Core and Attachments

The core is the structural element that is maintained throughout the combinatorial experiment. The attachment positions for each core were defined and the previously prepared reagents file was associated with each attachment point. The attachment point comprises bonds from the core structure that will be replaced in the build process. Considered the role of the hydrophobic side chain of the investigated HPCAR derivatives in binding both the viral DNA and the hydrophobic pocket within IN active site (Sirous et al., 2019), the replacement of benzyl and phenethyl amine moieties with different amine fragments from the reagents file, with other hydrophobic moieties, was performed.

TABLE 1 | Detailed description of predefined functional group type selected in the reagents preparation step of combinatorial library design.

Structure of reagent type	Name of reagent type	Definition of R
	Amine_Primary_ General_N_H	R can be alkyl, aryl. R cannot have a carbonyl carbon attached to the nitrogen of the amine.

Setting Up CombiGlide Docking Calculations

The docking step represents the main step of the combinatorial screening process in which a series of docking calculations are performed to screen out molecules that do not have satisfactory docking scores. In fact, on a core with a constant structure, CombiGlide attaches sidechains at defined positions of the core, and performs a docking calculation of the resulting compounds into the active site of HIV-1 IN, to assess the potential affinity of the new compounds. The grid box for the docking calculation was centered on the centroid between the two Mg^{2+} ions which roughly represents the center of the active site. The cubic grid box was adjusted based on a size capable of accommodating ligands with a length 15 Å. As part of grid generation procedure, metal constraints for the receptor grids were also applied. The other options and parameters in this step were set as default and then docking of the library members into the homology modeled HIV-1 IN active site was performed using the extra precision method (XP) in CombiGlide docking. At the end of the process a focused combinatorial library of more than 37,000 compounds was obtained for each studied core-containing molecule. The total structures obtained from combinatorial screening were sorted on the basis of their GlideScore (Glide, Version 5.7, Schrödinger, LLC, New York, NY, 2011; Friesner et al., 2004). The compounds with the better XP Glide scores compared with the corresponding core-containing molecules were selected for further studies. The interactions of these compounds, into HIV-1 IN active site, were assessed by using ligand-interaction diagram implemented in Maestro suite and visualized by PyMOL (PyMOL Molecular Graphics System, Version 1.6-alpha, Schrödinger, LLC, New York, NY, 2013).

Molecular Properties Prediction

The molecules selected from *in silico* combinatorial screening were evaluated using a series of filtering criteria for drug-like properties. In this regard, QikProp application (QikProp, version 3.4, Schrödinger, LLC, New York, NY, 2011) implemented in the Maestro suite was used for ADME-T properties predictions (Rostami et al., 2015; Zaccagnini et al., 2017). This step was performed to select compounds from each library with appropriate physico-chemical properties using the range values recommended by QikProp. Especially, Lipinski's rules of five, membrane permeability, lipophilicity, cardiotoxicity, or potential interaction with *h*ERG K^+ channel were considered as important criteria and investigated for filtering (Lipinski et al., 2001). Default settings were employed for these calculations. The compounds derived from the above-mentioned calculation were evaluated for their potential capability to behave as "Pan Assay Interference Compounds" (PAINS). This calculation was performed by means of FAFDrugs4.0 (http://fafdrugs4.mti.univ-paris-diderot.fr/) (Lagorce et al., 2008, 2011; Vallone et al., 2018; Brindisi et al., 2019). PAINS compounds are chemical compounds that tend to display activity against a wide range of targets by nonspecific interactions or by altering the results of the biological tests. The compounds containing this kind of

moieties, that are often present in PAINS compounds, could be false positive hits and in general should be removed from the designed series (Baell and Holloway, 2010).

Quantum Polarized Ligand Docking (QPLD)

In order to narrow down the number of the potential INIs and for improving the reliability of the protocol, the quantum polarized ligand docking (QPLD) calculations were performed for the resulting compounds with satisfactory physico-chemical properties. These compounds were docked into the modeled HIV-1 IN using QPLD protocol implemented in Schrödinger 2011 (Schrödinger Suite 2011: QM-Polarized Ligand Docking protocol; Glide, Version 5.7, Schrödinger, LLC, New York, NY, 2011; Jaguar, version 7.8, Schrödinger LLC, New York, NY, 2011; QSite version 5.7, Schrödinger LLC, New York, NY, 2011) (Irwin et al., 2012). This step was added to improve the accuracy of classical docking calculation. In fact, this procedure aims to improve the partial charges on ligand atoms by replacing them with charges derived from quantum mechanical calculations on the ligand in the field of the receptor (Paolino et al., 2018). Within the QPLD framework, the ligand atoms are treated at the Quantum Mechanical (QM) level, whereas the IN enzyme including the Mg^{2+} ions as Molecular Mechanical (MM) region are described using the OPLS force field parameters. In this way, the same grid file previously employed in the CombiGlide step was used. The best docked compounds obtained from the previous CombiGlide docking calculations followed by the evaluation of the physico-chemical properties, were selected in the ligand option. In the first step of the QPLD calculation, compounds were initially docked into the active site of IN enzyme. The initial docking calculations were carried out using Glide standard precision (SP) docking protocol, generating 5 poses per docked molecule. In the second step, the polarizable ligand charges induced by the protein field were calculated with QSite software which is coupled with Jaguar quantum mechanics engine (Jaguar, version 7.8, Schrödinger LLC, New York, NY, 2011). In this regard, the QM charge calculations of the best scoring poses for each ligand were carried out using density functional theory (DFT) method with the B3LYP/6-31G*/LACVP* basis set within the protein environment defined by the OPLS-2005 force field. Finally, the ligands with modified partial charges were redocked into the IN active site using Glide XP mode of docking considering 10 poses for each ligand. The potential inhibitors were selected based on the lower values of XP GlideScore and the key interactions between the ligand and HIV-1 IN active site.

Ligand Binding Energy Calculations

The best docked pose of ligands selected from previous QPLD calculations were subjected to a subsequent analysis with MM-GBSA process implemented in Prime software (Prime, version 3.0, Schrödinger LLC, New York, NY, 2011) (Brindisi et al., 2015). This method was employed to predict binding affinity and relative free-binding energy (ΔG_{bind}) between ligands and HIV-1 IN with further accuracy. The MM-GBSA approach combines MM energies with a continuum solvent generalized Born (GB) model for polar solvation and with a solvent-accessible surface area (SASA) for non-polar solvation term. In this way, the best ligand poses were subjected to energy minimization by local optimization feature in the Prime. During this process, the ligand strain energy was also considered. Ligand binding energies were calculated using the OPLS-2005 force field and generalized-Born/surface area continuum solvent model as previously reported by us (Brindisi et al., 2015, 2016; Maquiaveli et al., 2016; Brogi et al., 2017a; Vallone et al., 2018).

Molecular Dynamics Simulation

MD simulations studies were performed by means of Desmond 4.8 academic version, provided by D. E. Shaw Research ("DESRES"), using Maestro as graphical interface (Desmond Molecular Dynamics System, version 4.8, D. E. Shaw Research, New York, NY, 2016. Maestro-Desmond Interoperability Tools, version 4.8, Schrödinger, New York, NY, 2016). The calculation was performed using the Compute Unified Device Architecture (CUDA) API (Nickolls et al., 2008) employing two NVIDIA GPU (Brindisi et al., 2019). The calculation was performed on a system comprising 72 Intel Xeon E5-2695 v4@2.10 GHz processors and two NVIDIA GeForce 1070 GTX GPU. The complexes **HPCAR-28**/IN, **HPCAR-89**/IN, and **HPCAR-142**/IN were prepared by Protein Preparation Wizard protocol. The complexes were positioned into an orthorhombic box filled with water (TIP3P model). OPLS_2005 force field was used in MD calculation. The physiological concentration of monovalent ions (0.15 M) was simulated by adding Na^+ and Cl^- ions. Constant temperature (300 K) and pressure (1.01325 bar) were employed with NPT (constant number of particles, pressure, and temperature) as ensemble class. RESPA integrator (Humphreys et al., 1994) was used in order to integrate the equations of motion, with an inner time step of 2.0 fs for bonded interactions and nonbonded interactions within the short-range cut-off. Nose-Hoover thermostats (Hoover, 1985) were used to maintain the constant simulation temperature, and the Martyna-Tobias-Klein method (Martyna et al., 1994) was used to control the pressure. Long-range electrostatic interactions were evaluated adopting particle-mesh Ewald method (PME). The cut-off for van der Waals and short-range electrostatic interactions was set at 9.0 Å. The equilibration of the systems was performed with the default protocol provided in Desmond, which consists of a series of restrained minimizations and MD simulations used to slowly relax the system. By following this protocol, a single trajectory of 100 ns was obtained. We performed five independent MD runs for each mentioned complex with an aggregate simulation time of 0.5 μs to provide more reliable results. The trajectory files were investigated by simulation interaction diagram tools, simulation quality analysis and simulation event. The described applications were used to generate all plots regarding MD simulations analysis included in the manuscript as reported in the Results and Discussion section.

Chemistry

All reactants and reagents were purchased from Alfa Aesar and Sigma–Aldrich as "synthesis grade." Chemical reactions were monitored by analytical thin-layer chromatography (TLC) using

several solvent systems with different polarity on Merck Silica Gel 60 F_{254} (0.040–0.063 mm) with detection by UV. Merck Silica Gel 60 (0.040–0.063 mm) was used for column chromatography. ^{1}H NMR and ^{13}C NMR spectra were recorded on a Varian 300 MHz (USA) spectrometer using the residual signal of the deuterated solvent as internal standard. Splitting patterns of signals are indicated as singlet (s), doublet (d), triplet (t), multiplet (m), broad (br), and doublet of doublet (dd). The values of chemical shifts (δ) are reported in ppm and coupling constants (*J*) in hertz (Hz). Electrospray ionization-mass spectrometric (ESI-MS) were acquired with an Agilent 1100 series LC/MSD spectrometer equipped with a multimode ion source and by using methanol as solvent.

3-(Benzyloxy)-6-Methyl-4-oxo-4H-Pyran-2-Carboxylic Acid, BPCA

This key carboxylic acid intermediate was prepared according to a previously reported procedure (Sirous et al., 2019).

Procedures for the Synthesis of Amine Fragments AM (1–3)

With the aim of synthesizing the representative hit candidates, three kinds of different amine fragments were applied for amide coupling with the carboxylic acid functional group of intermediate **BPCA**. The required amine compounds were synthesized using the following methods.

Procedure for the Preparation of (3-Fluoro-5-(pyridin-2-yl)phenyl)methanamine (AM-1)

3-Fluoro-5-(pyridin-2-yl)benzonitrile (3)

Starting from 2-bromopyridine **1**, a Suzuki coupling with (3-cyano-5-fluorophenyl)boronic acid **2** catalyzed by tetrakis (triphenylphosphine) palladium (0) provided the phenyl-pyridine derivative **3**. In this reaction tetrakis was generated *in situ* from palladium (II) acetate and triphenylphosphine (Tan et al., 2014). To a vigorously stirred yellowish solution of palladium (II) acetate (0.28 g, 1.26 mmol, 0.2 eq) and triphenylphosphine (1.66 g, 6.33 mmol, 1 eq) in 4 mL dioxane/water (3:1), a premixed solution of 2-bromopyridine **1** (1.00 g, 6.33 mmol, 1 eq) in dioxane (50 mL), and a solution of potassium carbonate (2.62 g, 19 mmol, 3 eq) in water (28 mL) was added. The reaction mixture was allowed to stir for 15 min under N_2. This step was followed by the drop-wise addition of a solution of (3-cyano-5-fluorophenyl)boronic acid **2** (1.15 g, 6.96 mmol, 1.1 eq) in 8 mL dioxane/water (4:1) via a syringe. After the final addition, the reaction mixture was refluxed at 100°C in an oil bath under N_2. The progression of the reaction was monitored by TLC. The reaction was completed after 24 h. After that, the reaction mixture was cooled down to the room temperature and subsequently filtered through a short Celite pad. The filter cake was washed with dichloromethane (25 mL). The filtrate solution was diluted with water and extracted with dichloromethane (3 × 100 mL). The organic layers were combined, dried over sodium sulfate, filtered and then evaporated *in vacuo* to yield a crude product. The crude product was purified by flash chromatography on silica gel, eluting with 9:1 petroleum ether: ethyl acetate, to give **3** as a white solid (Tan et al., 2014). Yield: 64%, ^{1}H NMR (300 MHz, $CDCl_3$) δ (ppm): 8.75 (1H, d, J = 6.25 Hz, C6′-H), 8.42 (1H, d, J = 6.00 Hz, C4-H), 7.80 (2H, s, C6-H, and C2-H), 7.62–7.72 (1H, m, C4′-H), 7.22–7.40 (2H, m, C3′-H and C5′-H).

(3-Fluoro-5-(pyridin-2-yl)phenyl)methanamine (AM-1)

Dry $NiCl_2$ was prepared from hydrated $NiCl_2$. In this way, $NiCl_2 \bullet 6H_2O$ was used after drying in an oven at 250°C until its color turned from green to golden yellow. Then, it was powdered and stored in a vacuum desiccator for reaction. In a typical procedure (Caddick et al., 2003), nitrile compound **3** (0.2 g, 1.00 mmol, 1 eq) and anhydrous nickel (II) chloride (0.13 g, 1 mmol, 1 eq) were dissolved in dry ethanol (8 mL). Then, sodium borohydride (0.11 g, 3 mmol, 3 eq) was cautiously added in three portions to the vigorously stirred reaction mixture at room temperature. A black precipitate appeared during the addition of $NaBH_4$. When the addition of $NaBH_4$ was completed, stirring was continued and the progress of the reaction was monitored by TLC. After the complete disappearance of the nitrile compound in almost 15 min, the reaction mixture was filtered through a Celite pad. The filtered nickel boride precipitate was washed with ethanol (10 mL). The filtrate was collected, diluted with water (30 mL), and extracted with ethyl acetate (3 × 30 mL). The organic phase was combined, dried over sodium sulfate, filtered and concentrated on a rotary vacuum evaporator to afford a crude product. The crude product was purified by flash chromatography on silica gel, eluting with 4:1 chloroform:methanol to give the corresponding amine compound **AM-1** as a white solid (Caddick et al., 2003). Yield: 65%, ^{1}H NMR (300 MHz, DMSO-d_6) δ (ppm): 8.66 (2H, s, NH_2), 8.12 (1H, d, J = 6Hz, C6′-H), 7.69–7.98 (3H, m, Ar), 7.59–7.70 (1H, br, C4′-H), 7.20–7.40 (2H, m, C3′-H and C5′-H), 4.76 (2H, s, CH_2NH_2).

Procedure for the Preparation of (3-(2-methyl-1H-imidazol-4-yl)phenyl)methanamine (AM-2)

4,5-Dibromo-2-methyl-1H-imidazole (5)

2-Methyl-1*H*-imidazole **4** (10 g, 0.122 mol) was dissolved in 300 mL chloroform and cooled to the temperature between 0 and −5°C using salty ice bath. 48.66 g (15.60 mL, 0.305 mol) bromine was added to the reaction mixture drop-wise via a dropping funnel over 20 min. The reaction mixture was allowed to stir at room temperature. The progression of the reaction was monitored by TLC. The reaction was completed after 20 h. During this time, the reaction product precipitated as an orange solid. In the next step, the reaction mixture was cooled to 0°C in a salty ice bath and 250 mL of NaOH (2 N) was added drop-wise to the reaction mixture in order to quench unreacted bromine. The orange precipitate was filtered off and washed with water, dried *in vacuo* at 40°C for 12 h to give the yellow solid of **5** (Alonso-Alija et al., 2003). Yield: 43%, ^{1}H NMR (300 MHz, DMSO-d_6) δ (ppm): 8.68 (1H, brs, NH), 2.39 (3H, s, CH_3).

4-Bromo-2-methyl-1H-imidazole (6)

4,5-Dibromo-2-methyl-1*H*-imidazole **5** (5.0 g, 20.84 mmol) was suspended with sodium sulfite (80 g, 635 mmol) in 200 mL water

and 100 mL ethanol. The suspension was refluxed at 100°C in an oil bath. The progression of the reaction was monitored by TLC. The reaction was completed after 48 h and the reaction mixture was extracted with ethyl acetate (3 × 100 mL). The organic phases were collected, dried over sodium sulfate, filtered and then evaporated *in vacuo* to yield a white solid as pure product **6** (Alonso-Alija et al., 2003). Yield: 75%, ^{1}H NMR (300 MHz, DMSO-d_6) δ (ppm): 12.00 (1H, brs, NH), 7.05 (1H, d, J = 1.80 Hz, C5-H), 2.21 (3H, s, CH_3).

3-(2-Methyl-1H-imidazol-4-yl)benzonitrile (8)

Nitrile compound **8** was synthesized according to same Suzuki procedure described for the preparation of compound **3** (Tan et al., 2014). Briefly, starting from a solution of palladium (II) acetate (0.42 g, 1.86 mmol, 0.2 eq) and triphenylphosphine (2.44 g, 9.32 mmol, 1 eq) in 6 mL dioxane/water (5:1), a solution of bromo-imidazole derivative **6** (1.5 g, 9.32 mmol, 1 eq) in dioxane (70 mL), a solution of potassium carbonate (3.87 g, 28 mmol, 3 eq) in water (40 mL) and a solution of (3-cyanophenyl) boronic acid **7** (1.51 g, 10.252 mmol, 1.1 eq) in 10 mL dioxane/water (5:1) were used in Sequence. Crude product was purified by flash chromatography on silica gel using 3:1 petroleum ether: ethyl acetate as the eluent to yield **8** as a white solid (Tan et al., 2014). Yield: 70%, ^{1}H NMR (300 MHz, DMSO-d_6) δ (ppm): 11.95 (1H, brs, NH), 8.10 (1H, s, C2′-H), 8.00 (1H, d, J = 9.40 Hz, C6′-H), 7.45–7.68 (3H, m, C5-H, C4′-H, C5′-H), 2.20 (3H, s, CH_3). ESI-MS (+) m/z (%): 183.9 $[M+H]^+$ (100).

(3-(2-Methyl-1H-imidazol-4-yl)phenyl)methanamine (AM-2)

Amine compound **AM-2** was synthesized according to the same procedure described for reduction of nitrile **3** to the corresponding primary amine **AM-1** (Caddick et al., 2003). In this way, starting from nitrile compound **8** (0.5 g, 2.73 mmol, 1 eq), 0.353 gr (2.73 mmol, 1 eq) of anhydrous nickel (II) chloride, and 0.31 g (8.19 mmol, 3 eq) sodium borohydride were used. The purification was performed via silica flash column chromatography using 4:1 chloroform: methanol as the eluent to yield the corresponding amine product **AM-2** as a white solid (Caddick et al., 2003). Yield: 61%, ^{1}H NMR (300 MHz, DMSO-d_6) δ (ppm): 11.72 (1H, brs, NH), 7.70–8.50 (5H, m, C5-H, C2′-H, C4′-H, C5′-H and C6′-H), 3.70 (2H, s, CH_2NH_2), 2.25 (3H, s, CH_3).

Procedure for the Preparation of (3-(1H-pyrrol-1-yl)phenyl)methanamine (AM-3)

3-(1H-Pyrrol-1-yl)benzonitrile (11)

Phenylpyrrole **11** was synthesized through a Clauson–Kaas reaction with 3-aminobenzonitrile **9** (Chatzopoulou et al., 2013). To a solution of 3-aminobenzonitrile **9** (1.00 g, 8.46 mmol, 1 eq) in 15.0 mL of 1,4-dioxane, 2,5-dimethoxytetrahydrofuran **10** (1.23 g, 9.306 mmol, 1.1 eq) dissolved in 7.0 mL of 1,4-dioxane were added. The reaction mixture was refluxed for 5 min, and then 3 mL of hydrochloric acid 5 N was added drop-wise. The reaction mixture was refluxed for 25 min until the reaction was completed. After cooling to room temperature, water was added to the reaction mixture. The reaction mixture was extracted with dichloromethane (3 × 50 mL). The organic layer was dried over anhydrous sodium sulfate, filtered, and concentrated *in vacuo*. Flash chromatography on silica gel, eluting with 4:1 petroleum ether: ethyl acetate, afforded compound **11** as an amorphous white solid (Chatzopoulou et al., 2013). Yield: 82%, ^{1}H NMR (300 MHz, $CDCl_3$) δ (ppm): 7.48–7.52 (4H, m, Ar), 7.08 (2H, t, J = 3.00 Hz, C2′-H, C5′-H), 6.79 (2H, t, J = 3.00 Hz, C3′-H, C4′-H).

(3-(1H-Pyrrol-1-yl)phenyl)methanamine (AM-3)

Amine compound **AM-3** was synthesized according to the same procedure described for the reduction of nitrile **3** to the corresponding primary amine **AM-1** (Caddick et al., 2003). In this way, starting from nitrile **11** (0.5 g, 2.97 mmol, 1 eq), 0.385 g (2.97 mmol, 1 eq) of anhydrous nickel (II) chloride, and 0.337 g (8.91 mmol, 3 eq) sodium borohydride were used. The purification was performed via silica column chromatography using 4:1 chloroform: methanol as the eluent to yield the corresponding amine **AM-3** as a white solid (Caddick et al., 2003). Yield: 75%, ^{1}H NMR (300 MHz, DMSO-d_6) δ (ppm): 7.52 (1H, s, C_2-H), 7.27–7.42 (4H, m, 3H: C3-H, C4-H and C5-H; 1H: C2′-H), 7.18 (1H, d, J = 4.40 Hz, C5′-H), 6.24 (2H, s, C3′-H, C4′-H), 3.77 (2H, s, CH_2NH_2).

General Procedure for the Synthesis of 3-(benzyloxy)-6-methyl-N-(Substituted benzyl)-4-oxo-4H-pyran-2-carboxamide Derivatives, BPCAR

To a vigorously stirred suspension of intermediate **BPCA** (100 mg, 0.38 mmol, 1 eq) in dry dichloromethane (8 mL), 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDCI) (72.5 mg, 0.38 mmol, 1 eq) was added and the mixture was stirred for 30 min under N_2 to provide a clear yellow solution. Then, *N*-hydroxysuccinimide (NHS) (43.7 mg, 0.38 mmol, 1 eq) was added to the stirring solution and the mixture was allowed to stir for 3 h under N_2 to produce the activated ester **12** as reported in the Chemistry details in the Result and Discussion section. After this step and complete consumption of the starting material **BPCA**, desired prepared amine fragment (1.5 eq) was added to the reaction mixture. The reaction was stirred at room temperature for 3 days under N_2. During this time, the progression of the reaction was monitored by TLC. Then, the reaction mixture was poured into a separatory funnel and the dichloromethane layer was washed with water (2 × 20 mL). The organic phase was collected, dried over anhydrous sodium sulfate, filtered, and concentrated *in vacuo* to yield a crude solid. The obtained solid was purified via silica flash column chromatography to give the desired carboxamide product as a pure substance (Sheehan et al., 1965; Sirous et al., 2019).

3-(Benzyloxy)-N-(3-fluoro-5-(pyridin-2-yl)benzyl)-6-methyl-4-oxo-4H-pyran-2-carboxamide (BPCAR-28)

Carboxamide derivative **BPCAR-28** was prepared according to the general procedure, using amine compound **AM-1** (115 mg, 0.57 mmol, 1.5 eq). The purification of the crude product using silica flash column chromatography, eluting with 1:1 petroleum ether: ethyl acetate solution, afforded the carboxamide product

BPCAR-28 as a white solid. Yield: 44%, ^{1}H NMR (300 MHz, $CDCl_3$) δ (ppm): 8.74 (1H, d, J = 5.5 Hz, C6′-H), 8.18(1H, brs, $NHCH_2$), 7.88 (1H, d, J = 5.5 Hz, Ar: C4-H), 7.76 (2H, s, Ar: C2-H and C6-H), 7.18–7.40 (6H, m, 5H: $OCH_2C_6H_5$; 1H: C4′-H), 7.02–7.16 (2H, m, C3′-H and C5′-H), 6.24 (1H, s, C5-H), 5.36 (2H, s, OCH_2-C_6H_5), 4.44 (2H, d, J = 8.00 Hz, $NHCH_2$-C_6H_4), 2.38 (3H, s, 6-CH_3).

3-(Benzyloxy)-6-methyl-N-(3-(2-methyl-1H-imidazol-4-yl)benzyl)-4-oxo-4H-pyran-2-carboxamide (BPCAR-89)

Carboxamide derivative **BPCAR-89** was prepared according to the general procedure, using amine compound **AM-2** (0.107 g, 0.57 mmol, 1.5 eq). The purification of the crude product using silica flash column chromatography, eluting with 1:1 petroleum ether: ethyl acetate solution, afforded the carboxamide product **BPCAR-89** as a white solid. Yield: 38%, ^{1}H NMR (300 MHz, Acetone-d_6) δ (ppm): 8.41 (1H, brs, $NHCH_2$), 7.82 (1H, s, C2′-H), 7.70 (1H, d, J = 12.00 Hz, C4′-H),7.23–7.40 (7H, m, 5H:OCH_2-C_6H_5; 2H: C5′-H and C6′-H), 7.18 (1H, d, J = 12.00 Hz, C5″-H), 6.28 (1H, s, C5-H), 5.30 (2H, s, OCH_2-C_6H_5), 4.50 (2H, d, J = 6.00 Hz, $NHCH_2$-C_6H_4), 2.40 (3H, s, CH_3), 2.24 (3H, s, CH_3).

N-(3-(1H-Pyrrol-1-yl)benzyl)-3-(benzyloxy)-6-methyl-4-oxo-4H-pyran-2-carboxamide (BPCAR-142)

Carboxamide derivative **BPCAR-142** was prepared according to the general procedure, using amine compound **AM-3** (98.2 mg, 0.57 mmol, 1.5 eq). The purification of the crude product using silica flash column chromatography, eluting with 1:1 petroleum ether: ethyl acetate solution, afforded the carboxamide product **BPCAR-142** as a white solid. Yield: 48%, ^{1}H NMR (300 MHz, $CDCl_3$) δ (ppm): 8.12 (1H, t, J = 8.00 Hz, $NHCH_2$), 7.16–7.38 (8H, m, 5H: $OCH_2C_6H_5$; 3H: C2′-H, C4′-H and C5′-H), 7.01–7.08 (3H, m, 2H: C2″-H, C5″-H; 1H: C6′-H), 6.36 (2H, t, J = 3.00 Hz, C3″-H, C4″-H), 6.25 (1H, S, C5-H), 5.32 (2H, s, $OCH_2C_6H_5$), 4.46 (2H, d, J = 8.30 Hz, $NHCH_2C_6H_4$), 2.35 (3H, s, 6-CH_3).

General Procedure for the Synthesis of 3-hydroxy-6-methyl-N-(substituted benzyl)-4-oxo-4H-pyran-2-carboxamide derivatives, HPCAR

40 mg of each of the desired BPCAR derivatives was dissolved in dry dichloromethane (3 mL) and flushed with nitrogen. Then, the reaction mixture was cooled to the temperature between 0 and −5°C in salty ice bath and the 1 M solution of boron tribromide in dichloromethane (3 eq) was slowly added dropwise via a syringe. The reaction mixture was allowed to stir at room temperature and the reaction progress was monitored by TLC. The reaction was completed after almost 3 h. The excess BBr_3 was eliminated at the end of the reaction by the addition of cold methanol (5 mL) to the reaction mixture at 0°C and left to stir for half an hour. The mixture was concentrated to dryness in vacuum and the residue was dissolved several times in methanol and evaporated. This residue was purified by flash column chromatography to afford the final pure product (Ma and Hider, 2015; Sirous et al., 2019).

N-(3-Fluoro-5-(pyridin-2-yl)benzyl)-3-hydroxy-6-methyl-4-oxo-4H-pyran-2-carboxamide (HPCAR-28)

Compound **HPCAR-28** was prepared according to the general debenzylation procedure, starting from compound **BPCAR-28** (40 mg, 0.09 mmol, 1 eq) and 1 M solution of boron tribromide in CH_2Cl_2 (46 μL, 0.27 mmol, 3 eq). Purification using flash column chromatography (eluent: dichloromethane: methanol; 80:20 v/v) afforded a white solid as final product **HPCAR-28**. Yield: 71%, ^{1}H NMR (300 MHz, DMSO-d_6) δ (ppm): 8.60 (1H, d, J = 5.5 Hz, C6′-H), 8.12 (1H, brs, $NHCH_2$), 7.74 (1H, d, J = 5.5 Hz, Ar: C4-H), 7.62 (2H, s, Ar: C2-H and C6-H), 7.12–7.34 (1H, br, C4′-H), 7.00–7.13 (2H, m, C3′-H and C5′-H), 6.22 (1H, s, C5-H), 4.34 (2H, d, J = 8.00 Hz, $NHCH_2$-C_6H_4), 2.34 (3H, s, 6-CH_3). ^{13}C NMR (DMSO-d6) δ (ppm): 173.55 (4-C=O), 164.84 (CONH), 162.72 (C-3), 162.48 (Ar: $C_{3'}$-F, d, $^1J_{C-F}$: 220.4 Hz), 152.15 (Ar), 147.39 (Ar), 144.05 (C-2), 143.86 (Ar), 136.14 (C-6), 132.32 (Ar), 132.12 (Ar), 124.00 (Ar), 123.09 (Ar), 122.28 (Ar), 116.60 (Ar), 114.11 (Ar), 108.63 (C-5), 42.04 ($NHCH_2$), 19.31 (6-CH_3). ESI-MS (+) m/z (%): 355.3 $[M+H]^+$ (100).

3-Hydroxy-6-methyl-N-(3-(2-methyl-1H-imidazol-4-yl)benzyl)-4-oxo-4H-pyran-2 carboxamide (HPCAR-89)

Compound **HPCAR-89** was prepared according to the general procedure described above, starting from compound **BPCAR-89** (40 mg, 0.09 mmol, 1 eq) and 1 M solution of boron tribromide in CH_2Cl_2 (46 μL, 0.27 mmol, 3 eq). Purification using flash column chromatography (eluent: dichloromethane: methanol; 80:20 v/v) afforded a white solid as final product **HPCAR-89**. Yield: 82%, ^{1}H NMR (300 MHz, DMSO-d_6) δ (ppm): 11.80 (1H, brs, NH of imidazole ring), 10.63 (1H, brs, $NHCH_2$),7.60 (1H, s, C2′-H), 7.58 (1H, d, J = 12.00 Hz, C4′-H), 7.38 (1H, s, C5"-H), 7.20 (1H, t, J = 12.00 Hz, C5′-H), 7.05 (1H, d, J = 12.00 Hz, C6′-H), 6.20 (1H, s, C5-H), 4.40 (2H, d, J = 6.00 Hz, $NHCH_2$-C_6H_4), 2.24 (6H, s, CH_3-a and CH_3-b). ^{13}C NMR (DMSO-d6) δ (ppm): 173.53 (4-C=O), 164.78 (CONH), 162.72 (C-3), 156.60 ($C_3H_4N_2$: C2), 151.02 ($C_3H_4N_2$: C5), 147.44 (C-2), 142.54 (Ar), 136.14 (C-6), 134.64 (Ar), 129.47 (Ar), 128.30 (Ar), 125.52 (Ar), 123.14 ($C_3H_4N_2$: C4), 117.21 (Ar), 112.65 (C-5), 41.53 ($NHCH_2$), 19.30 (6-CH_3), 16.60 (CH_3-$C_3H_4N_2$). ESI-MS (+) m/z (%): 340.1 $[M+H]^+$ (100).

N-(3-(1H-Pyrrol-1-yl)benzyl)-3-hydroxy-6-methyl-4-oxo-4H-pyran-2-carboxamide (HPCAR-142)

Compound **HPCAR-142** was prepared according to the general procedure, starting from compound **BPCAR-142** (40 mg, 0.096 mmol, 1 eq) and 1 M solution of boron tribromide in CH_2Cl_2 (49 μL, 0.288 mmol, 3 eq). Purification using flash column chromatography (eluent: dichloromethane: methanol; 80:20 v/v) afforded a white solid as final product **HPCAR-142**. Yield: 86%, ^{1}H NMR (300 MHz, DMSO-d_6) δ (ppm): 10.30 (1H, brs,

$NHCH_2$), 7.2–7.56 (4H, m, C2′-H, C4′-H, C5′-H, C6′-H), 7.0–7.20 (2H, m, C2″-H, C5″-H), 6.23 (2H, d, $J = 7.50$ Hz, C3″-H, C4″-H), 6.18 (1H, s, C5-H), 4.50 (2H, d, $J = 7.00$ Hz, $NHCH_2$), 2.30 (3H, s, 6-CH_3). ^{13}C NMR (DMSO-d6) δ (ppm): 173.55 (4-C=O), 164.79 (CONH), 162.72 (C-3), 147.51 (C-2), 142.52 (Ar), 136.19 (C-6), 135.31 (Ar), 128.90 (Ar), 124.37 (Ar), 120.05 (C_4H_4N: C2, C5), 119.30 (Ar), 117.43 (Ar), 112.67 (C-5), 108.77 (C_4H_4N: C3, C4), 41.96 ($NHCH_2$), 20.65 (6-CH_3). ESI-MS (+) m/z (%): 324.9 $[M+H]^+$ (30), 346.7 $[M+Na]^+$ (100).

Biological Evaluation

Integrase Assays

The enzymatic integration reactions were carried out as previously described with minor modifications (Debyser et al., 2001; Christ et al., 2011). To determine the susceptibility of the HIV-1 IN enzyme to different compounds, an enzyme-linked immunosorbent assay (ELISA) adapted from Hwang et al. was used (Hwang et al., 2000). The overall integration assay uses an oligonucleotide substrate for which one oligonucleotide (5′-ACTGCTAGAGATTTTCCACACTGACTAAAAGGGTC-3′) is labeled with biotin at the 3′ end and the other oligonucleotide (5′-GACCCTTTTAGTCAGTGTGGAAAATCTCTAGCAGT-3′) is labeled with digoxigenin at the 5′ end. For the strand transfer assay, a pre-cleaved oligonucleotide substrate (the second oligonucleotide lacks GT [underlined] at the 3′ end) was used. The IN enzyme was diluted in 750 mM NaCl, 10 mM Tris (pH: 7.6), 10% glycerol, and 1 mM β-mercaptoethanol. To perform the reaction, 4 μL of diluted IN (corresponding to a concentration of 1.6 μM) and 4 μL of annealed oligonucleotides (7 nM) were added in a final reaction volume of 40 μL containing 10 mM $MgCl_2$, 5 mM dithiothreitol, 20 mM HEPES (pH 7.5), 5% polyethylene glycol, and 15% dimethyl sulfoxide. As such, the final concentration of IN in this assay was 160 nM. The reaction was carried out for 1 h at 37°C. Reaction products were denatured with 30 mM NaOH and detected by ELISA on avidin-coated plates. For determining the effect of compounds on the 3′-processing activity a classical cleavage assay with detection of products by denaturing gel electrophoresis was performed as described previously (Debyser et al., 2001; Christ et al., 2011). Briefly, 0.2 pmol of the radioactive labeled oligonucleotide substrate (INT1, ^{32}P-5′ TGTGGAAAATCTCTAGCAGT3′; INT2, 5′ACTGCTAGAGATTTTCCACA 3′) and 10 nmol IN in a final volume of 10 μL was incubated for 1 h at 37°C. The final reaction mixture contained 20 mM HEPES (pH 7.5), 5 mM dithiothreitol (DTT), 10 mM $MgCl_2$, 0.5% (v/v) polyethylene glycol 8000, 15% DMSO. IN was diluted previously in 750 mM NaCl, 10 mM Tris (pH 7.6), 10% glycerol and 1 mM β-mercaptoethanol. The reactions were stopped by the addition of formamide loading buffer (95% formamide, 0.1% xylene cyanol, 0.1% xylene cyanol, 0.1% bromophenol blue, and 0.1% sodium dodecyl sulfate). Samples were loaded on a 15% denaturing polyacrylamide/ureum gel. The extent of 3′-processing or DNA strand transfer was based on measuring the respective amounts of −2 bands or strand transfer products relative to the intensity of the total radioactivity present in the lane. These data were determined using the OptiQuant Acquisition and Analysis software (Perkin Elmer Corporate, Fremont, CA).

In vitro Anti-HIV and Drug Susceptibility Assays

The inhibitory effect of antiviral drugs on the HIV-induced cytopathic effect (CPE) in human lymphocyte MT-4 cell culture was determined by the MT-4/MTT-assay (Pauwels et al., 1988). This assay is based on the reduction of the yellow colored 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) by mitochondrial dehydrogenase of metabolically active cells to a blue formazan derivative, which can be measured spectrophotometrically. The 50% cell culture infective dose of the HIV strains was determined by titration of the virus stock using MT-4 cells. For the drug susceptibility assays, MT-4 cells were infected with 100–300 50% cell culture infective doses of the HIV strains in the presence of 5-fold serial dilutions of the antiviral drugs. The concentration of the compound achieving 50% protection against the CPE of HIV, which is defined as the 50% effective concentration (EC_{50}), was determined. In parallel, the concentration of the compound destroying 50% of the MT-4 cells, which is defined as the 50% cytotoxic concentration (CC_{50}), was determined as well.

RESULTS AND DISCUSSION

The main purpose of the present study is to identify novel chemical entities derived from HPCARs scaffold as new and useful hit compounds as HIV-1 INIs. Accordingly, an integrated computational protocol based on combinatorial library design protocol, physico-chemical properties prediction, molecular docking calculations, and MD simulation was developed in a stepwise filtering approach (**Figure 3**). The identified hit compounds were synthesized and submitted to biological evaluation in order to validate the proposed *in silico* strategy.

Generation of Combinatorial Hits Using CombiGlide

As the first step of the developed *in silico* protocol, HPCAR derivatives with $n = 1$ or 2 (**Figures 2**, **3**) were submitted to CombiGlide software as a combinatorial docking tool. In each case, combinatorial virtual screening was applied in order to replace the aromatic groups of the original core, applying side-chain hopping method. The prepared sets of amine fragments, available in the library of reagents, were used to replace the original substituents at each defined attachment point (**Figure 3**). Variation in aromatic group resulted in the generation of a combinatorial library of more than 37,000 hit compounds for each studied core-containing molecule. The compounds from each new combinatorial library were sorted by GlideScore values. Only derivatives with score values lower than −6.0 kcal/mol were considered. The selected molecules were further analyzed by visual inspection to find compounds with an appropriate binding mode according to the key interactions found for HIV-1 INIs. From this first filter, 1,803 combinatorial compounds were chosen for the next step.

Molecular Properties Prediction

One of the major goals in drug discovery is the identification of innovative small molecular scaffolds exhibiting high efficacy and selectivity against the desired target along with a satisfactory ADME-T profile. Thus, the second filter in the screening workflow consisted in the prediction of the ADME-T properties and drug-like behavior of the above-mentioned 1,803 compounds using QikProp software. This step was performed to select molecules possessing satisfactory predicted membrane permeability (QPPCaco-2 and QPPMDCK models > 100), appropriate lipophilicity (QPlogP) including capability to cross the blood brain barrier and drug-likeness properties in accordance with Lipinski's rule of five. The potential interaction with *h*ERG K^+ channel (QPlog-HERG) was another key parameter considered in this step of filtering. 146 out of 1,803 compounds were predicted to have pharmacokinetic properties in the appropriate range. Moreover, the resulting compounds were filtered for behaving as PAINS using FAF-Drugs4 tool. Among 146 compounds, only two molecules contain sub-structural features that marked them as "frequent hitters" in high throughput screens. Finally, 144 candidates passed this step of screening and were chosen for the next step. A list of these top candidates with improved ADMET properties was provided in the **Table S1**.

Quantum Polarized Ligand Docking Simulation

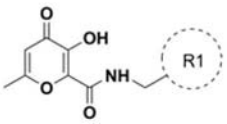
The resulting 144 potential hit molecules were further computationally analyzed using QPLD calculations for guaranteeing a better prediction of their binding mode into HIV-1 IN active site. This docking protocol could provide a more accurate treatment of electronic interactions especially within metalloproteins active site, leading to the improvement of the accuracy of the docking results (Cho et al., 2005; Illingworth et al., 2008; Paolino et al., 2018). In this step, the potential inhibitors were selected based on their lower XP GlideScore and on the ability to engage in critical interactions in the HIV-1 IN active site. At the end, a total of 76 hit candidates with the favorable XP GlideScore values were identified (**Table S2**). As reported in **Table S2**, the 76 selected hit molecules showed XP GlideScore values < −6 kcal/mol (the values of the cut-off filters for the *in silico* studies were chosen taking into consideration the values found for the reference compounds **RLT**, **EVG**, and **DTG**). The detailed analysis of QPLD results indicated that these compounds adopt a reasonable interfacial binding mode similar to that found for the approved HIV-1 INIs, namely **RLT**, **EVG**, and **DTG** (Rostami et al., 2015; Sirous et al., 2015, 2019). Consistent with docking models of HP derivatives previously reported (Sirous et al., 2019), the same interaction pattern was found for the best docked pose of all the selected hit molecules within HIV-1 IN active site. In this context, combinatorial hits perfectly occupied the DNA/IN interface with donor oxygen triad of MBG interacting with both Mg^{2+} ions through a bis-bidentate mode of chelation. This orientation enables the aromatic side chain of the molecule to sit in a hydrophobic pocket close to the active site generated by the displacement of the terminal adenosine on the 3′-end of the viral DNA. As a result, the terminal aromatic moiety of ligands participates in π-π stacking interactions with the viral DNA nucleosides, DC_{16} and DG_4, and favorable hydrophobic contacts with the amino acids residues of the catalytic loop, Pro_{145}, Gln_{146}, and Gly_{149}. Particularly, Pro_{145} and Gln_{146} are directly involved in separation of the viral DNA strands upon the ST reaction. This can reduce the catalytic loop mobility and thus physically hamper the binding of the host DNA (Dirac and Kjems, 2001; Dolan et al., 2009). In some cases, further stabilization of the ligand in the active site was mediated by H-bonds with Asn_{117}, Pro_{145}, Gln_{146}, and Glu_{152} as well as nucleoside residues DG_4, DC_{16}, and DA_{17}. For example, **HPCAR-40** was involved in hydrogen bond interactions with Asn_{117} and Glu_{152} and **HPCAR-144** formed hydrogen bonds with Gln_{146} and Glu_{152}. Furthermore, in most of the docking models, the position of 4-pyran core of ligands was suitably located to establish strong hydrophobic interactions such as a π-π stacking with 3′-deoxyadenosine A_{17} (**Table S2**).

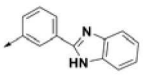
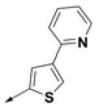
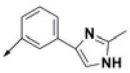
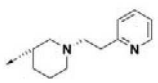
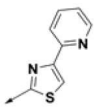
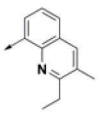
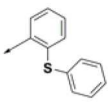
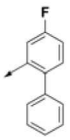
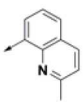
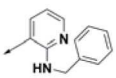
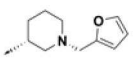
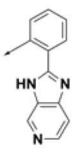
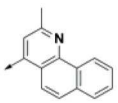
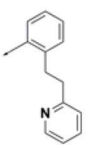
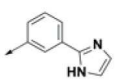
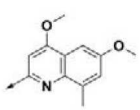
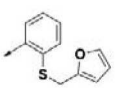
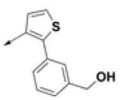
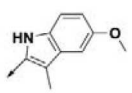
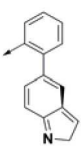
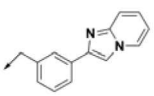
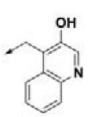
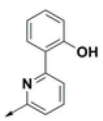
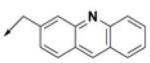
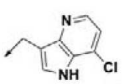
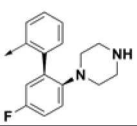
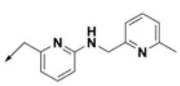
Prioritization of Hit Compounds Based on Relative Ligand Binding Energy

Although it is well established that docking calculations are highly successful in offering reliable ligand poses within the protein binding site, they often fail to rank compounds with respect to their binding affinities. This poor correlation may be due to severe approximations and simplifications employed by scoring functions of various docking tools. The scoring functions like GlideScore do not consider some essential thermodynamics factors in the ligand binding energy calculations such as protein and ligand solvation energy terms (Pearlman and Charifson, 2001; Taylor et al., 2002). Thus in the subsequent step of our computational workflow, relative ligand binding energy calculations using MM-GBSA rescoring method were carried out on the best docked pose of the ligands obtained from the previously described docking simulation. This approach may offer more reliable measuring criteria to prioritize screened HIV-1 INIs hits for chemical synthesis and biological evaluations as HIV-1 INIs (Huang et al., 2006). Rescoring using MM-GBSA leads to minor changes of the ligand conformations within receptor site. These changes result from minimization of the ligand in receptor's environment and consequent stabilization of receptor-ligand complex. The estimated binding energy values < −25 kcal/mol were considered to retrieve final set of combinatorial hits. Final ranking of the ligands in this step of screening workflow resulted in the identification of 40 top hit compounds as novel HIV-1 INIs, possessing relevant binding affinities for HIV-1 IN active site. The structures of these compounds are shown in **Table 2**. The calculated ΔG_{bind} of the final selected hits along with their contributions to total binding energy from various energy components are provided in **Table S3**.

Since the screened hit molecules share the same MBG, the main differences in ligand binding energies values between these inhibitors could be directly attributed to the hydrophobic aromatic moieties characterized by significant chemical diversity, including bicyclic and tricyclic structures. Inspection of energy

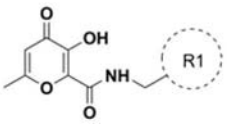
TABLE 2 | Chemical structures of 40 top combinatorial hits identified at the end of computational screening workflow applied in the present study.

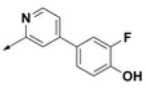
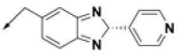
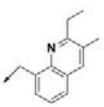
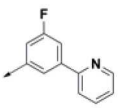
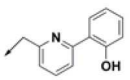
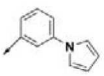
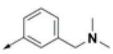
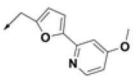
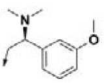
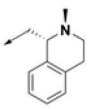


Cmpd	R_1	Cmpd	R_1	Cmpd	R_1
HPCAR-1		**HPCAR-33**		**HPCAR-89**	
HPCAR-2		**HPCAR-35**		**HPCAR-90**	
HPCAR-6		**HPCAR-37**		**HPCAR-91**	
HPCAR-7		**HPCAR-41**		**HPCAR-92**	
HPCAR-8		**HPCAR-44**		**HPCAR-108**	
HPCAR-14		**HPCAR-45**		**HPCAR-111**	
HPCAR-15		**HPCAR-46**		**HPCAR-114**	
HPCAR-22		**HPCAR-52**		**HPCAR-123**	
HPCAR-23		**HPCAR-54**		**HPCAR-126**	
HPCAR-25		**HPCAR-55**		**HPCAR-130**	

(Continued)

TABLE 2 | Continued



Cmpd	R_1	Cmpd	R_1	Cmpd	R_1
HPCAR-26		**HPCAR-56**		**HPCAR-140**	
HPCAR-28		**HPCAR-66**		**HPCAR-142**	
HPCAR-29		**HPCAR-69**		**HPCAR-144**	
HPCAR-30					

*Three hit compounds (**HPCAR-28**, **HPCAR-89** and **HPCAR-142**) were selected from this final set for synthesis and biological assessment.*

terms in **Table S3** revealed that all selected ligands showed high values of van der Waals interaction energy ($\Delta G_{bindVdW}$ values), contributing to the ligand binding energy which emphasizes critical importance of hydrophobic interactions in the stability of the ligand–protein complexes.

Validation of Computational Screening Workflow

In order to validate the developed computational protocol, three compounds (**HPCAR-28**, **HPCAR-89**, and **HPCAR-142** in **Table 2**) from final set of 40 hit candidates were selected and synthesized. The selection was performed considering the favorable computational scores, the binding modes, the structural differences and synthetic accessibility. The best docked poses along with the detailed interaction into the HIV-1 IN active site of these representative compounds are depicted in **Figure 4**.

The compounds were able to establish a bis-bidentate chelation of the Mg^{2+} ions, strong hydrophobic interactions (π-π stackings) with the nucleotide DC_{16} and Pro_{145}. Interestingly, compounds **HPCAR-28** and **HPCAR-89** were able to form H-bonds with DG_4 that can further stabilize the binding mode compared to **HPCAR-142**. Moreover, the fluorine atom of **HPCAR-28** can guarantee additional interactions within the binding site with Pro_{145} and the sidechains of Glu_{152}. This slightly different pattern of interactions is also highlighted by the differences in docking scores and ΔG_{bind} values (**HPCAR-28** GlideScore −7.980 kcal/mol and ΔG_{bind} −34.102 kcal/mol; **HPCAR-89** GlideScore −6.648 kcal/mol and ΔG_{bind} −26.777 kcal/mol; **HPCAR-142** −6.622 kcal/mol and ΔG_{bind} −25.759 kcal/mol; as reported in **Tables S2**, **S3**). Overall, the *in silico* analysis showed that **HPCAR-28** and **HPCAR-89** can better interact with the active site of HIV IN with respect to **HPCAR-142**.

Regarding the investigation of the binding modes of our derivatives, in our previous study (Sirous et al., 2019), we discussed about the mutation of Tyr_{143} that confers resistance to **RLT**. In particular, **RLT** established interactions with Tyr_{143} by its oxadiazole moiety in both binary PFV-IN and modeled HIV-1 IN complexes. Interestingly, our most promising derivatives do not possess a moiety that can establish interactions with this residue (i.e., oxadiazole in **RLT**). Furthermore, in this study we investigated also two additional mutations that could confer resistance to drugs including **RLT** and **DTG**, $Gln_{148}His$ and $Gly_{140}Ser$. The *in silico* analysis reported in **Figure S1** showed the superposition of the binding mode of **RLT**, **DTG**, and **HPCAR-28** into HIV-1 IN active site. Notably, the only binding mode that can be strongly influenced by these residues (Tyr_{143}, Gln_{148}, and Gly_{140}) is the one of **RLT**. **DTG** can marginally interact with the mentioned residues, while **HPCAR-28** is largely distant from the residues that are responsible of the resistance (distance from Me of **HPCAR-28** to Tyr_{143} over 5 Å, to Gln_{148} over 9 Å, to Gly_{140} over 10 Å; measured by the measurement tool available in PyMOL). Remarkably, our HPCAR derivatives (**HPCAR-28** and **HPCAR-89**) can additionally target the nucleotide DC16 and DG4 (**Figure 4**). This analysis is in perfect agreement with the experimental data showing a dramatic decrease of affinity of **RLT** for mutant HIV-1 IN and a lower decrease of affinity of **DTG**. Consequently, it was assumed that the possible mutations

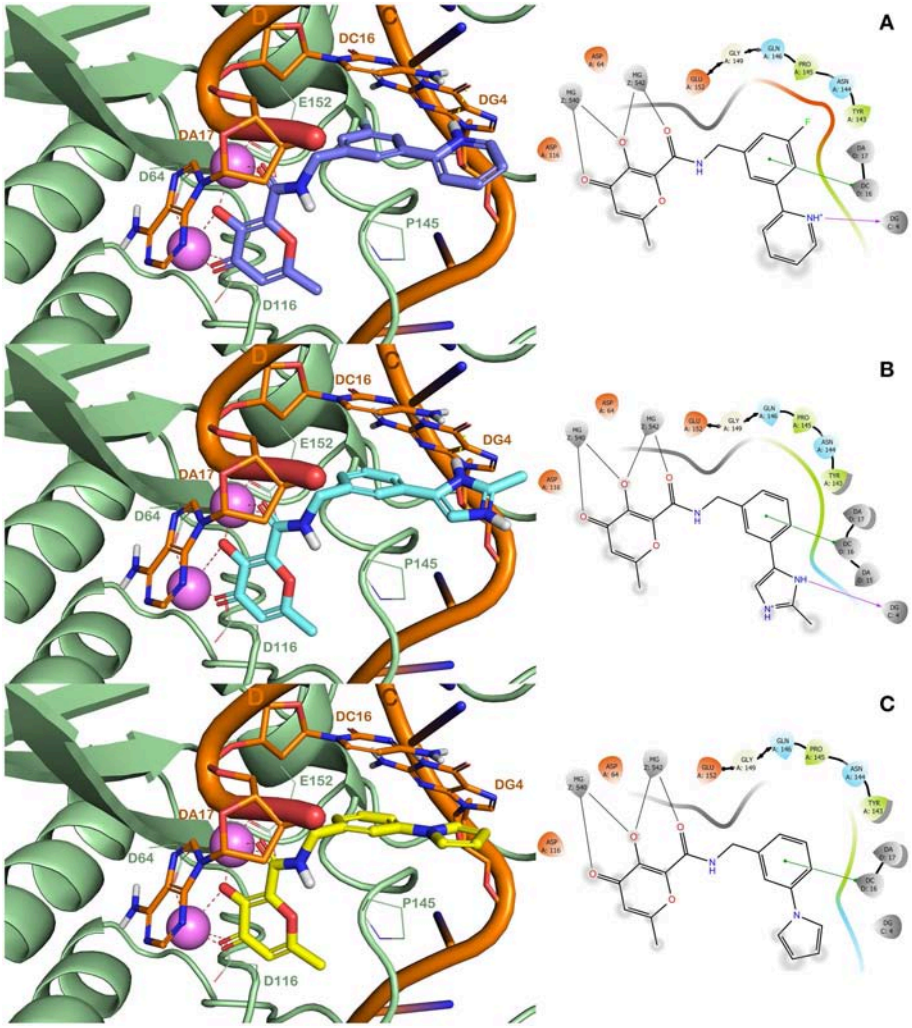


FIGURE 4 | (Left) Binding modes of **HPCAR-28** (purple sticks, **A**), **HPCAR-89** (cyan sticks, **B**), and **HPCAR-142** (yellow sticks, **C**) in the active site of the modeled HIV-1 IN, as found by QPLD method. The solid ribbon model shows the backbone of the modeled HIV-IN, and the key amino acids of the binding site are shown as lines. Two Mg^{2+} ions are displayed as violet spheres, while the viral DNA strands are depicted in orange. Hydrogen and metal coordination bonds are represented by dashed lines. These figures were prepared using PyMOL. (Right) Ligand interaction diagrams for the selected compounds. These figures were prepared using Maestro.

of residues Tyr_{143}, Gln_{148}, and Gly_{140} could not influence the binding of the HPCAR derivatives to IN.

Molecular Dynamics Simulation Studies

In order to better understand the behavior of the representative compounds into HIV-IN enzyme for providing more reliable results about the interactions of **HPCAR-28**, **HPCAR-89**, and **HPCAR-142** with HIV-IN, we performed MD simulations starting from the docked poses reported in **Figure 5** (see Experimental Section for further details).

The three INIs reached an overall stability about after 20 ns. We observed that the pattern of interaction indicated by the docking calculations are generally maintained during the MD, confirming **HPCAR-28** and **HPCAR-89** as more potent potential INIs with respect to the compound **HPCAR-142**. Accordingly, the three compounds maintained the coordination bond with the Mg^{2+} ions during the simulation as well as the hydrophobic interactions with DC_{16} and DA_{17}. We also observed that **HPCAR-28** and **HPCAR-89** were able to establish and maintain further contacts into the binding site with respect to the compound **HPCAR-142**. In fact, the presence of a nitrogen in the R_1 group as in of **HPCAR-28** and **HPCAR-89** allowed to the molecules to stabilize their binding mode by forming further H-bonds with DNA during the simulations. Briefly, the analysis of the computational studies allowed to propose **HPCAR-28** and **HPCAR-89** as more potent INIs with respect to the compound **HPCAR-142**.

In summary, combining different computational techniques for evaluating the affinity of the compounds for HIV IN binding site we provided a more comprehensive *in silico* protocol improving the probability to identify and select compounds with relevant affinity for the selected binding site (Brogi et al.,

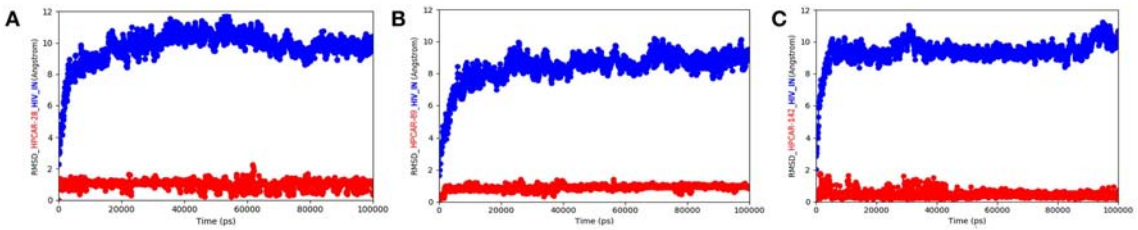


FIGURE 5 | RMSD of the complexes **HPCAR-28**/HIV IN **(A)**, **HPCAR-89**/HIV IN **(B)**, **HPCAR-142**/HIV IN **(C)** (RMSD of the protein in a line formed by blue circles; RMSD of the ligand in a line formed by red circles). RMSD were calculated between the final conformation and the starting conformation through the 100 ns of the MD simulation.

2009). Accordingly, the presented screening workflow allowed to select novel potential HIV-1 INIs based on HPCAR scaffold with improved predicted pharmacological profile. Taking into account also the synthetic feasibility three representative hit candidates were then chosen for the synthesis and then biologically evaluated for validating the applied computational approach.

Chemistry

In this study, 3-(benzyloxy)-6-methyl-4-oxo-4*H*-pyran-2-carboxylic acid, **BPCA** was used as a key intermediate material for the preparation of the three selected hit compounds. Thus, this intermediate was first synthesized starting from commercially available Kojic acid according to a synthetic procedure previously employed in our laboratory (Sirous et al., 2019). On the other hand, three different synthetic routes were developed for the preparation of the amine fragments needed for appending desired hydrophobic backbone to **BPCA**. The methodologies adopted for the synthesis of amine compounds **AM-(1-3)** are outlined in **Figures 6**–**9**, respectively.

As described in **Figure 6**, the synthesis of the amine fragment **AM-1** started using an efficient Suzuki–Miyaura cross-coupling (SMC) (Tan et al., 2014) by reacting 2-bromopyridine **1** with phenyl-boronic acid derivative **2** in the presence of tetrakis (triphenylphosphine) palladium (0) as the catalyst, affording the phenyl-pyridine derivative **3**. Subsequently the nitrile group of resulting compound **3** was reduced to the corresponding primary amine **AM-1** by treatment with sodium borohydride and dry nickel (II) chloride (Khurana and Gogia, 1997; Khurana and Kukreja, 2002; Caddick et al., 2003).

A four-step synthetic procedure was employed for the synthesis of the amine fragment **AM-2** starting from 2-methyl-imidazole **4** (**Figure 7**). The first step of the protocol consisted of the double bromination of the imidazole ring which afforded the dibromo-imidazole derivative **5** in a moderate yield. The selective debromination of the vicinal dibromide **5** employing sodium sulfite as the reducing agent in aqueous ethanol at reflux temperature effectively provided product **6** (yield: 75%) (Khurana and Gogia, 1997; Alonso-Alija et al., 2003). Bromo-imidazole **6** underwent a classical SM coupling reaction with phenyl-boronic acid **7**, providing phenyl-imidazole derivative **8**. Amine fragment **AM-2** was finally obtained by reduction of phenyl-imidazole derivative **8** using sodium borohydride and dry nickel (II) chloride as described for **AM-1** (Caddick et al., 2003).

The amine fragment **AM-3** was prepared from commercially available 3-aminobenzonitrile **9**, following **Figure 8**. A Clauson–Kaas reaction of 3-aminobenzonitrile **9** with dimethoxytetrahydrofuran **10** catalyzed by hydrochloric acid afforded the corresponding 1-phenylpyrrole derivative **11** in a good yield (82%) (Chatzopoulou et al., 2013). The nitrile functionality of the resulting product was then reduced in the presence of sodium borohydride and dry nickel (II) chloride to furnish the desired amine compound **AM-3** (Caddick et al., 2003).

Finally, the representative hit compounds were prepared by introduction of amine backbones to the benzyl-protected pyranone **BPCA** under standard amide coupling conditions (Sheehan et al., 1965; Sirous et al., 2019). The general procedure employed for the synthesis of the final compounds is summarized in **Figure 9**. The activation of carboxylic acid group of **BPCA** as the corresponding derivative **12** using EDCI and NHS as coupling reagents followed by treatment with the desired amines led to 2-amido substituted pyranone analogs **BPCAR**. The removal of the benzyl protecting group of the obtained amide derivatives was then accomplished by reaction with boron tribromide in dichloromethane at room temperature to obtain three final target products, **HPCAR-28, HPCAR-89**, and **HPCAR-142**, in 71–86% yield (Kosak et al., 2015; Sirous et al., 2019).

Biological Activity Evaluation

For validating the computational protocol herein presented, three representative compounds (**HPCAR-28**, **HPCAR-89**, and **HPCAR-142**) were synthesized and biologically assessed for HIV-1 IN catalytic inhibitory activity based on an *in vitro* enzymatic assay. Given that most HIV-1 INIs such as RLT target the ST step of the integration reaction, the inhibition of the ST activity of HIV-1 IN was examined in these assays in addition to overall HIV-1 IN inhibition. Moreover, assessment of the anti-HIV-1 potential in MT-4 cells was performed in a multiple round cell-based antiviral assay. Cytotoxicity of the selected compounds for the target host cell was also evaluated and their therapeutic indices were calculated. In these experiments, **RLT** was employed as a reference HIV-1 INI. The results for biological activities of these hits were summarized in **Table 3**.

As reported in **Table 3**, biological evaluation confirmed the favorable anti-HIV profile for the three newly synthesized

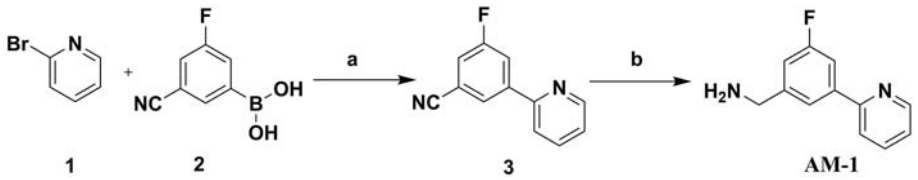


FIGURE 6 | Procedure applied for the synthesis of the amine fragment **AM-1**. Reagents and conditions: (a) $Pd(OAc)_2$, PPh_3, K_2CO_3, dioxane/water, 24 h reflux; (b) $NaBH_4$, $NiCl_2$ (dry), dry ethanol, 25°C.

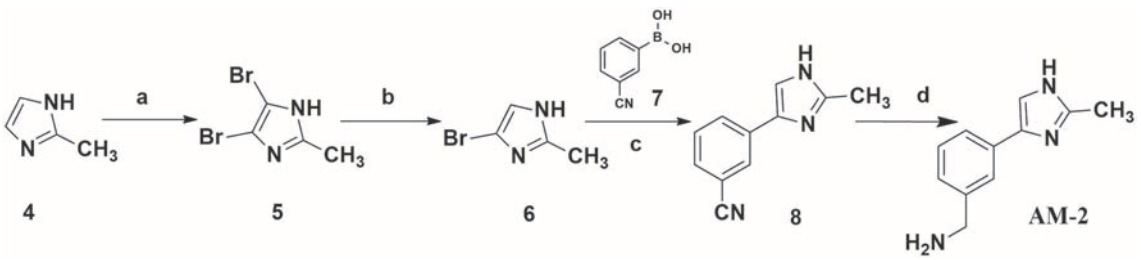


FIGURE 7 | Procedure applied for the synthesis of the amine fragment **AM-2**. Reagents and conditions: (a) Br_2, chloroform, 0–25°C; (b) Na_2SO_3, ethanol/water, 48 h reflux; (c) $Pd(OAc)_2$, PPh_3, K_2CO_3, dioxane/water, 24 h reflux (d) $NaBH_4$, $NiCl_2$ (dry), dry ethanol, 25°C.

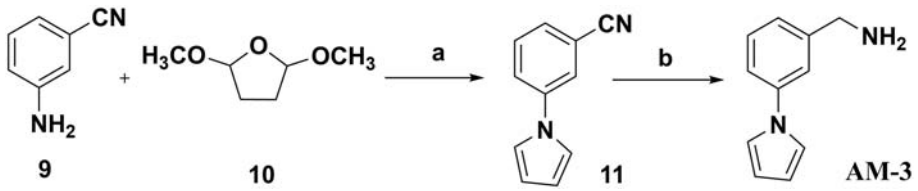


FIGURE 8 | Procedure applied for the synthesis of the amine fragment **AM-3**. Reagents and conditions: (a) 5 N HCl, dioxane, 30 min reflux; (b) $NaBH_4$, $NiCl_2$ (dry), dry ethanol, 25°C.

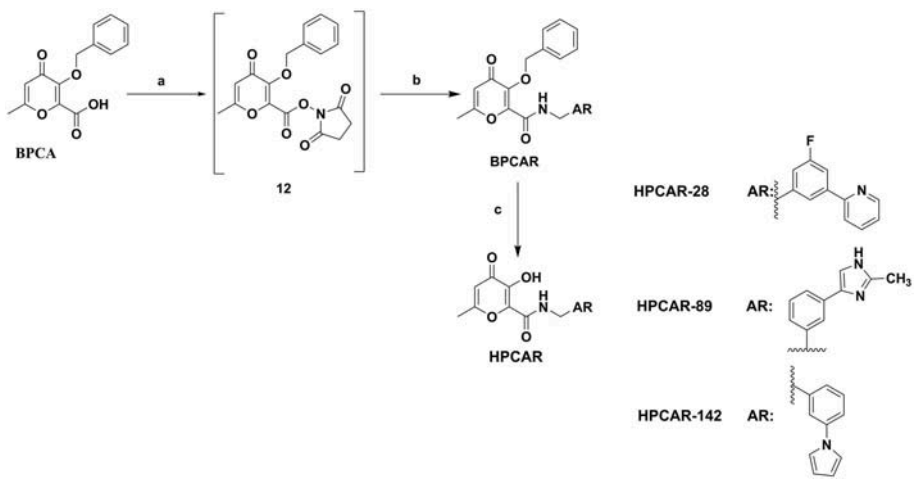


FIGURE 9 | General procedure applied for the final synthesis of three representative hit compounds based on HPCAR scaffold. Reagents and conditions: (a) NHS, EDCI, Dichloromethane (dry), 25°C, 3 h; (b) AR-CH_2NH_2 (amine compounds used for **HPCAR-28, HPCAR-89**, and **HPCAR-142** are **AM-1**, **AM-2**, and **AM-3**, respectively), 25°C, 3 days; (c) BBr_3, Dichloromethane (dry), 0–25°C.

TABLE 3 | Results for anti-HIV activity, inhibitory potential of HIV-1 IN and cytotoxicity of the synthesized compounds and **RLT**.

Entry	Overall HIV-1 IN activity inhibition IC_{50} (μM)[a]	HIV-1 IN strand transfer activity inhibition IC_{50} (μM)[b]	HIV-1 activity inhibition EC_{50} (μM)[c]	Cellular toxicity CC_{50} (μM)[d]	TI[e]
HPCAR-28	0.065 ± 0.01	3.37 ± 0.38	0.23 ± 0.18	>250	>1,087
HPCAR-89	0.27 ± 0.1	4.0 ± 0.34	2.34 ± 0.46	>125	>53.4
HPCAR-142	1.97 ± 1.06	17.49 ± 0.13	4.21 ± 0.52	>85.4	>20.3
HPb	0.37 ± 0.17	1.405 ± 0.01	1.72 ± 0.99	>250	>145
HPd	0.7 ± 0.31	1.54 ± 0.2	1.95 ± 0.93	195.5 ± 4.4	100.3
RLT	0.01	0.04	0.013 ± 0.008	>18.329	>1410

[a,b]*Concentration required to inhibit 50% of the in vitro overall and strand transfer integrase activities, respectively.*
[c]*Effective concentration in which 50% inhibition is observed.*
[d]*Cytotoxic concentration in which 50% of the cells are killed.*
[e]*Therapeutic index: defined by CC_{50}/EC_{50} ratio.*

entities, which demonstrated the ability to inhibit the catalytic activities of HIV-1 IN in the low micromolar range and highlighting the validity of the developed computational protocol for optimizing the previous developed compound. The positive influence of hydrophobic moiety modification of HPCAR derivatives on the inhibitory activity was particularly evident with phenyl-pyridine substituted derivative **HPCAR-28**. This compound emerged as the most potent inhibitor among three tested compounds with low nanomolar activity against HIV-1 IN (IC_{50} = 65 nM) as highlighted by the computational studies, and a 6-fold improvement in anti-IN potency compared to **HPb** (**Figure 2** and **Table 3**; IC_{50} = 0.37 μM) (Sirous et al., 2019). Compounds **HPCAR-89** and **HPCAR-142** also showed promising anti-IN activities in the low micromolar range. In this regard, the incorporation of phenyl-imidazole moiety at the carboxamide sidechain (compound **HPCAR-89**) proved to be also advantageous since it led to a slight enhancement in HIV-1 IN inhibitory activity (IC_{50} = 0.27 μM) compared to the respective *para*-fluorobenzyl amide analog. Although hit compound **HPCAR-142** bearing phenyl-pyrrole fragment was less active than two other hits in HIV-1 IN inhibition (IC_{50} = 1.97 μM), it is still a promising candidate compound for further structural optimization. It was also found that synthesized compounds have the capacity to inhibit the ST step of the HIV-1 IN in the low micromolar range. These observations were in agreement with the *in silico* analysis done, highlighting the fundamental validity of the developed computational protocol for optimizing the previous developed compound. Examination of anti-HIV-1 activity of representative compounds based on cell-based assay indicated that benzyl group replacement with desired hydrophobic moieties was well tolerated for the inhibition of HIV-1 replication as well. According to these results, antiviral activities are reasonably correlated with HIV-1 IN inhibition potencies thus confirming the mechanism of action of these anti-HIV-1 agents. Accordingly, **HPCAR-28** showed the best anti-IN activity and it is the most active hit against HIV-1 infected cells with an EC_{50} value of 0.23 μM.

The cytotoxicity assay also showed that the tested compounds are safe and possess anti-HIV activity at non-cytotoxic concentrations (CC_{50} values ranging from 85.4 to >250 μM), thus resulting in favorable therapeutic indices for the investigated compounds. In particular, the most promising hit compound (**HPCAR-28**) revealed an appreciable therapeutic index (TI > 1,087) comparable to that found for RLT (TI > 1,410). On the contrary, the limited toxicity (>85.4 μM) showed by **HPCAR-142** is potentially ascribable to the presence of the pyrrole moiety that often presents some toxicity.

Overall, these results confirmed that three representative hit compounds are able to achieve the desired level of biological activities in terms of reduced toxicity and optimum inhibitory activities against HIV-1 IN and HIV-1 in cell culture. Furthermore, it was clearly verified that modification of the hydrophobic aromatic moiety within the HPCAR derivatives can lead to differences in HIV-1 IN inhibitory profiles. Moreover, this research clearly confirms the key role of the *in silico* drug design in medicinal chemistry to optimize compounds for a selected binding site. Remarkably the presented protocol could be easily translated to different targets in order to find suitable decoration for optimizing promising hit compounds.

CONCLUSION

In the present study, we have reported the development of a computational protocol for identifying novel analogs based on recently disclosed 3-hydroxyl-pyran-4-one-2-carboxamides (HPCAR) scaffold (Sirous et al., 2019) with improved activity against HIV-1 IN. In particular, the *in silico* protocol allowed us to replace the aromatic hydrophobic moiety of HPCAR with appropriate hydrophobic aromatic/hetero-aromatic fragments. To this end, we used a combinatorial side chain hopping strategy. The resulting compounds (>37,000) were filtered using different computational methodologies. Filtering criteria included: appropriate calculated physico-chemical properties, satisfactory docking score values, visual inspection, lower ligand binding energies and proper behaviors into the HIV IN binding site assessed by MD. By using these subsequent filtering tools, we reduced the number of compounds from 1,803 to 40. Among the 40 top hit compounds, three HPCAR derivatives were chosen according to the relevant computational outputs coupled to a synthetic accessibility. After the synthesis of **HPCAR-28**, **HPCAR-89**, and **HPCAR-142**, the compounds underwent to biological evaluation in order to validate the described *in silico*

protocol. Gratifyingly, the results of pharmacological studies showed that the representative hit compounds inhibited HIV-1 IN in the low micromolar range. Among them, compound **HPCAR-28** showed the best inhibitory activity against HIV-1 IN as well as the best inhibitory activity against HIV-1 replication and HIV-1 IN strand transfer process along with a notable therapeutic index and no appreciable cell toxicity.

These promising and encouraging results provide further solid support for the potential exploitation of HPCAR scaffold in the development of anti-retroviral drugs, paving the way to the discovery of a new class of drugs against HIV-1 IN for treating HIV infection.

DATA AVAILABILITY

All datasets generated for this study are included in the manuscript/**Supplementary Files**.

AUTHOR CONTRIBUTIONS

HS carried out the computational experiments and the synthesis of the selected compounds, contributing in writing the manuscript. GCh carried out the computational experiments and performed the acquisition, analysis, interpretation of data, contributing in writing, and revising the manuscript. SG and SBu advised in the synthesis of the compounds, contributing in revising the manuscript. ZD and FC performed the biological evaluation of the selected compounds. LS advised in the synthesis of the compounds. SBr conceived, designed, and performed the computational experiments, supervised the overall work, wrote and revised the manuscript. AF advised in the synthesis of the compounds and in the computational experiments, contributing in revising the manuscript. GCa supervised the overall work and contributing in revising the manuscript. MB conceived the synthetic strategy for the synthesis of the compounds and supervised the synthesis of them, contributing in revising the manuscript.

ACKNOWLEDGMENTS

The authors wish to thank the Department of Biotechnology Chemistry and Pharmacy at University of Siena for hosting HS during her Ph.D. training.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00574/full#supplementary-material

REFERENCES

Alonso-Alija, C., Micheles, M., Schirok, H., Schlemmer, K. H., Dodd, S., Fitgerald, M., et al. (2003). *Imidazopyridinones as p38 Map Kinase Inhibitors.* WO2003008413 A1. Available online at: https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2003008413

Baell, J. B., and Holloway, G. A. (2010). New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 53, 2719–2740. doi: 10.1021/jm901137j

Brindisi, M., Gemma, S., Kunjir, S., Di Cerbo, L., Brogi, S., Parapini, S., et al. (2015). Synthetic spirocyclic endoperoxides: new antimalarial scaffolds. *Medchemcomm* 6, 357–362. doi: 10.1039/C4MD00454J

Brindisi, M., Maramai, S., Brogi, S., Fanigliulo, E., Butini, S., Guarino, E., et al. (2016). Development of novel cyclic peptides as pro-apoptotic agents. *Eur. J. Med. Chem.* 117, 301–320. doi: 10.1016/j.ejmech.2016.04.001

Brindisi, M., Ulivieri, C., Alfano, G., Gemma, S., de Asis Balaguer, F., Khan, T., et al. (2019). Structure-activity relationships, biological evaluation and structural studies of novel pyrrolonaphthoxazepines as antitumor agents. *Eur. J. Med. Chem.* 162, 290–320. doi: 10.1016/j.ejmech.2018.11.004

Brogi, S., Brindisi, M., Butini, S., Kshirsagar, G. U., Maramai, S., Chemi, G., et al. (2018). (S)-2-Amino-3-(5-methyl-3-hydroxyisoxazol-4-yl)propanoic Acid (AMPA) and kainate receptor ligands: further exploration of bioisosteric replacements and structural and biological investigation. *J. Med. Chem.* 61, 2124–2130. doi: 10.1021/acs.jmedchem.8b00099

Brogi, S., Fiorillo, A., Chemi, G., Butini, S., Lalle, M., Ilari, A., et al. (2017a). Structural characterization of Giardia duodenalis thioredoxin reductase (gTrxR) and computational analysis of its interaction with NBDHEX. *Eur. J. Med. Chem.* 135, 479–490. doi: 10.1016/j.ejmech.2017.04.057

Brogi, S., Kladi, M., Vagias, C., Papazafiri, P., Roussis, V., and Tafi, A. (2009). Pharmacophore modeling for qualitative prediction of antiestrogenic activity. *J. Chem. Inf. Model.* 49, 2489–2497. doi: 10.1021/ci900254b

Brogi, S., Ramunno, A., Savi, L., Chemi, G., Alfano, G., Pecorelli, A., et al. (2017b). First dual AK/GSK-3beta inhibitors endowed with antioxidant properties as multifunctional, potential neuroprotective agents. *Eur. J. Med. Chem.* 138, 438–457. doi: 10.1016/j.ejmech.2017.06.017

Caddick, S., Judd, D. B., Lewis, A. K. D., Reich, M. T., and Williams, M. R. V. (2003). A generic approach for the catalytic reduction of nitriles. *Tetrahedron* 59, 5417–5423. doi: 10.1016/S0040-4020(03)00858-5

Cappelli, A., Manini, M., Valenti, S., Castriconi, F., Giuliani, G., Anzini, M., et al. (2013). Synthesis and structure-activity relationship studies in serotonin 5-HT(1A) receptor agonists based on fused pyrrolidone scaffolds. *Eur. J. Med. Chem.* 63, 85–94. doi: 10.1016/j.ejmech.2013.01.044

Chatzopoulou, M., Kotsampasakou, E., and Demopoulos, V. J. (2013). Clauson-kaas-type synthesis of pyrrolyl-phenols, from the hydrochlorides of aminophenols, in the presence of nicotinamide. *Synth. Commun.* 43, 2949–2954. doi: 10.1080/00397911.2012.753460

Chiu, T. K., and Davies, D. R. (2004). Structure and function of HIV-1 integrase. *Curr. Top. Med. Chem.* 4, 965–977. doi: 10.2174/1568026043388547

Cho, A. E., Guallar, V., Berne, B. J., and Friesner, R. (2005). Importance of accurate charges in molecular docking: quantum mechanical/molecular mechanical (QM/MM) approach. *J. Comput. Chem.* 26, 915–931. doi: 10.1002/jcc.20222

Christ, F., Busschots, K., Hendrix, J., Engelborghs, Y., and Debyser, Z. (2011). "Assays for the evaluation of HIV-1 integrase enzymatic activity, DNA-binding and co-factor interaction," in *HIV-1 Integrase*, ed N. Neamati (John Wiley & Sons, Inc), 151–163. doi: 10.1002/9781118015377

Debyser, Z., Cherepanov, P., Pluymers, W., and De Clercq, E. (2001). Assays for the evaluation of HIV-1 integrase inhibitors. *Methods Mol. Biol.* 160, 139–155. doi: 10.1385/1-59259-233-3:139

Debyser, Z., Cherepanov, P., Van Maele, B., De Clercq, E., and Witvrouw, M. (2002). In search of authentic inhibitors of HIV-1 integration. *Antivir. Chem. Chemother.* 13, 1–15. doi: 10.1177/095632020201300101

Delelis, O., Carayon, K., Saib, A., Deprez, E., and Mouscadet, J. F. (2008). Integrase and integration: biochemical activities of HIV-1 integrase. *Retrovirology* 5:114. doi: 10.1186/1742-4690-5-114

Di Santo, R. (2014). Inhibiting the HIV integration process: past, present, and the future. *J. Med. Chem.* 57, 539–566. doi: 10.1021/jm400674a

Dirac, A. M., and Kjems, J. (2001). Mapping DNA-binding sites of HIV-1 integrase by protein footprinting. *Eur. J. Biochem.* 268, 743–751. doi: 10.1046/j.1432-1327.2001.01932.x

Dolan, J., Chen, A., Weber, I. T., Harrison, R. W., and Leis, J. (2009). Defining the DNA substrate binding sites on HIV-1 integrase. *J. Mol. Biol.* 385, 568–579. doi: 10.1016/j.jmb.2008.10.083

Dyda, F., Hickman, A. B., Jenkins, T. M., Engelman, A., Craigie, R., and Davies, D. R. (1994). Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science* 266, 1981–1986. doi: 10.1126/science.7801124

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* 47, 1739–1749. doi: 10.1021/jm0306430

Gasser, A., Brogi, S., Urayama, K., Nishi, T., Kurose, H., Tafi, A., et al. (2015). Discovery and cardioprotective effects of the first non-peptide agonists of the G protein-coupled prokineticin receptor-1. *PLoS ONE* 10:e0121027. doi: 10.1371/journal.pone.0121027

Hoover, W. G. (1985). Canonical dynamics - equilibrium phase-space distributions. *Physical Rev. A* 31, 1695–1697. doi: 10.1103/PhysRevA.31.1695

Huang, N., Kalyanaraman, C., Irwin, J. J., and Jacobson, M. P. (2006). Physics-based scoring of protein-ligand complexes: enrichment of known inhibitors in large-scale virtual screening. *J. Chem. Inf. Model.* 46, 243–253. doi: 10.1021/ci0502855

Humphreys, D. D., Friesner, R. A., and Berne, B. J. (1994). A multiple-time-step molecular-dynamics algorithm for macromolecules. *J. Phys. Chem.* 98, 6885–6892. doi: 10.1021/j100078a035

Hwang, Y., Rhodes, D., and Bushman, F. (2000). Rapid microtiter assays for poxvirus topoisomerase, mammalian type IB topoisomerase and HIV-1 integrase: application to inhibitor isolation. *Nucleic Acids Res.* 28, 4884–4892. doi: 10.1093/nar/28.24.4884

Illingworth, C. J., Morris, G. M., Parkes, K. E., Snell, C. R., and Reynolds, C. A. (2008). Assessing the role of polarization in docking. *J. Phys. Chem. A* 112, 12157–12163. doi: 10.1021/jp710169m

Irwin, J. J., and Shoichet, B. K. (2005). ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 45, 177–182. doi: 10.1021/ci049714+

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. (2012). ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* 52, 1757–1768. doi: 10.1021/ci3001277

Johns, B. A., and Svolto, A. C. (2008). Advances in two-metal chelation inhibitors of HIV integrase. *Expert. Opin. Ther. Pat.* 18, 1225–1237. doi: 10.1517/13543776.18.11.1225

Jorgensen, W. L., Maxwell, D. S., and TiradoRives, J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* 118, 11225–11236. doi: 10.1021/ja9621760

Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001). Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* 105, 6474–6487. doi: 10.1021/jp003919d

Katlama, C., and Murphy, R. (2012). Dolutegravir for the treatment of HIV. *Expert. Opin. Investig. Drugs* 21, 523–530. doi: 10.1517/13543784.2012.661713

Kawasuji, T., Fuji, M., Yoshinaga, T., Sato, A., Fujiwara, T., and Kiyama, R. (2006a). A platform for designing HIV integrase inhibitors. Part 2: a two-metal binding model as a potential mechanism of HIV integrase inhibitors. *Bioorg. Med. Chem.* 14, 8420–8429. doi: 10.1016/j.bmc.2006.08.043

Kawasuji, T., Yoshinaga, T., Sato, A., Yodo, M., Fujiwara, T., and Kiyama, R. (2006b). A platform for designing HIV integrase inhibitors. Part 1: 2-hydroxy-3-heteroaryl acrylic acid derivatives as novel HIV integrase inhibitor and modeling of hydrophilic and hydrophobic pharmacophores. *Bioorg. Med. Chem.* 14, 8430–8445. doi: 10.1016/j.bmc.2006.08.044

Khurana, J. M., and Gogia, A. (1997). Synthetically useful reactions with Nickle boride. A review. *Org. Prep. Proced. Int.* 29, 1–32. doi: 10.1080/00304949709355171

Khurana, J. M., and Kukreja, G. (2002). Rapid reduction of nitriles to primary amines with nickel boride at ambient temperature. *Synth. Commun.* 32, 1265–1269. doi: 10.1081/SCC-120003619

Kosak, T. M., Conrad, H. A., Korich, A. L., and Lord, R. L. (2015). Ether cleavage re-investigated: elucidating the mechanism of BBr3-facilitated demethylation of aryl methyl ethers. *Eur. J. Org. Chem.* 2015, 7460–7467. doi: 10.1002/ejoc.201501042

Lagorce, D., Maupetit, J., Baell, J., Sperandio, O., Tuffery, P., Miteva, M. A., et al. (2011). The FAF-Drugs2 server: a multistep engine to prepare electronic chemical compound collections. *Bioinformatics* 27, 2018–2020. doi: 10.1093/bioinformatics/btr333

Lagorce, D., Sperandio, O., Galons, H., Miteva, M. A., and Villoutreix, B. O. (2008). FAF-Drugs2: free ADME/tox filtering tool to assist drug discovery and chemical biology projects. *BMC Bioinformatics* 9:396. doi: 10.1186/1471-2105-9-396

Liao, C., Marchand, C., Burke, T. R. Jr., Pommier, Y., and Nicklaus, M. C. (2010). Authentic HIV-1 integrase inhibitors. *Future Med. Chem.* 2, 1107–1122. doi: 10.4155/fmc.10.199

Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.* 46, 3–26. doi: 10.1016/S0169-409X(00)00129-0

Ma, Y., and Hider, R. C. (2015). Novel synthetic approach to fluoro- and amido-disubstituted 3-hydroxypyridin-4-ones. *J. Fluor. Chem.* 173, 29–34. doi: 10.1016/j.jfluchem.2015.02.005

Maquiaveli, C. C., Lucon-Junior, J. F., Brogi, S., Campiani, G., Gemma, S., Vieira, P. C., et al. (2016). Verbascoside inhibits promastigote growth and arginase activity of *Leishmania amazonensis*. *J. Nat. Prod.* 79, 1459–1463. doi: 10.1021/acs.jnatprod.5b00875

Martyna, G. J., Tobias, D. J., and Klein, M. L. (1994). constant-pressure molecular-dynamics algorithms. *J. Chem. Phys.* 101, 4177–4189. doi: 10.1063/1.467468

Neamati, N., Lin, Z., Karki, R. G., Orr, A., Cowansage, K., Strumberg, D., et al. (2002). Metal-dependent inhibition of HIV-1 integrase. *J. Med. Chem.* 45, 5661–5670. doi: 10.1021/jm0201417

Nickolls, J., Buck, I., Garland, M., and Skadron, K. (2008). Scalable parallel programming with CUDA. *Queue* 6, 40–53. doi: 10.1145/1365490.1365500

Paolino, M., Brindisi, M., Vallone, A., Butini, S., Campiani, G., Nannicini, C., et al. (2018). Development of potent inhibitors of the *Mycobacterium tuberculosis* virulence factor Zmp1 and evaluation of their effect on mycobacterial survival inside macrophages. *ChemMedChem* 13, 422–430. doi: 10.1002/cmdc.201700759

Pauwels, R., Balzarini, J., Baba, M., Snoeck, R., Schols, D., Herdewijn, P., et al. (1988). Rapid and automated tetrazolium-based colorimetric assay for the detection of anti-HIV compounds. *J. Virol. Methods* 20, 309–321. doi: 10.1016/0166-0934(88)90134-6

Pearlman, D. A., and Charifson, P. S. (2001). Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system. *J. Med. Chem.* 44, 3417–3423. doi: 10.1021/jm0100279

Pommier, Y., Johnson, A. A., and Marchand, C. (2005). Integrase inhibitors to treat HIV/AIDS. *Nat. Rev. Drug. Discov.* 4, 236–248. doi: 10.1038/nrd1660

Rogolino, D., Carcelli, M., Sechi, M., and Neamati, N. (2012). Viral enzymes containing magnesium: metal binding as a successful strategy in drug design. *Coord. Chem. Rev.* 256, 3063–3086. doi: 10.1016/j.ccr.2012.07.006

Rostami, M., Sirous, H., Zabihollahi, R., Aghasadeghi, M. R., Sadat, S. M., Namazi, R., et al. (2015). Design, synthesis and anti-HIV-1 evaluation of a series of 5-hydroxypyridine-4-one derivatives as possible integrase inhibitors. *Med. Chem. Res.* 24, 4113–4127. doi: 10.1007/s00044-015-1443-4

Rowley, M. (2008). The discovery of raltegravir, an integrase inhibitor for the treatment of HIV infection. *Prog. Med. Chem.* 46, 1–28. doi: 10.1016/S0079-6468(07)00001-X

Santos, M. A., Marques, S. M., and Chaves, S. (2012). Hydroxypyridinones as "privileged" chelating structures for the design of medicinal drugs. *Coord. Chem. Rev.* 256, 240–259. doi: 10.1016/j.ccr.2011.08.008

Sato, M., Kawakami, H., Motomura, T., Aramaki, H., Matsuda, T., Yamashita, M., et al. (2009). Quinolone carboxylic acids as a novel monoketo acid class of human immunodeficiency virus type 1 integrase inhibitors. *J. Med. Chem.* 52, 4869–4882. doi: 10.1021/jm900460z

Sheehan, J. C., Preston, J., and Cruickshank, P. A. (1965). A rapid synthesis of oligopeptide derivatives without isolation of intermediates. *J. Am. Chem. Soc.* 87, 2492–2493. doi: 10.1021/ja01089a034

Sirous, H., Fassihi, A., Brogi, S., Campiani, G., Christ, F., Debyser, Z., et al. (2019). Synthesis, molecular modelling and biological studies of 3-hydroxy-pyrane-4-one and 3-hydroxy-pyridine-4-one derivatives as HIV-1 integrase inhibitors. *Med. Chem.* doi: 10.2174/1573406415666181219113225. [Epub ahead of print].

Sirous, H., Zabihollahi, R., Aghasadeghi, M. R., Sadat, S. M., Saghaie, L., and Fassihi, A. (2015). Docking studies of some 5-hydroxypyridine-4-one derivatives: evaluation of integrase and ribonuclease H domain of reverse transcriptase as possible targets for anti-HIV-1 activity. *Med. Chem. Res.* 24, 2195–2212. doi: 10.1007/s00044-014-1289-1

Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* 112, 6127–6129. doi: 10.1021/ja00172a038

Tan, J., Chen, Y., Li, H., and Yasuda, N. (2014). Suzuki-Miyaura cross-coupling reactions of unprotected haloimidazoles. *J. Org. Chem.* 79, 8871–8876. doi: 10.1021/jo501326r

Taylor, R. D., Jewsbury, P. J., and Essex, J. W. (2002). A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.* 16, 151–166. doi: 10.1023/A:1020155510718

Vallone, A., D'Alessandro, S., Brogi, S., Brindisi, M., Chemi, G., Alfano, G., et al. (2018). Antimalarial agents against both sexual and asexual parasites stages: structure-activity relationships and biological studies of the Malaria Box compound 1-[5-(4-bromo-2-chlorophenyl)furan-2-yl]-N-[(piperidin-4-yl)methyl]methanamine (MMV019918) and analogues. *Eur. J. Med. Chem.* 150, 698–718. doi: 10.1016/j.ejmech.2018.03.024

Zaccagnini, L., Brogi, S., Brindisi, M., Gemma, S., Chemi, G., Legname, G., et al. (2017). Identification of novel fluorescent probes preventing PrP(Sc) replication in prion diseases. *Eur. J. Med. Chem.* 127, 859–873. doi: 10.1016/j.ejmech.2016.10.064

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a shared affiliation, though no other collaboration, with one of the authors MB.

Copyright © 2019 Sirous, Chemi, Gemma, Butini, Debyser, Christ, Saghaie, Brogi, Fassihi, Campiani and Brindisi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

PERSPECTIVE
published: 16 August 2019
doi: 10.3389/fchem.2019.00564

Pushing the Ligand Efficiency Metrics: Relative Group Contribution (RGC) Model as a Helpful Strategy to Promote a Fragment "Rescue" Effect

Andrés Felipe Vásquez[1,2] *and Andrés González Barrios*[1]*

[1] Grupo de Diseño de Productos y Procesos (GDPP), School of Chemical Engineering, Universidad de los Andes, Bogotá, Colombia, [2] Laboratorio de Fisiología Molecular, Instituto Nacional de Salud, Bogotá, Colombia

The ligand efficiency (LE) indexes have long been used as decision-making criteria in drug discovery and development. However, in the context of fragment-based drug design (FBDD), these metrics often exhibit a strong emphasis toward the selection of highly efficient "core" fragments for potential optimization, which are not usually considered as parts of a larger molecule with a size typical for a drug. In this study, we present a relative group contribution (RGC) model intended to predict the efficiency of a drug-sized compound in terms of its component fragments. This model could be useful not only in rapidly predicting all the possible combinations of promising fragments from an earlier hit discovery stage, but also in enabling a relatively low-LE fragment to become part of a drug-sized compound as long as it is "rescued" by other high-LE fragments.

Keywords: ligand efficiency metrics, fragment-based screening, property-based design, drug discovery, protein-ligand interactions, structure-activity relationship, fragment library

OPEN ACCESS

Edited by:
Jose L. Medina-Franco, National Autonomous University of Mexico, Mexico

Reviewed by:
Sergio Hidalgo Figueroa, Instituto Potosino de Investigación Científica y Tecnológica (IPICYT), Mexico
Salvatore Guccione, University of Catania, Italy

***Correspondence:**
Andrés González Barrios
andgonza@uniandes.edu.co

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *19 May 2019*
Accepted: *24 July 2019*
Published: *16 August 2019*

Citation:
Vásquez AF and González Barrios A (2019) Pushing the Ligand Efficiency Metrics: Relative Group Contribution (RGC) Model as a Helpful Strategy to Promote a Fragment "Rescue" Effect. Front. Chem. 7:564. doi: 10.3389/fchem.2019.00564

INTRODUCTION

Ligand efficiency metrics have been applied as a decision-making strategy nearly universally accepted during the last two decades (Hopkins et al., 2014). They are intended to compare the quality of hits and leads during hit to lead and lead optimization phases of drug discovery (Cavalluzzi et al., 2017). Among these metrics, the ligand efficiency (LE), firstly described by Hopkins et al. (2004), continues to be the most widely used index to discriminate between promising molecules and those which are not (Reynolds, 2015). LE is defined by the following formula:

$$LE = \frac{\Delta G}{N} \quad (1)$$

where $\Delta G = -RTInK_d$, N (also known as HAC or heavy atom count) represents the number of heavy, non-hydrogen atoms, and K_d corresponds to the equilibrium dissociation constant.

As shown in Equation (1), LE normalizes the potency by size, specifically representing the average contribution ΔG (Gibbs free energy) per heavy atom. LE is typically used in FBDD as cut-off criterion to retrieve just high-LE fragments in a screening process (Murray and Verdonk, 2006; Schultes et al., 2010). Intriguingly, because LE usually considers fragments as independent chemical entities, some other metrics have emerged to consider the change in affinity as a fragment is developed into a larger, high-affinity drug-sized compound. One prime example is

the group efficiency (GE), described in 2008 (Verdonk and Rees, 2008), which allows measuring the contribution to the binding efficiency of a particular group of atoms added to an existing lead molecule:

$$GE = \frac{\Delta\Delta G}{\Delta N} \quad (2)$$

where $\Delta\Delta G = \Delta G(B) - \Delta G(A)$ and $\Delta N = (N(B) - N(A)$; in other words, $\Delta\Delta G$ is equal to the difference between the Gibbs free energy of the existing molecule (or fragment) "A" and the new combined molecule "B" and ΔN corresponds to the difference between the number of non-hydrogen atoms of molecules "A" and "B." However, GE is based on a pairwise comparison of structurally closely related compounds and, hence, it is frequently applied for optimization of a high-efficient fragment (Hopkins et al., 2014). Therefore, a rapid and simple method for comparing efficiency of different fragments as part of a whole, including if they are dissimilar to each other (or if they occupy different pockets in the target molecule), needs to be developed.

Several independent studies in the past two decades have indicated an overemphasis on potency by the pharmaceutical industry (Albert et al., 2007; Hopkins et al., 2014). Still, other factors such as chemical novelty (Medina-Franco et al., 2014), selectivity fine-tuning (Costantino and Barlocco, 2018), structural alerts avoidance (Jasial et al., 2017), and synthetic accessibility (Fukunishi et al., 2014) are increasingly playing a key role in drug design. Considering these non-mutually exclusive events, we hypothesize that fragments that not necessarily exhibit a high efficiency level during a screening procedure, either virtual or experimental, would still have the possibility of taking part in a complete drug-sized compound.

In this study, we propose that a relative group contribution (RGC) model based on the efficiency of its component fragments may estimate the efficiency of a drug-sized compound. This model calculates the minimum efficiency required for unknown fragments by considering the efficiency of those already known, which facilitates a rapid elucidation of the best combinations of fragments. Likewise, this model facilitates that fragments with a relatively low efficiency may not necessarily be eliminated at an early stage of the screening process and, consequently, may become eventually represented as chemical moieties within the final candidate compound -a phenomenon herein referred to as fragment "rescue" effect.

THEORETICAL FRAMEWORK OF THE RGC MODEL

The proposed model is based on three main assumptions:

1. The efficiency of an entire molecule may be estimated as the weighted root mean square (WMRS) of the efficiency of its component fragments. The rational for using this type of mean is intended to (1) cope with the negative values of ΔG and (2) consider effectively the potentially different number of non-hydrogen atoms (N) for each component fragment.
2. The efficiency of each fragment (regarded as the entire ratio and not just the quotient) is, in principle, dependent on each other, excepting in cases of two fragments when the N for them is equal to each other (and then their weight is equivalent), or in cases when three or more fragments are involved.
3. The efficiency of each fragment is directly calculated from the ΔG resulting in its direct interaction with a specific location (i.e., binding site or pocket) in a particular receptor.

Our hypothesis assumes, according to its first principle, that the WRMS of the LEs of the fragments composing an entire molecule ($\overline{LE_q}$) corresponds to the actual LE of this latter (LE_T) (A comprehensive list of mathematical terms is shown in **Supplementary Material**). Because this mean is intended to be proportionally similar to the real, total LE of an entire molecule (LE_T), we refer herein to it as the apparent total LE (LE_T^{app}):

$$\overline{LE}_q = LE_T^{app} \approx LE_T \quad (3)$$

For clarity of the RGC concept, we consider first the LE_T^{app} as a simple arithmetic mean:

$$LE_T^{app} = \frac{1}{x}(LE_1 + LE_2 + LE_3 + \ldots + LE_x) \quad (4)$$

where LE_i corresponds to the LE of the component fragments (LE_1, LE_2, etc.), and x refers to the number of fragments composing the molecule. Therefore, once LE is expressed in terms of the Equation (1):

$$LE_T^{app} = \frac{1}{x}\left(\frac{\Delta G_1}{N_1} + \frac{\Delta G_2}{N_2} + \frac{\Delta G_3}{N_3} + \ldots + \frac{\Delta G_x}{N_x}\right) \cong \frac{\Delta G_T}{xN_T} \quad (5)$$

where ΔG_i is the change in Gibbs free energy for each composing fragment (ΔG_1, ΔG_2, etc.) up to a maximum number of fragments x and N_i correspond to the number of non-hydrogen atoms of each fragment. Similarly, ΔG_T and N_T represent the change in Gibbs free energy and number of non-hydrogen atoms for the entire molecule, respectively. Should be remembered that (5) is based on an arithmetic mean and hence it assumes that N is equal among all composing fragments, so that it would just be applicable in this specific scenario.

If we consider the Equation (5) for a molecule composed by a single fragment:

$$LE_T^{app} = \frac{1}{x}\sum_{i=1}^{x}\left(\frac{\Delta G}{N}\right)_i = LE_1 \quad (6)$$

we can observe that LE_T^{app} correspond to the LE value of the unique component fragment, namely LE1, which supports a scenario where the component fragment is also the entire "final" molecule. However, expressing (5) for a molecule composed by two fragments:

$$LE_T^{app} = \frac{1}{x}\sum_{i=1}^{x}\left(\frac{\Delta G}{N}\right)_i = \frac{1}{2}\sum_{i=1}^{2}\left(\frac{\Delta G}{N}\right)_i = \frac{1}{2}(LE_1 + LE_2) \quad (7)$$

we could notice that, in contrast to the one-fragment case, the existence of more than one compound allows for the solution of the equation in terms of a particular fragment:

$$LE_2 = 2LE_T^{app} - LE_1 \quad (8)$$

The last equation poses a simple but important principle: Starting from an "ideal" LE_T^{app}, a particular low-LE fragment can be successfully chosen or "rescued" by one or more high-LE fragments. Now, if we consider a three-fragment case:

$$LE_T^{app} = \frac{1}{x}\sum_{i=1}^{x}\left(\frac{\Delta G}{N}\right)_i = \frac{1}{3}\sum_{i=1}^{3}\left(\frac{\Delta G}{N}\right)_i = \frac{1}{3}(LE_1 + LE_2 + LE_3) \quad (9)$$

Interestingly, for this case, even if we assume in this example that we know LE_1, it is still possible to consider a LE value for the unknown fragments LE_2 and LE_3 grouping them together into a single term:

$$LE_T^{app} = \frac{1}{3}(LE_1 + 2LE_\delta) \quad (10)$$

where the LE delta (LE_δ) corresponds to a transient, "ideal" value intended to be equal for all the fragments which individual LE is still unknown. Therefore, as we will discuss below, this value will be modified as long as new LE values are known for fragments, independently of their position.

Likewise, assuming also that we just know LE_1, we could express LE_T^{app} for the two-fragment case:

$$LE_T^{app} = \frac{1}{2}(LE_1 + LE_\delta) \quad (11)$$

which would indicate that:

$$LE_\delta = LE_2 \quad (12)$$

This result suggests, as we also discuss below, that a LE_δ is expected to equal the LE value of a last fragment to be known (LE_u), independently of the number of fragments and their position. This behavior corresponds to a subtractive average (SA). Remarkably, although the nature of LE_δ is somewhat similar to the cumulative average (CA) or moving average (MA), the number of fragments with unknown LE is continuously decreasing and LE_δ does not "run" within a predetermined window size.

Finally, assuming also that we just know LE_1, we could express LE_T^{app} for the one-fragment case:

$$LE_T^{app} = [\,LE_1 + 0\,(LE_\delta)\,] = LE_1 \quad (13)$$

indicating that LE_δ can only be calculated if there are at least two starting fragments and, more importantly, it is especially useful in cases of three or more of them.

At this point, think of the LE_T^{app} as a whole for an undetermined series of fragments:

$$LE_T^{app} = \frac{1}{x}(LE_1 + \ldots + LE_x) \quad (14)$$

If we assume again that we just know LE_1, it is possible to rearrange LE_T^{app} using LE_δ:

$$LE_T^{app} = \frac{1}{x}[LE_1 + (x-1)\,LE_\delta] \quad (15)$$

where (x-1) corresponds to the coefficient of the LE_δ value independently of the number of starting fragments as shown in Equations (10, 11, 13). Hence, if we resolve for LE_δ:

$$LE_\delta = \frac{1}{x-1}(xLE_T^{app} - LE_1) \quad (16)$$

Now, if we take into account that the number 1 in this equation actually corresponds to the number of known LE values of fragments or a, we can observe that:

$$if\ (x - a) \to 1,\ then\ LE_\delta \to LE_u \quad (17)$$

What means that the more (x-a) tends to 1, the more LE_δ tends to LE_u, just as we saw previously in Equation (12). Finally, considering the formula for LE_δ in terms of an undetermined number of fragments with different known and unknown LE values:

$$LE_\delta = \frac{1}{x - a}\left[\sum_{i=1}^{x}\left(\frac{\Delta G}{N}\right)_i - \sum_{j=0}^{a}\left(\frac{\Delta G}{N}\right)_j\right] \quad (18)$$

$$if\ and\ only\ if \begin{cases} 1 \le x < \infty \\ x \in \mathbb{N} \\ 0 \le a < x \\ a \in \mathbb{Z}_0^+ \end{cases}$$

where the first summation term indicates the "ideal" sum of LE values for the existing fragments (LE_i) as if they would have the same value, and the second summation term refers to the "real" sum of all fragments which LE value is already known (LE_j). On the other hand, if a = o, LE_0 would not proceed as a real value (and by extension the second summation term). Therefore, in this specific case a consequence would be that:

$$LE_\delta = LE_T^{app} \quad (19)$$

Now, after having explained the basic concepts of RGC and LE_δ, we could express LE_T^{app} in terms of the WMRS according to our hypothesis:

$$LE_T^{app} = \sqrt{\sum_{i=1}^{x} LE_i^2 w_i / \sum_{i=1}^{x} w_i} = \sqrt{\sum_{i=1}^{x}\left(\frac{\Delta G}{N}\right)_i^2 w_i / \sum_{i=1}^{x} w_i} \quad (20)$$

where LE_i corresponds to the LE of each component fragment, w_i refers to the weight of each fragment (depending on the N of each one) and x refers to the number of fragments.

Likewise, our formula for LE_δ would be:

$$LE_\delta = \frac{1}{\sqrt{w_\delta (x - a)}}\left[\sqrt{\sum_{i=1}^{x}\left(\frac{\Delta G}{N}\right)_i^2 w_i - \sum_{j=0}^{a}\left(\frac{\Delta G}{N}\right)_j^2 w_j}\right] \quad (21)$$

$$if\ and\ only\ if \begin{cases} 1 \le x < \infty \\ x \in \mathbb{N} \\ 0 \le a < x \\ a \in \mathbb{Z}_0^+ \end{cases}$$

The additional term in this equation, namely W_δ, corresponds to the "ideal" weight of all fragments with unknown LE value, as if they would have the same value. Just as with LE_δ, this parameter is expected to change with every new LE value of fragment known, until the value (weight) corresponding to the last fragment with unknown LE value is adopted.

FRAGMENT SELECTION AND LE_T PREDICTION BY THE RGC MODEL

As presented in a hypothetical example (**Figure 1**), three central premises may be elucidated for the RGC model by selecting hit compounds, starting from a cut-off LE value:

1. A low-efficiency fragment (i.e., with a LE $< LE_T^{app}$) can be rescued IF there exists high-efficiency fragments (i.e., with a LE $> LE_T^{app}$) in ALL the other positions (i.e., binding sites or pockets) of the target molecule, usually a protein. If this condition is not satisfied, the fragment could be automatically rejected from the set of possible combinations of fragments for an entire molecule.
2. If there are no high-efficiency fragments in all the other positions simultaneously, a molecular fragment with low-efficiency fragment can still be rescued IF at least one fragment in any other position with an efficiency high enough to reach exists, in average with the first, the LE_T^{app}. This implies, therefore, that once a fragment is rescued by one or more high-efficiency fragments in other positions, any fragment in subsequent unexplored positions is just required to have a mid-efficiency (with a LE $\cong LE_\delta$ ideal) as a minimum. Additionally, each time LE_δ changes, a differential "pushing" over the LE of unknown fragments occurs in terms of ΔG and/or N, which could be potentially modified to achieve an acceptable LE value and could therefore be selected.
3. Even if a low-efficiency fragment is not rescued after implementing the strategies stated in Equation (1, 2), a potential rescue could still take place exploring a more diverse library sample of fragments, assuming that (1) you are dealing with a number of fragments well-below under the maximum theoretical chemical space for them (a population of about 10^7 molecules) and (2) the fragment is not an outlier compared to other fragments intended to combine.

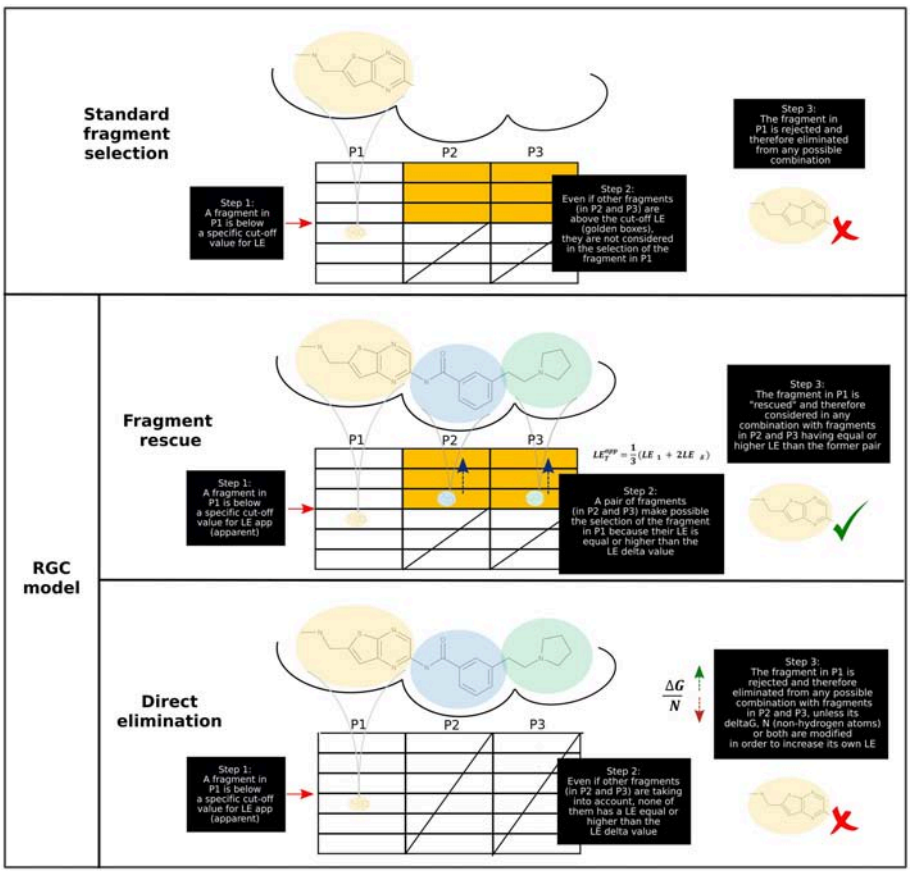


FIGURE 1 | Application of RGC model to a hypothetical fragment-based drug design (FBDD) campaign. (**Upper**) In the "standard" or classical screening approach, a fragment is selected (i.e., can be part of a final drug-size compound) depending exclusively upon their own LE. If this parameter is not equal or greater than a pre-established cut-off value, the fragment is rejected. (**Lower**) According to the RGC model, a fragment is selected depending on the fragments on the other positions. Based on the presence of high-LE fragments in alternative positions (illustrated by yellow boxes), a low-LE fragment may become either rescued or rapidly discarded (using the dynamic LE_δ value in both cases).

In addition, and according to our preliminary results, we found that LE_T^{app} values predicted by this model were consistent with the LE_T values calculated experimentally from a set of 16 drug-sized molecules taken from scientific literature (**Figure 2**, **Table 1S**).

DISCUSSION

The present study pretends to propose the RGC model as an innovative and effective approach to apply in drug design. This model and, especially, the fragment "rescue" effect that is conceptually implicit, offer an alternative for the long-standing FBDD paradigm of designing compounds merely based on the intrinsic binding energy of fragments, facilitating the introduction of other decision–making criteria that are becoming increasingly common.

If the principles of the RGC model are considered together, it is possible to elucidate two major advantages. First, we count on a limited amount of data and in order to more clearly reveal any trends, LE_T^{app} appears to increase as much as the LE_T, and there appears to be no dramatic shift toward higher efficiencies for particular fragments or protein targets. Secondly, since a particular fragment could be directly rejected early in the process and there are many fragments by pocket in a typical FBDD campaign, this model might dramatically reduce the computational and synthetic costs, respectively (which is especially true in cases of three or more pockets).

The RGC model is, however, not free of inherent shortcomings. As a LE-derived metric, all fragments are assumed to maintain equal orientations both individually and as part of a larger chemical compound (Zartler and Shapiro, 2008), and phenomena such as hot spots (Zerbe et al., 2012; Rathi et al., 2017) or synergy (also called "super-additivity") (Hebeisen et al., 2008; Nazaré et al., 2012) are not directly considered. Likewise, because its average-based nature, ΔG of each fragment is normalized not only at the number of non-hydrogen atoms but also on the number of component fragments. Therefore, the less accurate (or more extreme) ΔG values for fragments in each position are, the greater the difference expected between the LE_T^{app} and the LE_T. However, we believe that the impact of these hurdles could be minimized, improving the confidence of any potential fragment "rescue," if additional energy terms derived from rigid body barrier (ΔG_{rigid}), linker binding ($\Delta G_{binding}$) or strain (ΔG_{strain}) (Murray and Verdonk, 2006; Cherry and Mitchell, 2008) are included and the ΔG values are both accurate and above a reasonable cut-off value.

A final examination about the implications of this work leads us to assert that the fragment "rescue" phenomenon is far from being new: it has already occurred and continues to

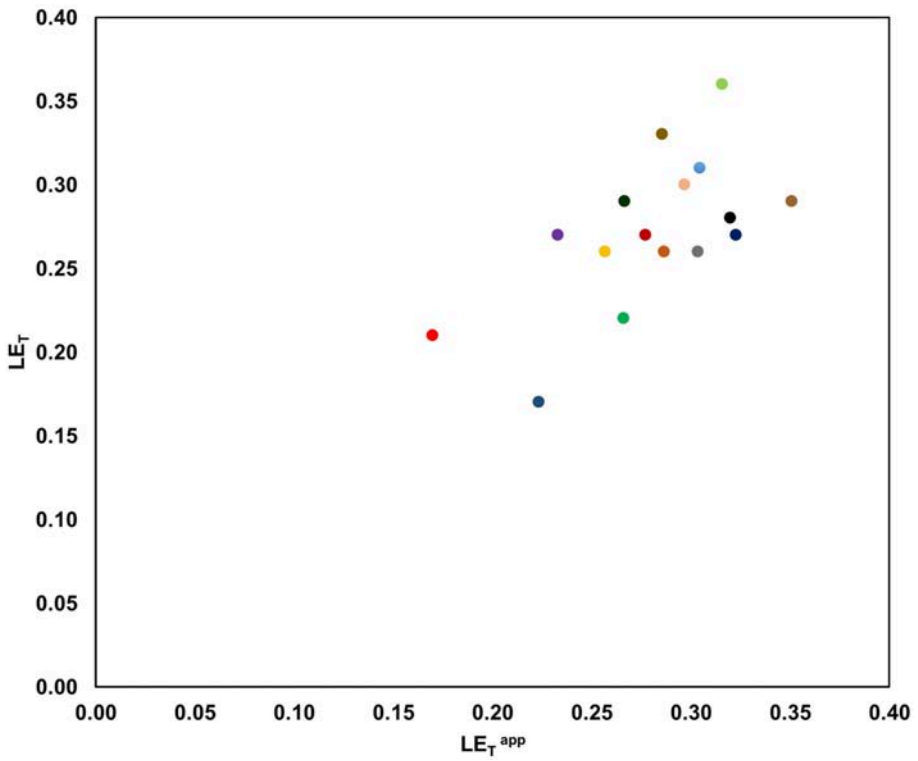


FIGURE 2 | Preliminary comparison between LE_T^{app} and LE_T for a set of 16 drug-sized literature compounds developed during FBDD studies. All compounds were elaborated using a linking strategy (on two fragments) for 10 different protein targets [●/●LDHA ●Replication Protein A (RPA) ●/●Blood coagulation factor Xa ●Bcl-2 ●DOT1L ●Hsp90 ●/●Pantothenate synthetase (PtS) ●Blood coagulation factor XIa ●CK2 ●BACE1 ●Endothiapepsin (Epn) ●Bcl-x_L ●PKM2].

occur, but usually during a lead optimization instead an earlier hit discovery phase. The rationale for this statement lies in two central facts. First, we currently know that the LE value of a compound tends to decrease during optimization process (Bembenek et al., 2009), while both its lipophilicity and its MW (and, hence, the number of non-hydrogen atoms -N) tends to increase (Ferenczy and Keseru, 2016). Second, the energy of supramolecular interactions is widely known to be largely different depending on the chemical moiety involved and, thus, ionic and hydrogen bonds are expected to account for a larger part of the drug-receptor binding energy (i.e., its ΔG) compared with hydrophobic interactions (Ermondi and Caron, 2006). Therefore, it is decidedly inviting to believe that, in many drug discovery initiatives, both the increase in LE and the decrease in MW and lipophilicity observed during lead optimization-could be explained by the addition of large and hydrophobic chemical moieties such as those cyclic aliphatic, which have a much smaller $\Delta G/N$ ratio compared to other fragments. Interestingly, these aliphatic moieties have been recently suggested by some authors to be more "developable" compared its aromatic analogs, which could be an additional factor behind this phenomenon (Lovering et al., 2009).

The RGC model presented in this study is based on the assumption that the LE of a drug-sized molecule may be estimated using the relative contribution of each component fragment. We believe this model could serve as a complementary benchmark for medicinal chemists in experimental or virtual fragment-based screening campaigns. Likewise, we consider that the RGC model could be implemented with other metrics based on either LE or a potency/size ratio and could be eventually adjusted to consider not only "linking" but also "growing" or "merging" as alternative fragment elaboration strategies.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

AV planned and performed the entire *in silico* work presented in this study, contributed to the analysis and interpretation of data, and assisted the writing, editing, and submission of this manuscript. AG made substantial contributions to the data analysis, critical revision for important intellectual content, and document editing. All authors have read and approved the final manuscript.

FUNDING

We acknowledge funds from the Colombian Department of Science, Technology and Innovation COLCIENCIAS Grant No. 727.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Professor Marco de Vivo from Istituto Italiano di Tecnologia (IIT), in Genoa (Italy) for helpful discussions and recommendations.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00564/full#supplementary-material

REFERENCES

Albert, J. S., Blomberg, N., Breeze, A., Brown, A., Burrows, J., Edwards, P., et al. (2007). An integrated approach to fragment-based lead generation: philosophy, strategy and case studies from AstraZeneca's Drug Discovery Programmes. *Curr. Top. Med. Chem.* 7, 1600–1629. doi: 10.2174/1568026077823 41091

Bembenek, S. D., Tounge, B. A., and Reynolds, C. H. (2009). Ligand efficiency and fragment-based drug discovery. *Drug Discov. Today* 14, 278–283. doi: 10.1016/j.drudis.2008.11.007

Cavalluzzi, M. M., Mangiatordi, G. F., Nicolotti, O., and Lentini, G. (2017). Ligand efficiency metrics in drug discovery: the pros and cons from a practical perspective. *Expert Opin. Drug Discov.* 12, 1087–1104. doi: 10.1080/17460441.2017.1365056

Cherry, M., and Mitchell, T. (2008). "Introduction to fragment-based drug discovery," in *Fragment-Based Drug Discovery*, eds E. R. Zartler and M. J. Shapiro, 1–32. doi: 10.1002/9780470721551.ch1

Costantino, L., and Barlocco, D. (2018). "Designing Approaches to Multitarget Drugs," in *Drug Selectivity: An Evolving Concept in Medicinal Chemistry*, 161–205.

Ermondi, G., and Caron, G. (2006). Recognition forces in ligand-protein complexes: blending information from different sources. *Biochem. Pharmacol.* 72, 1633–1645. doi: 10.1016/j.bcp.2006.05.022

Ferenczy, G. G., and Keseru, G. M. (2016). "Ligand efficiency metrics and their use in fragment optimizations," in *Fragment-Based Drug Discovery Lessons and Outlook*, eds D. A. Erlanson and W. Jahnke (Wiley-VCH Verlag GmbH & Co. KGaA), 75–98. doi: 10.1002/9783527683604.ch04

Fukunishi, Y., Kurosawa, T., Mikami, Y., and Nakamura, H. (2014). Prediction of synthetic accessibility based on commercially available compound databases. *J. Chem. Inf. Model.* 54, 3259–3267. doi: 10.1021/ci500568d

Hebeisen, P., Kuhn, B., Kohler, P., Gubler, M., Huber, W., Kitas, E., et al. (2008). Allosteric FBPase inhibitors gain 10(5) times in potency when simultaneously binding two neighboring AMP sites. *Bioorganic Med. Chem. Lett.* 18, 4708–4712. doi: 10.1016/j.bmcl.2008.06.103

Hopkins, A. L., Groom, C. R., and Alex, A. (2004). Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* 9, 430–431. doi: 10.1016/S1359-6446(04)03069-7

Hopkins, A. L., Keser,ü, G. M., Leeson, P. D., Rees, D. C., and Reynolds, C. H. (2014). The role of ligand efficiency metrics in drug discovery. *Nat. Rev. Drug Discov.* 13, 105–121. doi: 10.1038/nrd4163

Jasial, S., Hu, Y., and Bajorath, J. (2017). How frequently are pan-assay interference compounds active? Large-scale analysis of screening data reveals diverse activity profiles, low global hit frequency, and many consistently inactive compounds. *J. Med. Chem.* 60, 3879–3886. doi: 10.1021/acs.jmedchem.7b00154

Lovering, F., Bikker, J., and Humblet, C. (2009). Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* 52, 6752–6756. doi: 10.1021/jm901241e

Medina-Franco, J. L., Martinez-Mayorga, K., and Meurice, N. (2014). Balancing novelty with confined chemical space in modern drug discovery. *Expert Opin. Drug Discov.* 9, 151–165. doi: 10.1517/17460441.2014.872624

Murray, C. W., and Verdonk, M. L. (2006). "Entropic consequences of linking ligands," in *Fragment-Based Approaches in Drug Discovery*, eds R. Mannhold, H. Kubinyi, G. Folkers, W. Jahnke, and D. A. Erlanson (Weinheim: Wiley-VCG Verlag GmbH & Co. KGaA), 55–66. doi: 10.1002/3527608761.ch3

Nazaré, M., Matter, H., Will, D. W., Wagner, M., Urmann, M., Czech, J., et al. (2012). Fragment deconstruction of small, potent factor xa inhibitors: exploring the superadditivity energetics of fragment linking in protein-ligand complexes. *Angew. Chemie Int. Ed.* 51, 905–911. doi: 10.1002/anie.201107091

Rathi, P. C., Ludlow, R. F., Hall, R. J., Murray, C. W., Mortenson, P. N., and Verdonk, M. L. (2017). Predicting "Hot" and "Warm" Spots for Fragment Binding. *J. Med. Chem.* 60, 4036–4046. doi: 10.1021/acs.jmedchem.7b00366

Reynolds, C. H. (2015). Ligand efficiency metrics: why all the fuss? *Future Med. Chem.* 7, 1363–1365. doi: 10.4155/fmc.15.70

Schultes, S., De Graaf, C., Haaksma, E. E. J., De Esch, I. J. P., Leurs, R., and Krämer, O. (2010). Ligand efficiency as a guide in fragment hit selection and optimization. *Drug Discov. Today Technol.* 7, 157–162. doi: 10.1016/j.ddtec.2010.11.003

Verdonk, M. L., and Rees, D. C. (2008). Group efficiency: a guideline for hits-to-leads chemistry. *Chem. Med. Chem.* 3, 1179–1180. doi: 10.1002/cmdc.200800132

Zartler, E. R., and Shapiro, M. J. (eds.). (2008). "Designing a fragment process to fit your needs," in *Fragment-Based Drug Discovery* (John Wiley & Sons, Ltd.), 15–37. doi: 10.1002/97804707215 51.ch2

Zerbe, B. S., Hall, D. R., Vajda, S., Whitty, A., and Kozakov, D. (2012). Relationship between hot spot residues and ligand binding hot spots in protein-protein interfaces. *J. Chem. Inf. Model.* 52, 2236–2244. doi: 10.1021/ci30 0175u

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Vásquez and González Barrios. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 04 September 2019
doi: 10.3389/fchem.2019.00602

Structural, Thermodynamic, and Kinetic Traits of Antiestrogen-Compounds Selectively Targeting the Y537S Mutant Estrogen Receptor α Transcriptional Activity in Breast Cancer Cell Lines

Matic Pavlin[1†], Luca Gelsomino[2†], Ines Barone[2], Angelo Spinello[1], Stefania Catalano[2], Sebastiano Andò[2] and Alessandra Magistrato[1*]*

[1] National Research Council – Institute of Materials (IOM) at International School for Advanced Studies (ISAS), Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy, [2] Department of Pharmacy, Health and Nutrition Sciences, Centro Sanitario, University of Calabria, Rende, Italy

The most frequently diagnosed cancers in women are the estrogen receptor (ER)-positive breast cancer subtypes, which are characterized by estrogen dependency for their growth. The mainstay of clinical treatment for this tumor relies on the modulation of ERα action or on the suppression of estrogen biosynthesis via the administration of Selective ERα Modulators/Down-regulators (SERMs/SERDs) or aromatase inhibitors, respectively. Nevertheless, *de novo* and acquired resistance to these therapies frequently occurs and represents a major clinical concern for patient survival. Recently, somatic mutations affecting the hormone-binding domain of ERα (i.e., Y537S, Y537N, D538G) have been associated with endocrine resistance, disease relapse and increased mortality rates. Hence, devising novel therapies against these ERα isoforms represents a daunting challenge. Here, we identified five molecules active on recurrent Y537S ERα polymorphism by employing *in silico* virtual screening on commercial databases of molecules, complemented by ER-transactivation and MTT assays in MCF7 and MDA-MB-231 breast cancer cells expressing wild type or mutated ERα. Among them, one molecule selectively targets Y537S ERα without inducing any cytotoxicity in breast cell lines. Multi-microseconds (4.5 μs) of biased and unbiased molecular dynamics provided an atomic-level picture of the structural, thermodynamics (i.e., binding free energies) and the kinetic (i.e., dissociation free energy barriers) of these active ligands as compared to clinically used SERM/SERDs upon binding to wild type and distinct ERα variants (Y537S, Y537N, D538G). This study contributes to a dissection of the key molecular traits needed by drug-candidates to hamper the agonist (active)-like conformation of ERα, normally selected by those polymorphic variants. This information can be useful to discover mutant specific drug-candidates, enabling to move a step forward toward tailored approaches for breast cancer treatment.

Keywords: estrogen receptor, breast cancer, SERM, SERD, molecular dynamics, Y537S, resistant breast cancers

OPEN ACCESS

Edited by:
Simone Brogi,
Department of Pharmacy, University of Pisa, Italy

Reviewed by:
Marcus Scotti,
Federal University of Paraíba, Brazil
Vincent Zoete,
Swiss Institute of Bioinformatics (SIB), Switzerland
Ahmed H. E. Hassan,
Kyung Hee University, South Korea
Stevan Armakovic,
University of Novi Sad, Serbia

***Correspondence:**
Sebastiano Andò
sebastiano.ando@unical.it
Alessandra Magistrato
alessandra.magistrato@sissa.it

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *24 June 2019*
Accepted: *15 August 2019*
Published: *04 September 2019*

Citation:
Pavlin M, Gelsomino L, Barone I, Spinello A, Catalano S, Andò S and Magistrato A (2019) Structural, Thermodynamic, and Kinetic Traits of Antiestrogen-Compounds Selectively Targeting the Y537S Mutant Estrogen Receptor α Transcriptional Activity in Breast Cancer Cell Lines. Front. Chem. 7:602. doi: 10.3389/fchem.2019.00602

INTRODUCTION

Breast Cancer (BC) is the most frequent cancer type and the second leading cause of death in women, representing 25% of all cancers. In ∼70% of the BC cases detected after the menopause, cellular proliferation is mediated by estrogens (**1**; **Figure 1**) binding to their specific nuclear hormone receptor [Estrogen Receptor α (ERα, *ESR1*)] (Fanning and Greene, 2019).

This latter is a ligand-activated transcription factor, which upon estrogen binding, decreases apoptosis and/or promotes cell proliferation, ultimately playing a pro-oncogenic role. Hence, in the most diffused BC type cell proliferation relies on the expression of ERα, and on the presence of blood circulating estrogens, being hence classified as ER sensitive (ER+). Gold standard endocrine treatments against ER+ BC consist in suppressing estrogen biosynthesis, via the administration of aromatase inhibitors, or in counteracting ERα pro-oncogenic action via the drugging of selective ERα modulators (SERMs) or downregulators (SERDs). Namely, SERMs [tamoxifen and its most abundant metabolite endoxifen (END)] act as antagonists, occupying the estrogen binding site and inducing a conformational change of ERα toward an inactive conformation. SERDs [fulvestrant (FULV)], instead, also foster ERα ubiquitination and degradation (Fanning et al., 2016; Pavlin et al., 2018).

Similarly to other nuclear receptors, ERα presents a puzzling tridimensional structure which atomic-level organization remains controversial (Huang et al., 2018). This is composed out of five distinct functional domains (**Supplementary Figure 1**), among which only the structures of the DNA-binding domain and the ligand-binding domain (LBD) have been characterized. The LBD is active as a homodimer with each monomer hosting a ligand binding cavity (LBC). The LBD crystal structures (**Supplementary Figure 2**) revealed that upon binding of an agonist or an antagonist molecule, helix 12 (H12) can undergo a conformational switch between the active and inactive form, respectively (Joseph et al., 2016). In the agonist (active) state, H12 faces helices H3, H5/6, and H11, closing the LBC (**Supplementary Figure 2A**; Robinson et al., 2013). Conversely, in the antagonist (inactive) form, H12 rearranges, as a consequence of SERM/SERD-binding, moving toward the groove lined by H3 and H5 (**Supplementary Figure 2B**; van Kruchten et al., 2015; Joseph et al., 2016).

In the last decades, SERMs have been proved to be highly beneficial, substantially decreasing the mortality rates of woman affected by BC cancer type by 25–30 %. The most effective ERα antagonists in clinical use are: (i) tamoxifen (**2**; **Figure 1**), a SERM, which in spite of its beneficial action in breast tissues, is plagued by agonistic effects in peripheral ones (endometrium), and is active through its metabolite, endoxifen (**3**; **Figure 1**), and (ii) FULV (**4**; **Figure 1**), a potent SERD (Nilsson and Gustafsson, 2011; van Kruchten et al., 2015) characterized, however, by poor solubility. This makes its administration arduous and therefore probably limiting its efficacy. These adjuvant therapies are administered over extended time frame (5–10 years) to control tumor growth, or, even, to prevent disease in case of BC-prone

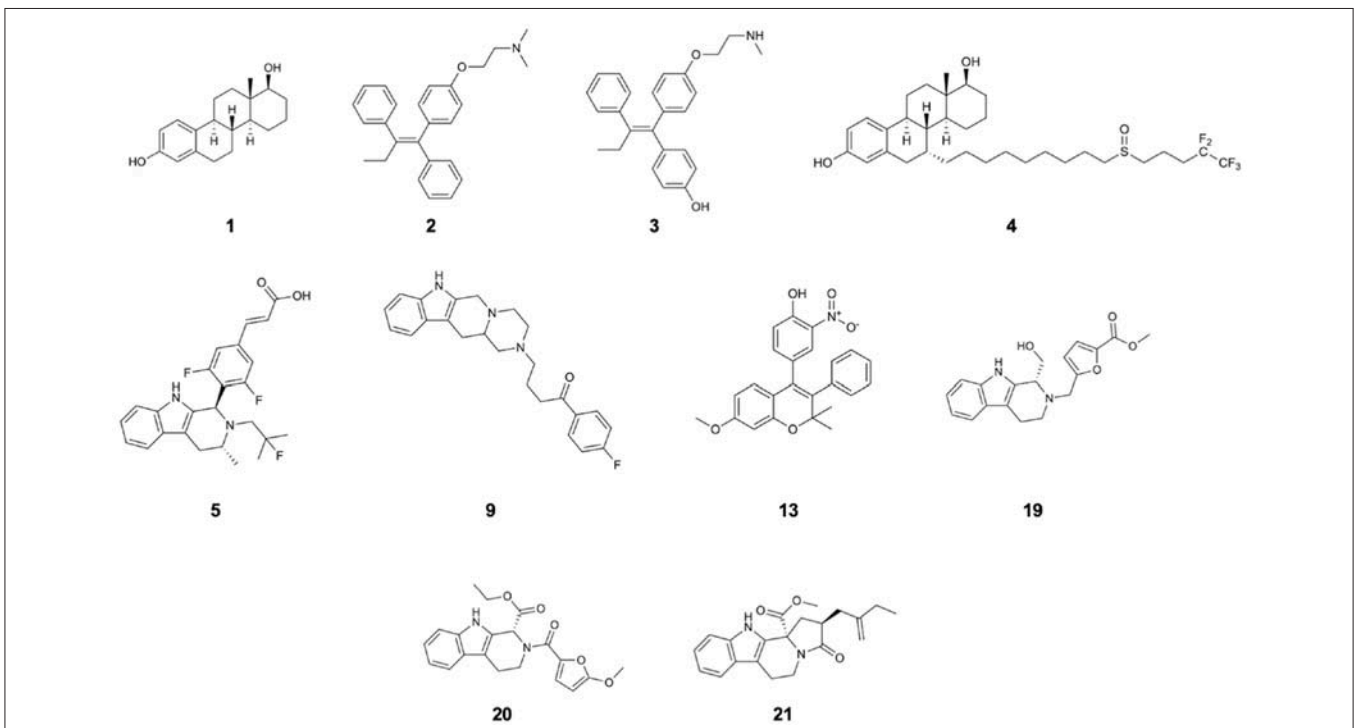


FIGURE 1 | Structures of 17β-estradiol (**1**), tamoxifen (**2**), endoxifen (**3**), fulvestrant (**4**), AZD-9496 (**5**), and compounds **9**, **13**, **19**, **20**, **21** which are the identified drug-candidates active against Y537S ERα.

genetic profiles. Nevertheless, after prolonged exposure to these therapies, tumors evolve by adapting to the pharmacological pressure. Distinct studies highlighted a stunningly complex, composite, and multifactorial genomic landscape as responsible of tumor refractoriness to treatments (Spinello et al., 2019b). This has mostly been associated to an alteration of mitogen activated kinase pathway (MAPK) or of a deregulated estrogen receptor transcriptional activity (Razavi et al., 2018). This latter takes place when ERα acquires new structural traits, eventually leading to resistance and relapse to therapies. Scaringly, almost 50 % of all ER+ BC patients, initially benefiting from first-line therapy, will eventually develop resistance after prolonged treatments (acquired resistance). This ultimately results in a shortening of their survival time.

It is nowadays well-established (Merenbakh-Lamin et al., 2013; Robinson et al., 2013; Toy et al., 2013; Jeselsohn et al., 2014) that distinct ERα polymorphisms (mERαs), located in the vicinity of LBC (i.e., E380Q), between H9 and H10, or in the loop connecting H11 and H12 (i.e., L536Q, L536R, Y537C, Y537N, Y537S, and D538G), are recurrently observed in metastatic BC patients relapsing after extended treatments regimens. The most abundant ERα polymorphisms observed in LBD are D538G (occurrence of 21–36 % of cases (Jeselsohn et al., 2014; Chandarlapaty et al., 2016; Toy et al., 2017), Y537N (5–33 %) (Jeselsohn et al., 2014; Toy et al., 2017), and Y537S (13–22 %) (Jeselsohn et al., 2014; Chandarlapaty et al., 2016; Toy et al., 2017). This latter remains the most aggressive isoform (survival rate of 20 months as compared to 26 of D538G) (Chandarlapaty et al., 2016). Most of these mutations are insensitive to tamoxifen, while still responding to FULV. As such, tumors proliferation still depends on ERα expression, underlying the still unmet oncological need of a complete inhibition/abrogation of its signaling pathway (Busonero et al., 2019).

Targeted therapy counteracting mERα is a current object of intense preclinical and early clinical interest, as also evidenced by the significant number of studies aiming at identifying orally bioavailable SERDs, eventually able to overcome resistance (De Savi et al., 2015; Fanning et al., 2018; Hamilton et al., 2018; Sharma et al., 2018; Kahraman et al., 2019; Scott et al., 2019). Among these GDC-810, AZD9496 (**5**; **Figure 1**) (hereafter AZD, a drug in preclinical use as oral SERD, for which a first clinical study have been recently accomplished) (Weir et al., 2016; Hamilton et al., 2018), and LSZ102 have been identified (Scott et al., 2019).

Aiming at selectively targeting specific and aggressive ERα variants, we have recently meticulously annotated the structural and dynamical alterations induced to ERα structure by each recurrent polymorphism, disclosing that each of them triggers the acquisition of a different agonist-like (intrinsically active) conformation of H12. As a result, tumors bearing these isoforms proliferate irrespectively of estrogen production, with SERMs and aromatase inhibitors' efficacy being lost after their appearance (Spinello et al., 2019b). Since these ER+ BC cells are sensitive to FULV and AZD, we also performed extensive MD simulations typifying the structural features responsible of distinct efficacy of FULV and AZD on mutant (m)ERα, as opposite to END.

Stunningly, our computational assays disclosed that the FULV and AZD's elongated shape, owing to their aliphatic and carboxylic moiety, respectively, was the key structural determinant counteracting the acquisition of an H12 agonist-like conformation (Fanning et al., 2016; Pavlin et al., 2018). Understanding the structural signature needed by drugs able to effectively fight metastatic and refractory BC types was a necessary prerequisite for the rational discovery of mutant-specific SERMs/SERDs. Building on this knowledge, here, we performed virtual screenings on existing database of molecules, and we tested their efficacy on MCF7 and MDA-MB-231 cell lines harboring Wild Type (WT) and Y537S, D538G, and Y537N ERα variants. As a result, we identified five molecules able to counteract the enhanced transcriptional activity of Y537S ERα isoform, with one being non-cytotoxic and preferentially active toward the Y537S variant. The structural and dynamical impact of the five active molecules, as well as their thermodynamic and kinetic properties (of the best molecule) were also explored via biased and unbiased classical Molecular Dynamics (MD) simulations, and compared with those of END, AZD, and FULV in order to dissect the source of their distinct efficacy profiles. These outcomes may lead toward the discovery of isoform-selective drug-candidates, providing a therapeutic option for the specific genomic profile of ER+ BC patients relapsing mainstay therapies.

MATERIALS AND METHODS

In silico Screening

The NCI (https://cactus.nci.nih.gov/download/nci/) library Release 4 Files Series—May 2012 (containing 265,242 structures) was used for virtual screening (VS) studies. Compounds were filtered using the Schrodinger Suite 2017-1 Ligfilter tool (2017). In order to eliminate molecules possessing poor absorption and permeation we applied Lipinski's rule of five (Lipinski et al., 2001). Further, filtering was applied to compounds bearing more than 10 rotatable bonds, since high ligand flexibility implies higher entropic contributions and reduces oral availability (Veber et al., 2002). Next, QikProp (2017) was employed to predict LogP values of the compounds to assess information on their solubility in water.

After ligand preparation *in silico* screening of the library was performed on different mERα structures. Namely, in order to account for receptor's flexibility in the screening we considered five different ERα conformations as target structures (Pavlin et al., 2018). These were selected from the populated cluster extracted from 500 ns-long classical MD simulations trajectories of AZD and FULV in complex with the Y537S, Y537N, and D538G isoforms obtained in our previous study (Pavlin et al., 2018). In this respect, we employed two conformations for Y537S ERα (in complex with AZD and FULV), one for Y537N (in complex with FULV, since conformation of this mutant complex with AZD was similar to that of Y537S), and two structures for D538G (in complex with AZD and FULV). A van der Waals (vdW) radius scaling factor of 0.80 Å for protein and ligands atoms having a partial charge < 0.15 was used to account for protein

flexibility. Size of the box for *in silico* screenings was determined by considering the residues interacting with different antagonists placed inside the binding site and those residues pinpointed as critical for antagonizing ERα activation in our previous work (Pavlin et al., 2018).

In order to obtain a set of promising ligands for experimental testing, we followed two protocols of VS. First, a workflow based on three subsequent steps of docking with increasing level of accuracy for each ERα conformation, was adopted by using the Glide program (Friesner et al., 2004). Namely, (i) a fast high-throughput virtual screening (HTVS) was initially performed in order to efficiently select promising ligands among ∼220,000 of compounds from the pre-filtered NCI library; (ii) 10 % of the best ranked ligands (∼22,000 structures per each ERα conformation) were retained and a single precision (SP) docking calculation was done; (iii) the top 10 % of the resulting compounds (∼2,200 structures per each ERα conformation) were screened using the extra precision (XP) protocol. This latter should eliminate false positives by using a more extensive sampling and more accurate scoring functions. END, AZD, and FULV were also docked to assess the quality of our results as reported in **Supplementary Table 1**. The molecules resulting from the screening were sorted according to GlideScore scoring function. The selection criterion for further investigation was that the screened compounds had docking score lower than −8.5 kcal/mol and that displayed favorable interactions with at least one of the five mERαs target structures (i.e., two structures extracted from the MD trajectory of Y537S ERα in complex with AZD and FULV, one for Y537N in complex with FULV, and two structures for D538G in complex with AZD and FULV). This was done in order to find a good compromise between the number of molecules selected for experimental screening and the quality of the docking score. Moreover, our reference molecules FULV and AZD exhibit on the target structure the same range of docking score values.

In the second protocol, we initially performed ligand-based screening using the CANVAS program (Duan et al., 2010; Sastry et al., 2010). Here the searching criteria were based on the scaffold that antagonist should possess. The latter was defined considering the common structural features that an effective SERD should have according to our previous study (Pavlin et al., 2018), more precisely, selected ligand should have scaffold based either on the END scaffold or on the tri-membered ring scaffold of AZD in order to stabilize ligand inside LBC, together with a polar tail that is able to form hydrogen bonds (H-bonds) with the H11-12 loop. The 415 selected ligands were then screened to all five mERαs conformations by using XP protocol. In this second case, the cut-off docking score for selection of ligands was −7.5 kcal/mol, and the ligands were selected only when displaying favorable interactions with at least two distinct mERα structures among the five target structures used and at least one was within the cut-off range. These molecules were available as donation of the National Cancer Institute USA. Other known activities of the five molecules observed to be active on Y537S in this study are reported in **Supplementary Table 7**.

Reagents, Antibodies, and Plasmids

17β-estradiol was purchased from Sigma (St. Louis, MO, USA). Antibodies against ERα and GAPDH were from Santa Cruz Biotechnology (Santa Cruz, CA, USA). Yellow fluorescent protein (YFP)-tagged expression constructs, YFP-WT, YFP-Y537S, YFP-Y537N and YFP-D538G ERα were generated as previously described (Gelsomino et al., 2016, 2018) XETL plasmid, containing an estrogen-responsive element, was provided by Dr. Picard (University of Geneva, Geneva, Switzerland).

Cell Cultures

Human MCF-7 and MDA-MB-231 BC cells were acquired in 2015 from American Type Culture Collection, stored and cultured according to supplier's instructions. Cells were used within six-months after frozen-aliquot resuscitations and regularly tested for Mycoplasma-negativity (MycoAlert, Lonza, Basilea, Switzerland).

Immunoblot Analysis

Equal amounts of proteins were resolved on 10% SDS-PAGE as previously described (Giordano et al., 2016). The antigen-antibody complex was revealed using the ECL System (Bio-rad, Hercules, CA, USA). Images were acquired using Odissey FC from Licor (Lincoln, Nebraska, USA). Blots are representative of three independent experiments.

ERα Transactivation Assay

ERα transactivation assay was performed as previously reported (Barone et al., 2011). Briefly, MCF-7 and MDA-MB-231 cells (50,000/well) were plated in phenol red-free with 5 % charcoal-stripped FBS in 24-well plates. After 24 h, cells were co-transfected with 0.5 μg of reporter plasmid XETL plus 0.1 μg of YFP-tagged expression constructs and 20 ng of TK Renilla luciferase plasmid as an internal control. Transfection was performed using the Lipofectamine 2000 reagent (Life Technologies, Carlsbad, CA, USA) as recommended by the manufacturer. Six hours after transfection, the medium was changed and the cells were treated as indicated for 24 h. Firefly and Renilla luciferase activities were measured using a Dual Luciferase kit (Promega, Madison, WI, USA). The firefly luciferase data for each sample were normalized on the basis of transfection efficiency measured by Renilla luciferase activity (Rizza et al., 2014). Data represent three independent experiments, carried out in triplicate.

MTT Cell Viability Assay

1,000 cells were plated into 96-well plates in phenol red-free medium containing 5 % charcoal-stripped FBS. After 24 h, cells were exposed to the different treatments as indicated. One day later, cell viability was assessed by (3-(4,5-Dimethylthiazol-2-yl)-2,5-Diphenyltetrazolium-Bromide) (MTT, Sigma-Aldrich) as described (Covington et al., 2013). Results are expressed as fold change relative to vehicle-treated cells. Data represent three-independent experiments, performed in triplicate.

Real-Time RT-PCR Assays

Total RNA was extracted from cells using TRIzol reagent (Life Technologies). Purity and integrity of the RNA were confirmed spectroscopically and by gel electrophoresis before use. One microgram of total RNA was reverse transcribed in a final volume of 20 μL using the RETROscript kit (Life Technologies) and cDNA was diluted 1:3 in nuclease-free water. The evaluation of *TFF1*, *CTSD*, *CCND1* and *MYC* mRNA expression was performed by real-time RT-PCR, using SYBR Green Universal PCR Master Mix (Bio-rad). The relative gene expression levels were calculated using the ΔΔCt method as described (Catalano et al., 2015). Primers are listed in **Supplementary Table 2**.

Statistical Analysis of Experimental Data

Data were analyzed for statistical significance using two-tailed student's Test using GraphPad-Prism5 (GraphPad Software, Inc., San Diego, CA). Standard deviations (S. D.) are shown.

Classical MD Simulations

Of the 17 tested molecules (**Figure 1**, **Supplementary Figure 3**) obtained by both protocols of VS (four from the structure-based strategy and the rest by the ligand-based strategy as reported in **Supplementary Table 1**) five resulted to be active in the *in vitro* tests. These latter were also docked to the WT ERα using the XP protocol. All active molecules were subjected to MD simulations in complex with the WT and Y537S ERα variants. Additionally, compound **19**, showing the most promising results in experimental tests, was also docked and simulated in complex with Y537N and D538G ERα.

Physiological protonation states of ERα were already determined previously (Pavlin et al., 2018) using the webserver H++ (Anandakrishnan et al., 2012). Parm99SB AMBER force field (FF) with ILDN modification was employed for the protein (Wickstrom et al., 2009; Lindorff-Larsen et al., 2010), and the general Amber FF (GAFF) (Wang et al., 2004) was used for ligands. ESP charges (Bayly et al., 1993) were calculated by performing geometry optimizations of the ligands at Hartree-Fock level of theory using a 6-31G* basis set with the Gaussian 09 software (Frisch et al., 2016) and were later transformed in RESP charges with the Antechamber module of Ambertools16 (Wang et al., 2006). Since the dockings were performed on monomers, while the LBD in physiological conditions is a dimer, we built each dimer by superimposing the monomer on each of the two dimers of the corresponding mERα conformation from our previous work (Pavlin et al., 2018).

Each system was solvated using TIP3P waters (Jorgensen et al., 1983) in a truncated octahedron box with minimum distance of 12 Å between solute and the edge of the box, leading to a total of ∼95,000 atoms. MD simulations were performed with

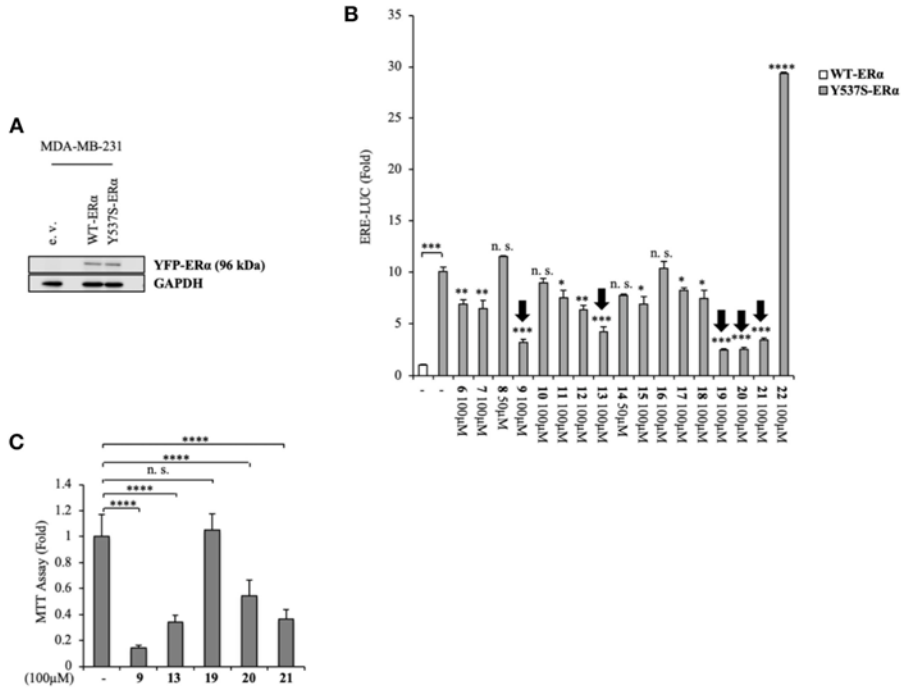


FIGURE 2 | Inhibition of Y537S ERα activation exerted by the compound **19** in MDA-MB-231 breast cancer cells. **(A)** Immunoblotting showing exogenous ERα protein expression in ERα-negative MDA-MB-231 breast cancer cells transiently transfected with an empty vector (e. v.), YFP-WT or YFP-Y537S ERα expression vectors. GAPDH was used as a control for equal loading and transfer. **(B)** ERα-transactivation assay in MDA-MB-231 cells transiently transfected with YFP-WT or YFP-Y537S ERα expressing vectors plus an ERE-luciferase reporter (XETL), and treated with vehicle (-) or the different compounds (**6**-**22**), as indicated. Data are reported as fold change relative to WT-ERα expressing cells. **(C)** MTT cell viability assay in MDA-MB-231 cells treated with vehicle (-) or the different compounds (**9**, **13**, **19**, **20**, **21**, 100 μM), as indicated for 24 h. Results are expressed as fold change relative to vehicle-treated cells. The values represent the mean ± S. D. of three different experiments, each performed in triplicate. n. s., non-significant; *$P < 0.05$; **$P < 0.005$; ***$P < 0.0005$; ****$P < 0.00005$.

GROMACS 5.0.4 (Abraham et al., 2015). An integration time step of 2 fs was used and all covalent bonds involving hydrogen atoms constrained with the LINCS algorithm (Hess et al., 1997). Particle Mesh Ewald algorithm (Darden et al., 1993) was used in order to account electrostatic interactions. Simulations were performed in the isothermal-isobaric NPT ensemble, at a temperature of 310 K, under control of a velocity-rescaling thermostat (Bussi et al., 2007). Preliminary energy minimization was done with the steepest descend algorithm. Next, all systems were heated to the final temperature of 310 K using 40 steps of simulated annealing (0-90 K in steps of 5 K/25 ps; 90-310 K in steps of 10 K/25 ps). WT ERα models underwent 300 ns long classical MD simulations (last 200 ns were used for analysis), while mERα ones underwent 400 ns long simulations and last 300 ns were used for further analysis.

Metadynamics

In order to further refine the binding poses and better dissect the impact of the kinetic properties on efficacy and selectivity, we performed FF-based Metadynamics (MTD) simulations of AZD and **19**. In particular, MTD runs of 60–130 ns were done to refine the binding pose and study ligand dissociation with GROMACS 5.0.4 using the PLUMED 2.0 plugin (Tribello et al., 2014). Two collective variables (CVs) were used: the first CV (CV1) describes the number of either hydrogen bonds (AZD) or hydrophobic contacts (**19**) between the ligands and the LBC, computed as a coordination number; the second CV (CV2) corresponds to the distance between the center of masses (COM) of the protein and the ligand. Gaussian hills having a height of 0.6 kJ/mol in all systems, while the widths of 0.06 and 0.015 (AZD), 0.40 and 0.025 (**19**), were added, respectively, for CV1 and CV2 every 4 ps of MD. A harmonic wall was used to restrain the exploration of the FES on CV2 at the value of 3.5 nm. Three replicas of the MTD simulations were performed, starting from different frames as extracted from the equilibrated MD trajectory and the uncertainty of the dissociation free energy barriers ($\Delta G_b^{\#}$) were estimated from the standard deviation of the barriers obtained out of the three replicas, following a protocol adopted in previous studies (Bisha et al., 2013; Sgrignani and Magistrato, 2015; Spinello et al., 2018, 2019a).

Simulation Analysis

Cluster analysis and root mean square deviation (RMSD) of the MD trajectories were done with the g_cluster tool, based on the Daura et al algorithm (Daura et al., 1999), and g_rms, as implemented in the GROMACS 5.0.4 program (Abraham

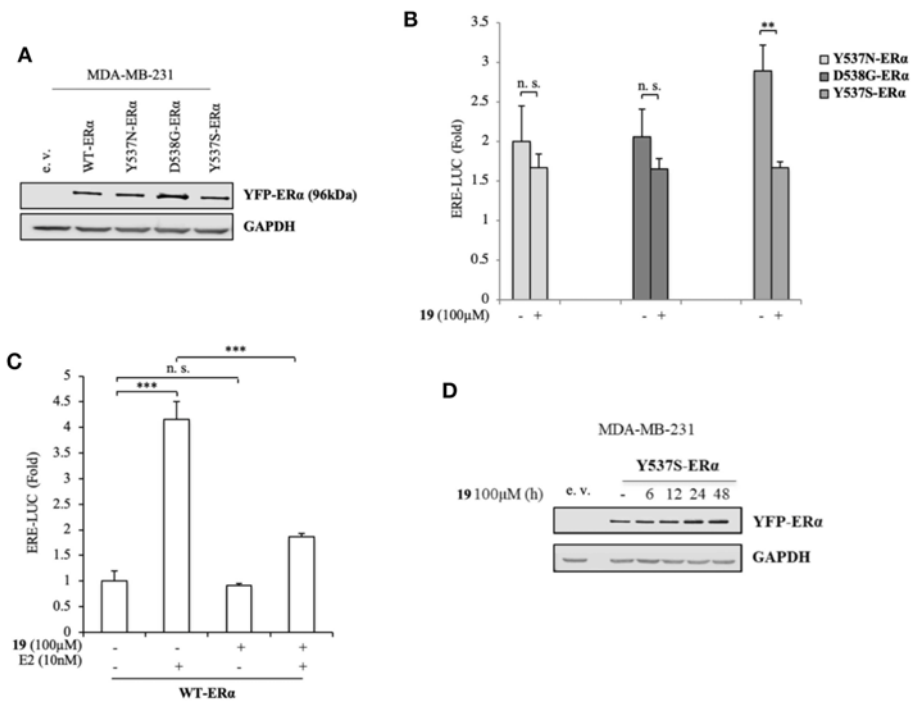


FIGURE 3 | Impact of compound **19** on activation of WT, Y537N, and D538G ERα in MDA-MB-231 breast cancer cells. **(A)** Immunoblotting showing exogenous ERα protein expression in ERα-negative MDA-MB-231 breast cancer cells transiently transfected with an empty vector (e. v.), YFP-WT, YFP-Y537N, YFP-Y537S, or YFP-D538G ERα expressing vectors. GAPDH was used as a control for equal loading and transfer. **(B)** ERα-transactivation assay in MDA-MB-231 cells transiently transfected with YFP-Y537N, YFP-Y537S, or YFP-D538G ERα expressing vectors plus an ERE-luciferase reporter (XETL), and treated with vehicle (-) or the compound **19** at 100 μM. Data are reported as fold change relative to WT-ERα expressing cells. **(C)** ERα-transactivation assay in MDA-MB-231 cells transiently transfected with YFP-WT plus XETL plasmid, and treated with vehicle (-) or the compound **19** (100 μM) in the presence or not of 17β-estradiol (E2, 10 nM). Data are reported as fold change relative to vehicle (-)-treated cells. The values represent the mean ± S. D. of three different experiments, each performed in triplicate. n. s., non-significant; $^{**}P < 0.005$; $^{***}P < 0.0005$. **(D)** Immunoblotting showing ERα protein expression in MDA-MB-231 breast cancer cells transiently transfected with YFP-Y537S ERα expressing vector and treated with the compound **19** (100 μM) at the indicated time. GAPDH was used as a control for equal loading and transfer.

et al., 2015). Molecular Mechanics Generalized Born Surface Area (MM-GBSA) free energy calculation were performed with the MM_PBSA.py tool of Amber 18 program, following a procedure successfully applied in previous studies (Spinello et al., 2019a,c). Visualization of the MD trajectories was done with the VMD program (Humphrey et al., 1996), while the images were prepared using UCSF Chimera1.12 visualization tool (Pettersen et al., 2004).

Correlation Analysis

The covariance matrices were constructed from the atoms position vectors upon an RMS-fit to the starting configuration of the MD run as to remove the rotational and translational motions. Each element in the covariance matrix is the covariance between atoms i and j, defining the i, j position of the matrix. The covariance C_{ij} is defined as

$$C_{ij} = \left\langle \left(\vec{r}_i - \left\langle \vec{r}_i \right\rangle \right) \left(\vec{r}_j - \left\langle \vec{r}_j \right\rangle \right) \right\rangle, \quad (1)$$

where $\vec{r}_i$ and $\vec{r}_j$ are the position vectors of atoms i and j, and the brackets denote an average over the sampled time period. The diagonalization of the covariance matrix leads to a set of orthogonal collective eigenvectors, each associated to a corresponding eigenvalue. The eigenvalues denote how much each eigenvector is representative of the system dynamics.

The cross-correlation matrices (or normalized covariance matrices) based on the Pearson's correlation coefficients (CC_{ij}) were calculated with the *cpptraj* module of Ambertools 18 from the calculated covariance matrices. Each element of the cross-correlation matrix in the i,j position corresponds to a Pearson's CC_{ij}, i.e., the normalized covariance between atoms i and j calculated with the formula:

$$C_{ij} = \frac{\left\langle \left(\vec{r}_i - \left\langle \vec{r}_i \right\rangle \right) \left(\vec{r}_j - \left\langle \vec{r}_j \right\rangle \right) \right\rangle}{\left[\left(\left\langle \vec{r}_i^{\,2} \right\rangle - \left\langle \vec{r}_i \right\rangle^2 \right) \left(\left\langle \vec{r}_j^{\,2} \right\rangle - \left\langle \vec{r}_j \right\rangle^2 \right) \right]}, \quad (2)$$

here the normalization factor is the product between the standard deviations of the two position vectors. As a result, CC_{ij} range from a value of -1, for a totally negatively correlated motion between two atoms, and a value of $+1$, which instead means a positively correlated lockstep motion. Here we have also calculated the correlation scores (CSs) between each LBD helix and all the others, dividing each as depicted

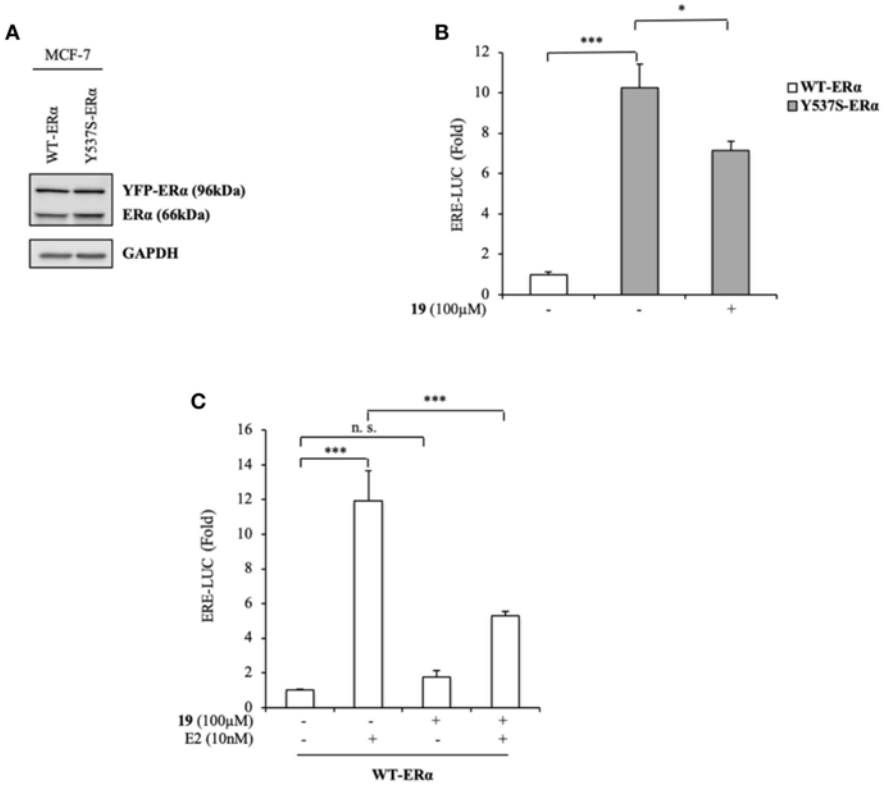


FIGURE 4 | Effects of compound **19** on WT and Y537S ERα activation in MCF-7 breast cancer cells. **(A)** Immunoblotting showing exogenous and endogenous ERα protein expression in ERα-positive MCF-7 breast cancer cells transiently transfected with YFP-WT or YFP-Y537S ERα expressing vectors. GAPDH was used as a control for equal loading and transfer. **(B)** ERα-transactivation assay in MCF-7 cells transiently transfected with YFP-WT or YFP-Y537S ERα expressing vectors plus an ERE-luciferase reporter (XETL), and treated with vehicle (-) or the compound **19** at 100 μM, as indicated. Data are reported as fold change relative to WT-ERα expressing cells. **(C)** ERα-transactivation assay in MCF-7 cells transiently transfected with YFP-WT plus XETL plasmid, and treated with vehicle (-) or the compound **19** (100 μM) in the presence or not of 17β-estradiol (E2, 10 nM). Data are reported as fold change relative to vehicle (-)-treated cells. The values represent the mean ± S. D. of three different experiments, each performed in triplicate. n. s., non-significant; $^{*}P < 0.05$; $^{***}P < 0.0005$.

in **Supplementary Figure 2** (Pavlin et al., 2018). Then, we calculated the sum of CC_{ij} between the residues i belonging to the helix I and the residues j belonging to the helix J. Importantly, the values $-0.6 < CC_{ij} < +0.6$ were discarded in order to eliminate the noise due to uncorrelated motions (Palermo et al., 2016; Casalino et al., 2018) and the sum of the cross-correlation score was divided by the product of the number of residues contributing to the score as a correlation density.

RESULTS

In silico Screening and *in vitro* Studies

Building on our previous classification of structural traits of drugs effectively targeting mERα (Pavlin et al., 2018), we performed *in silico* screening on the structures obtained from MD simulations of Y357S, Y537N, and D538G mutants hosting AZD and FULV in the LBC. A detailed list of the molecules tested as well as their docking score on each specific target structure is reported in **Supplementary Figure 3** and **Supplementary Table 1**, respectively. Remarkably our newly developed VS strategy allows not only to account for the distinct conformations that the receptor can adopt at finite temperature, as in the ensemble docking, but it also encompasses the induced fit effects exerted by the binding of efficacious drugs to distinct ERα isoforms (Spinello et al., 2019a). The best-ranked 17 compounds (**6**-**22**), that were binding to more than one ERα isoform in VS, were then experimentally tested.

Namely, their effect on the transcriptional activity of the mutant Y537S ERα was investigated in cell-based assays, using a standard genomic transcriptional output method (i.e., the estrogen response element (ERE) - luciferase-based gene transactivation system) for assessing their ability to bind ERα and, subsequently, transactivate an ERE-mediated transcription, allowing an assessment of the transcriptional responses of each receptor separately. Thus, human ERα-negative MDA-MB-231 BC cells were cotransfected with either YFP-WT or YFP-Y537S ERα expression vectors along with an ERE-luciferase reporter plasmid (XETL) and treated with the vehicle or the selected compounds (**6**-**22**). As shown in **Figure 2A**, cells expressed similar levels of the 96 kDa protein representing the exogenously added WT or Y537S mutant receptor tagged with YFP. In line with previous results (Toy et al., 2013), reporter gene transactivation assays showed that control basal activity of Y537S ERα was more elevated than that of WT (**Figure 2B**). Importantly, the tested compounds exerted different effects on Y537S ERα transcriptional activity, with **9**, **13**, **19**, **20**, and **21** showing the highest efficacy in reducing the activity of the Y537S mutant (76–57 % decrease) when used at 100 μM concentration. Hence, these were chosen to evaluate their potential toxicity in MDA-MB-231 cells by using MTT cell survival assay (**Figure 2C**). As a result, compounds **9**, **13**, **20**, and **21** markedly reduced cell viability even in MDA-MB-231 cells, whereas the compound **19** did not provoke any significant effects at the dose tested. Thus, among these compounds, **19** represents the best-candidate for further studies. Among these compounds, **9**, **19**, **20**, and **21** share the same chemo-type of the parent compounds AZD.

MDA-MB-231 BC cells transfected with Y537S-ERα vectors and treated with compound **19** at increasing doses (from 1 nM to 100 μM) displayed a dose-dependent decrease of Y537S-ERα transactivation, with the highest inhibition registered at 100 μM concentration (65 ± 10 % inhibition as compared to vehicle) (**Supplementary Figure 4**).

To better clarify the activity of compound **19**, we also evaluated its ability to affect ERα transactivation in cells expressing other two frequently-occurring mutations: YFP-Y537N and YFP-D538G ERα mutations (**Figure 3A**). Surprisingly, **19** has a smaller effect in hampering the transactivation of these mutants (**Figure 3B**), stunningly pinpointing its selectivity in antagonizing preferentially the transcriptional activity of the Y537S ERα isoform.

Next, we inspected its effects on YFP-WT ERα expressing cells in the presence/absence of 17β-estradiol (E2), the endogenous ERα ligand (**Figure 3C**). As expected, E2 treatment was able to trigger luciferase expression through the ERE interaction. Notably its treatment with compound **19**, while not significantly altering WT-ERα transactivation, was associated with a drastic reduction in E2-mediated effects. This suggests that ligand **19** may compete with E2 for the LBC. Interestingly, treatment with **19** was not associated with a down-regulation of Y537S ERα levels (**Figure 3D**).

To expand our investigation, the potency of the compound **19** in affecting Y537S ERα activity was also tested in ER+ MCF-7 BC cells bearing the YFP-WT and YFP-Y537S receptor. These expressed a 66 kDa endogenous ERα, along with a 96 kDa receptor represent the exogenously added WT and mERα tagged with YFP (**Figure 4A**). As previously shown for MDA-MB-231 BC cells, we found a significant increase of YFP-Y537S receptor transcriptional activity as compared to that of YFP-WT ERα and this induction was reduced upon exposure to compound **19** (**Figure 4B**). In addition, **19** antagonized

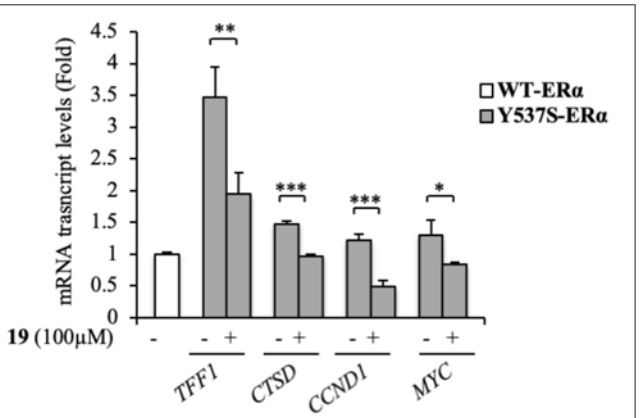


FIGURE 5 | Effects of compound **19** on ERα target genes in YFP-Y537S ERα expressing MDA-MB-231 breast cancer cells. Real time RT-PCR assay for Trefoil Factor 1 (*TFF1*), cathepsin D (*CTSD*), cyclin D1 (*CCND1*) and c-Myc (*MYC*) mRNA expression. The values represent the mean ± S. D. of three different experiments, each performed in triplicate. n. s., non-significant; $^{*}P < 0.05$; $^{**}P < 0.005$; $^{***}P < 0.0005$.

E2-mediated effects also in YFP-WT ERα MCF-7 expressing cells, without exerting any action on basal WT ERα activity (**Figure 4C**).

At a molecular level, ERα activation and association with the ERE result in an enhanced expression profiles of a number of downstream target genes, including those for trefoil factor 1/pS2, cathepsin D, cyclin D1, and c-Myc (Barone et al., 2010). The biological correlation of the inhibition of Y537S ERα transactivation induced by **19** is the down-regulation of the classical estrogen-regulated genes in MDA-MB-231 cells (**Figure 5**), confirming the binding of this antagonist to ERα, well-fitting with *in silico* predictions.

Atomic-Level Understanding of Drugs Efficacy

In order to identify the structural and dynamics features responsible of the efficacy and the selectivity of compound **19** toward Y537S ERα, while being inefficacious and/or displaying limited efficacy on WT, D538G and Y357N, we performed extensive MD simulations of the five active molecules in complex with the WT and Y537S ERα isoforms, starting from binding poses obtained from docking simulations.

MD simulations revealed two important and common structural traits among the inspected compounds, also shared by AZD and FULV. All molecules occupy the binding cavity

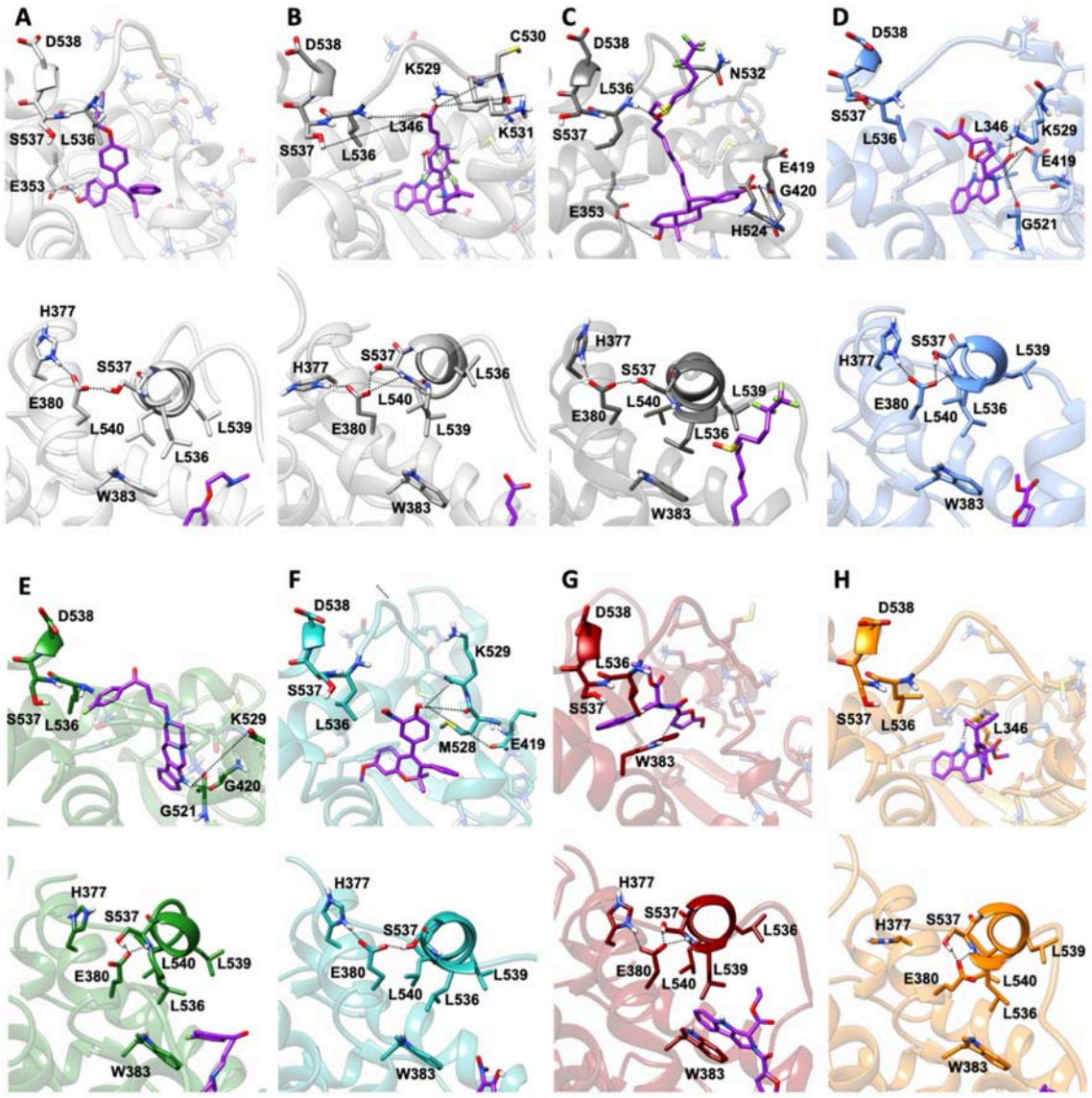


FIGURE 6 | Binding of the five active molecules to Y537S ERα as compared with END, AZD, and FULV. **(A)** END; **(B)** AZD; **(C)** FULV; **(D)** compound **19**; **(E) 9**; **(F) 13**; **(G) 20**; **(H) 21**. Top panels show their placement in the ligand binding cavity, while bottom panels display a close view of E380 H-bond network induced by each ligand. Inhibitors are shown in licorice with carbon atoms in purple color, while oxygen and nitrogen are shown in red and blue, respectively. Protein is depicted in gray new cartoons for END, AZD, FULV, and blue, green, light green purple orange new cartoons for **19**, **9**, **13**, **20**, **21**, respectively.

protruding toward the H11-12 loop, which hosts the Y537S variant (**Figure 6**). Three of them (**9**, **19, 20**) establish π-π interactions with W386 in WT (**Supplementary Figure 5**). Due to their different shapes, each ligand engages distinct H-bonds patterns (**Supplementary Table 3**). This network in compounds **13** and **20** involves residues G521, M528, and C530, while **9** and **21** persistently H-bond either to L346, similarly to AZD, or to E419 and G420 (**Supplementary Table 3**). These results show that the selected compounds can either bind in LBC (**9** and **21**) or interact with H11-12 loop (**13** and **20**).

Conversely, **19** is the only ligand firmly anchored to E419 and L346 (**Supplementary Table 3**), at tract H-bonding to K529, similarly to AZD in complex with to Y537S (**Figure 6**). This H-bonding motif in our previous paper was indicated as an essential signature of drug-efficacy. Nevertheless, **19** forms a set of low-persistent H-bonds, underlying its high mobility and the need for further optimization in order to improve its efficacy. Surprisingly, compound **19** establishes a well-defined and stable H-bond network only in one LBCs of WT ERα (**Supplementary Table 3**).

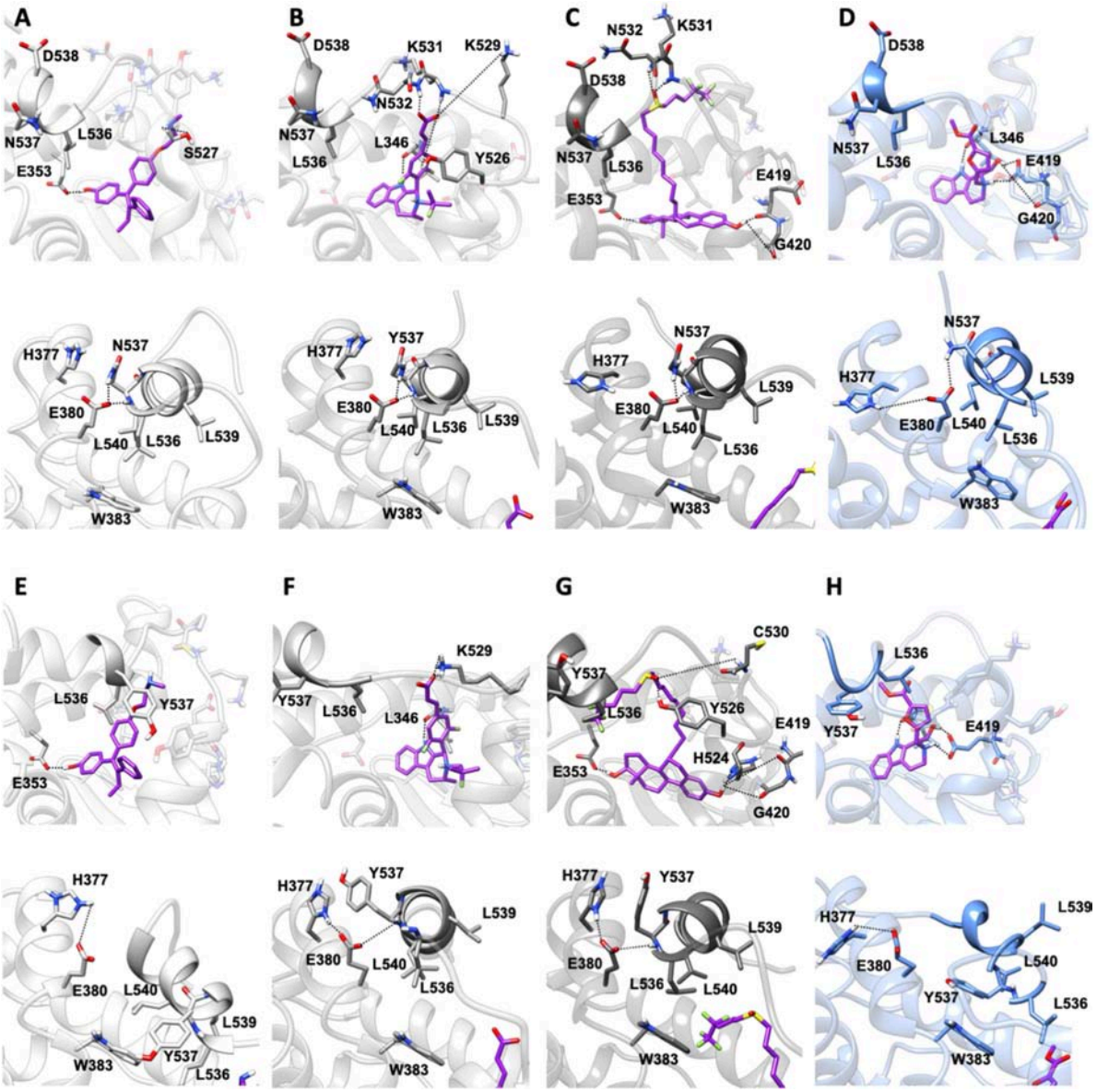


FIGURE 7 | Binding of compound **19** to Y537N and D538G ERα as compared with END, AZD and FULV. Binding to Y537N: **(A)** END; **(B)** AZD; **(C)** FULV; **(D)** Compound **19**. Binding to D538G: **(E)** END; **(F)** AZD; **(G)** FULV; **(H)** Compound **19**. Top panels show their placement in the ligand binding cavity, while bottom panels display a close view of E380 H-bond network induced by each ligand. Inhibitors are shown in licorice with carbon atoms in purple color, while oxygen and nitrogen in red and blue, respectively. Protein is shown in gray new cartoons for END, AZD, FULV, and in blue new cartoons for **19**, respectively.

Next, we inspected how the active drugs counteract the H-bond network responsible of the ERα agonist-like conformation induced by mERα (**Supplementary Table 4**, **Figure 6**). A decrease of the E380-Y537S interaction, previously indicated as structural signature of an intrinsic ERα activation, occurs with all ligands, even if this is less effective than upon FULV or AZD binding. What is more, E380, which strongly H-bonds to S537 in the aggressive Y537S ERα variant, upon binding of **9**, **19**, and **21,** rearranges and engages persistent H-bond to L536. Additionally, in the presence of **19**, there is significant change in the H-bond network of L536 backbone. While this latter strongly interacts with backbone of L539 in the presence of all other compounds, **19** weakens it and, as a result, L536 H-bonds to the backbone of L540 (**Supplementary Table 4**). Remarkably, **19** triggers formation of these H-bonds only in Y537S, but not in WT ERα.

In both Y537N and D538G variants **19** establishes a H-bond network in the binding cavity and in the H11-12 loop region similarly to Y537S (**Supplementary Tables 3**, **4**, and **Figure 7**). Hence, **19** exclusively H-bonds to E419 in all mutants, while only in Y537S can establish week H-bond to K529, similarly to AZD.

We have also calculated the binding free energy (ΔG_b) (**Table 1**) of the five active compounds to WT, Y537S and, for compound **19,** also to the Y537N, D538G variants. Stunningly, **19** dissociates from one monomer of WT ERα due to the lack of H-bonds, rationalizing its preference toward the pathogenic variants. Instead, its ΔG_b is similar in all tested mutants. All other active ligands, but **20**, display a slightly higher affinity for WT ERα, and their ΔG_b is slightly larger than that of **19** toward Y537S, most probably because of their larger size. Nonetheless, the tested ligands do not strongly bind to the LBC, as shown by a comparison of the calculated ΔG_b compared to that AZD. Thus, even small differences in the position of their H-bonding moieties may result in the different binding poses observed for **20** and **13**. To monitor the impact of the distinct ligands size on ΔG_b, we also computed the ligand efficiency (LE, **Supplementary Table 5**), calculated as ΔG_b divided by the number of non-hydrogen atoms. LE differences among ligands are smaller than that of ΔG_bs. **19** presents comparable LE for all mutants tested. Interestingly, compounds **9**, **20**, **21** have a slightly larger LE for Y537S than **19**, suggesting that other features, besides LE or ΔG_b, may be important for ligand selectivity toward the distinct ERα isoforms.

Structural Signatures of (m)ERα Activation/Inactivation

The cross-correlation matrix calculated on the basis of the Pearson correlation coefficients (CC_{ij}) was computed to qualitatively identify the linearly coupled motions between couples of residues along the MD trajectory. A simplified version of this matrix, based on the sum of the of correlation scores (CSs) between each structural elements of (m)ERα (**Supplementary Figure 2**), has been calculated to decrypt the correlation pattern in complex systems (Casalino et al., 2018), among which ERα (Pavlin et al., 2018). In this analysis, a positive/negative score corresponds to a positively (correlated) / negatively (anti-correlated) motion.

In our previous study, the presence of a positive correlation score between H12 and H3-H5 was taken as a structural signature of Y537S ERα intrinsic activation. This was persistent upon END binding, while only FULV and, partially, AZD were able to remove it, in line with the proved activity of these SERDs on the Y537S mutant (Fanning et al., 2016). Hence, we also inspected if the ligands differently affect the internal cross-correlation map. All compounds binding to Y537S remove the contacts between H12 and H3, reducing, in most cases, the cross-correlation score in both monomers

TABLE 1 | Binding free energies (ΔG_b, kcal/mol) of the ligands **9**, **13**, **19**, **20**, and **21**.

	WT		Y537S		Y537N		D538G	
	A	B	A	B	A	B	A	B
Endoxifen	−51.7 ± 3.6	−46.9 ± 3.3	−45.6 ± 3.5	−45.2 ± 3.5	−43.8 ± 3.3	−49.9 ± 4.3	−47.1 ± 3.4	−50.5 ± 3.6
AZD−9496	−45.2 ± 3.6	−43.9 ± 3.7	−42.9 ± 3.7	−53.7 ± 7.7	−45.1 ± 4.1	−42.9 ± 3.9	−45.1 ± 4.0	−43.7 ± 4.3
Fulvestrant	−69.0 ± 6.4	−66.9 ± 6.2	−65.1 ± 5.1	−76.0 ± 5.6	−68.2 ± 6.1	−66.6 ± 5.7	−63.5 ± 6.0	−69.7 ± 6.0
9	−42.7 ± 4.2	−36.7 ± 4.4	−36.9 ± 5.7	−37.3 ± 4.6	/	/	/	/
13	−44.8 ± 4.3	−42.6 ± 3.3	−41.4 ± 3.2	−45.2 ± 4.6	/	/	/	/
19	−21.3 ± 7.2*	−40.9 ± 8.1	−32.3 ± 5.8	−32.3 ± 6.2	−32.9 ± 3.6	−43.0 ± 6.7	−30.8 ± 5.7	−46.8 ± 6.0
20	−38.6 ± 3.3	−35.1 ± 2.6	−40.0 ± 2.9	−38.9 ± 5.0	/	/	/	/
21	−45.4 ± 3.3	−42.5 ± 3.8	−43.6 ± 3.2	−39.2 ± 3.0	/	/	/	/

Absolute values of molecular mechanics Generalized bond surface are (MM-GBSA).‡

‡Number of atoms/heavy atoms in each ligand:

Endoxifen – 55/28.

AZD-9496 – 57/32.

Fulvestrant – 88/41.

***9** – 56/29.*

***13** – 51/30.*

***19** – 46/25.*

***20** – 47/27.*

***21** – 50/26.*

**Ligand exits from the binding pocket.*

(**Supplementary Figure 6**). Moreover, compound **19** decreases these contacts also when binding to Y537N and, to a minor extent, to D538G (**Supplementary Figures 7**, **8**). Conversely, in WT ERα a smaller positive correlation among H5 and H12 can be observed only for (**20** and **21**; **Supplementary Figure 9**). In order to capture more quantitatively the relative differences among the activity exerted by these ligands we also analyzed how H12 correlates with all other ERα structural elements in the presence of the distinct active compounds. This analysis clearly shows that ligands **9** and **20** present a cross-correlation

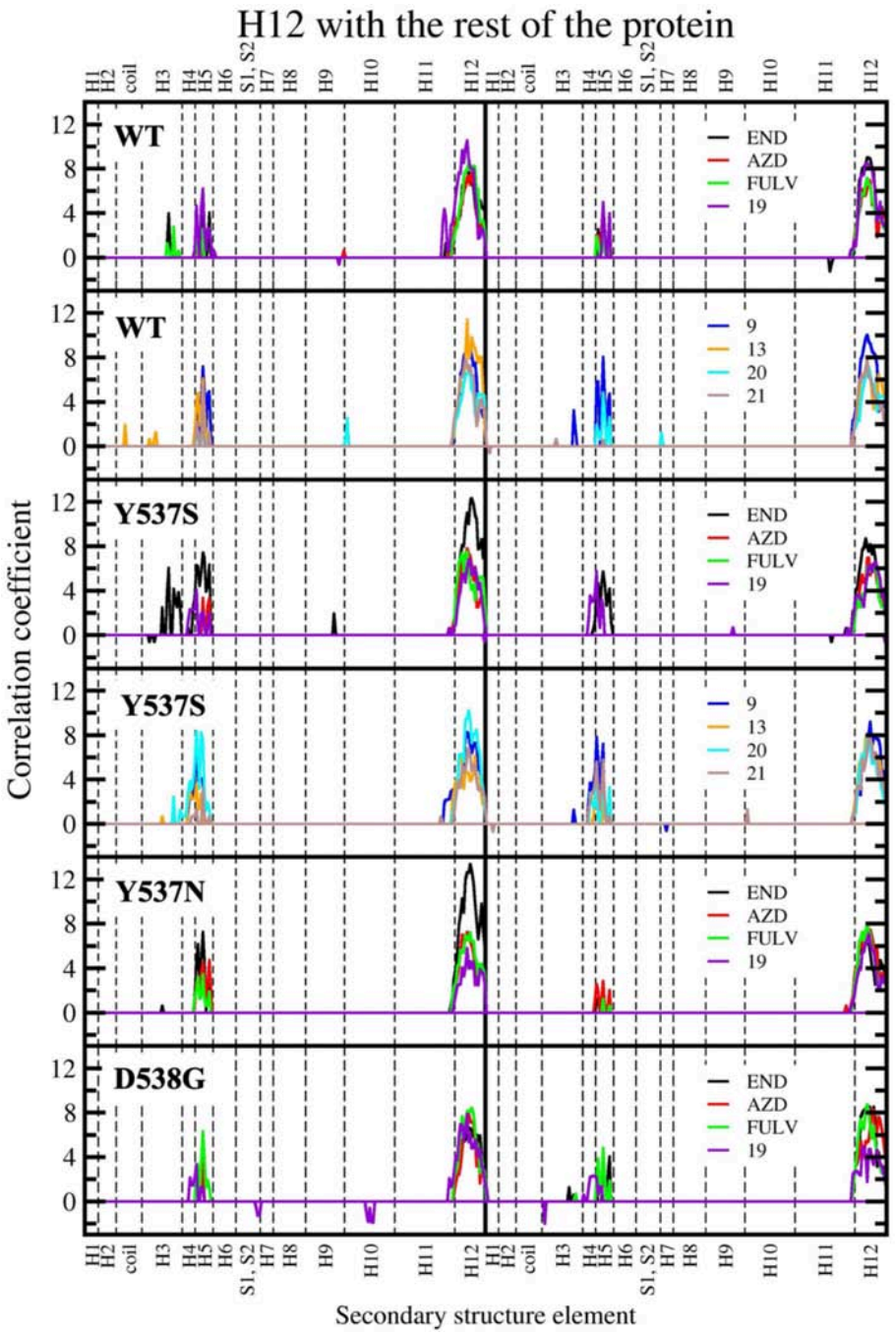


FIGURE 8 | Sum of per-residue cross-correlation coefficients. Left and right columns refer to monomers A and B, respectively. From top to bottom: WT (in complex with END, AZD, FULV, and **19**), WT (with **9**, **13**, **20**, and **21**), Y537S (with END, AZD, FULV, and **19**), Y537S (with **9**, **13**, **20**, and **21**), Y537N, and D538G (in complex with END, AZD, FULV, and **19**). END, AZD, FULV, and compounds **9**, **13**, **19**, **20**, **21** are shown as black, red, green, blue, orange, purple, cyan, and brown lines, respectively.

coupling between H12 and H4-H5 higher or similar to END at least in one Y537S ERα monomer, while **19** effectively reduces this positive correlation in both monomers. This correlation coupling is completely abolished in Y537N and reduced even in D538G (**Figure 8**), pointing to an activity of **19** also against these mutants, even if to a minor extent.

As a result, **19** appears to reduce the transcriptional activity of Y537S cells thanks to its capability of binding in the LBC of both LBD monomers only in the mERαs, where it establishes H-bonds with E415, L346 and K529, similarly to AZD.

Kinetic Characterization of Active Compounds

Since increasing evidences pinpoint the dissociation free energy barriers ($\Delta G_d^{\#}$) of a ligand from its binding cavity to be strongly entwined with the residence time and, thus, with drugs' efficacy (Magistrato et al., 2017), the observed preferential activity of **19** toward Y537S ERα fostered the investigations of its kinetic properties as compared to those of AZD.

The free energy surface (FES) obtained from MTD simulations inducing the dissociation of AZD from the LBC of Y537S ERα shows a wide minimum at Center of Mass (COM) distance between ligand and protein at ∼1.2 nm, which, instead, spans the coordination number (CN) 0.2-0.4. A second, narrower, minimum appears at CN around 0 and COM distance ≥ 2.5 nm. By inspecting two-dimensional FES plots of both CVs one can estimate a $\Delta G_d^{\#}$ of 14.1 ± 2.0 kcal/mol for AZD dissociation (**Figure 9A**, **Supplementary Figures 10**, **11**, and **Supplementary Table 6**). The main barrier observed for AZD dissociation is due to the breaking of its H-bond interactions between the carboxylic group of AZD with K529 and C530. These, therefore, appear as pivotal residues for increasing the residence time of this drug in the LBC and possibly its efficacy.

On the other hand, FES for the dissociation of **19** reveals a rather wide minimum at CN = 0.2–0.4, lying at higher distance (COM) between ligand and the LBC as compared to AZD (1.5–1.7 nm) (**Figure 9B**). The second minimum is located in a similar position to that of AZD. In this case, however, the $\Delta G_d^{\#}$ is rather small (3.7 ± 1.9 kcal/mol) (**Supplementary Figures 10**, **11** and **Supplementary Table 6**) and it is associated to the breaking of the H-bond between the hydroxyl group of the ligand and the E418 residue. This latter, therefore, appears to be a distinctive feature of this ligand.

These simulations pinpoint the most important substituents of the ligand that may contribute to improve the kinetic properties of the drugs, and the residues of the LBC that must be engaged in specific interactions for the discovery of mutant-specific anti-estrogen compounds.

DISCUSSION

Breast cancer remains the most diagnosed (1 over 8) and the second leading cause of cancer induced mortality in women. The majority (70 %) of BC is hormone dependent and its proliferation relies on the presence of ERα, which has a pro-oncogenic effect in the presence of estrogens. The gold standard treatment in this type of BC is the hormone adjuvant therapy, which either suppresses estrogen production (aromatase inhibitors) or modulates/degrades the ERα (SERMs/SERDs). The prolonged exposure to these therapies, usually administered consecutively for 5–10 years' time-frame, leads to resistance in half of all luminal BCs after 5 years, in spite of the ERα expression (Toy et al., 2017; Fanning et al., 2018; Busonero et al., 2019; Spinello et al., 2019b).

While the genomic profile of inherited and somatic alterations characterizing each type of BC is well-established, the evolution of the BC's genomic landscape under the evolutionary pressure of systemic therapies is not clearly understood. As well as how this landscape impacts on the clinical outcome of endocrine therapies remains poorly characterized and is currently object of intense research efforts. Resistance onset is, in fact, responsible of refractory BCs and of an increased mortality rate. In this worrisome scenario, the therapeutic options to intervene with personalized treatments based on the patients' evolution of the genomic profiles remains a daunting challenge. This has spurred substantial efforts to characterize the phenotype responsible of drug resistance and propose innovative therapeutic options.

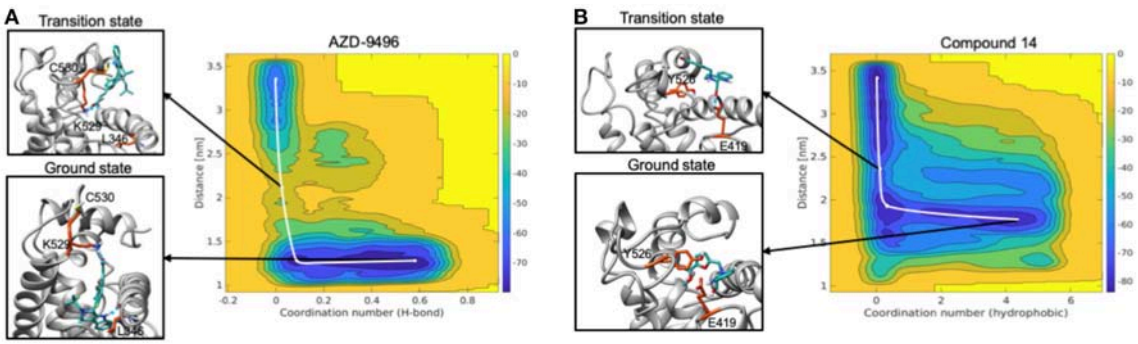


FIGURE 9 | Free energy surface maps from metadynamics simulations for the dissociation of AZD **(A)** and compound **19 (B)** from the ligand binding cavity. On x axis is coordination number (H-bond for AZD and hydrophobic interactions for **19**) and on y axis is COM distance between ligand and receptor. Color scale represents free energy values in kJ/mol. In black squares are encircled structures corresponding to the ligand bound state (ground state), the transition state.

Distinct studies indicated that frequent mutations present in the loop connecting H11 and H12 of the LBD trigger the acquisition of an intrinsically active (agonist-like) ERα conformation, even in the absence of E2. This conformation remains even in the presence of SERMs (Fanning et al., 2016; Pavlin et al., 2018). Our recent computational attempt to identify the key common structural traits that drugs should possess in order to effectively fight resistant BCs was the grounding knowledge for the present study. Indeed, here we carried out *in silico* screening on the structural scaffolds of the Y537S, Y537N, and D538G mutants adapted to known mERα degraders (AZD and FULV), seeking for the structural elements able to protrude toward loop connecting H11-H12. This should allow the ligand to counteract the intrinsic and mutant dependent ERα activation (Pavlin et al., 2018).

From a consensus docking study, we selected 17 molecules (**Supplementary Figure 3**) effectively binding in at least two mutants, among which five resulted to be active on BC cell lines. Some of them were known to be active also on other targets and diseases (**Supplementary Table 7**). Among these compounds, **19** was selective exclusively toward those expressing Y537S (and to a minor extent to Y537N, D538G) ERα (**Figures 2–4**). In spite of its ability to block the transcriptional activity of the receptor only in the high μM range, thus requiring further optimization, the structural scaffold of compound **19** encompasses all the motifs required by an active and mutant-specific drug-candidate. Namely, **19** forms number of H-bonds in the ligand binding cavity (L346 and E419) and with K529. Conversely, E380, a key residue involved in the structural transition toward an agonist-like state of the receptor, persistently H-bonds to H377. This is a previously annotated structural feature able to impede the pro-oncogenic effect of resistant phenotypes. Indeed, compound **19**, to the best of our knowledge, is the only mutant specific modulator of ERα transcriptional activity identified so far. However, its ΔG_b and $\Delta G_d^{\#}$ are remarkably smaller than the parent AZD compound. A detailed comparison among the residues, which optimize these thermodynamics and kinetic properties of the **19** with respect to those of AZD is informative for future knowledge-based drug-design efforts aimed at discovering drug-candidates with superior efficacy.

Since ESR1 mutations are potential clinical biomarkers to guide therapeutic decisions, identification of small molecules able to block proliferation of metastatic tumors expressing one prevalent mERα resistant phenotypes may result in counteracting, preventing and/or delaying their occurrence in early disease stage. In this scenario, our study contributes to move a step forward toward precision and personalized medicine tailored against metastatic and resistant ER+ BCs.

DATA AVAILABILITY

All datasets generated for this study are included in the manuscript/**Supplementary Files**.

AUTHOR CONTRIBUTIONS

LG and IB performed experiments. MP and AS performed simulations and analysis. IB, SA, SC, and AM designed research and wrote the manuscript.

FUNDING

This work was supported by My First AIRC Grant (MFAG) #16899 to IB, MFAG Grant #17134 to AM, AIRC Investigator Grant (IG) #21414 to SC, AIRC IG #18602 to SA.

ACKNOWLEDGMENTS

The authors thank AIRC for financial support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00602/full#supplementary-material

SUPPORTING INFORMATION

Complementary figures and tables are reported in the supporting information. Structures of the docked compounds before and after molecular dynamics simulation are available upon request to the authors.

REFERENCES

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2, 19–25. doi: 10.1016/j.softx.2015.06.001

Anandakrishnan, R., Aguilar, B., and Onufriev, A. V. (2012). H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucl Acids Res.* 40, W537–41. doi: 10.1093/nar/gks375

Barone, I., Brusco, L., and Fuqua, S. A. (2010). Estrogen receptor mutations and changes in downstream gene expression and signaling. *Clin. Cancer Res.* 16, 2702–2708. doi: 10.1158/1078-0432.CCR-09-1753

Barone, I., Brusco, L., Gu, G., Selever, J., Beyer, A., Covington, K. R., et al. (2011). Loss of Rho GDIalpha and resistance to tamoxifen via effects on estrogen receptor alpha. *J. Natl. Cancer Inst.* 103, 538–552. doi: 10.1093/jnci/djr058

Bayly, C. I., Cieplak, P., Cornell, W. D., and Kollman, P. A. (1993). A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges - the RESP model. *J. Phys. Chem.* 97, 10269–10280. doi: 10.1021/J100142a004

Bisha, I., Laio, A., Magistrato, A., Giorgetti, A., and Sgrignani, J. (2013). A candidate ion-retaining state in the inward-facing conformation of sodium/galactose symporter: clues from atomistic simulations. *J. Chem. Theory Comput.* 9, 1240–1246. doi: 10.1021/ct3008233

Busonero, C., Leone, S., Bartoloni, S., and Acconcia, F. (2019). Strategies to degrade estrogen receptor alpha in primary and ESR1 mutant-expressing

metastatic breast cancer. *Mol. Cell. Endocrinol.* 480, 107–121. doi: 10.1016/j.mce.2018.10.020

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126:014101. doi: 10.1063/1.2408420

Casalino, L., Palermo, G., Spinello, A., Rothlisberger, U., and Magistrato, A. (2018). All-atom simulations disentangle the functional dynamics underlying gene maturation in the intron lariat spliceosome. *Proc. Natl. Acad. Sci. U.S.A.* 115, 6584–6589. doi: 10.1073/pnas.1802963115

Catalano, S., Leggio, A., Barone, I., De Marco, R., Gelsomino, L., Campana, A., et al. (2015). A novel leptin antagonist peptide inhibits breast cancer growth *in vitro* and *in vivo*. *J. Cell. Mol. Med.* 19, 1122–1132. doi: 10.1111/jcmm.12517

Chandarlapaty, S., Chen, D., He, W., Sung, P., Samoila, A., You, D., et al. (2016). Prevalence of ESR1 Mutations in cell-free DNA and outcomes in metastatic breast cancer: a secondary analysis of the BOLERO-2 clinical trial. *JAMA Oncol.* 2, 1310–1315. doi: 10.1001/jamaoncol.2016.1279

Covington, K. R., Brusco, L., Barone, I., Tsimelzon, A., Selever, J., Corona-Rodriguez, A., et al. (2013). Metastasis tumor-associated protein 2 enhances metastatic behavior and is associated with poor outcomes in estrogen receptor-negative breast cancer. *Breast Cancer Res. Treat.* 141, 375–384. doi: 10.1007/s10549-013-2709-5

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh ewald: An N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98, 10089–10092. doi: 10.1063/1.464397

Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W. F., and Mark, A. E. (1999). Peptide folding: when simulation meets experiment. *Angew. Chem. Int. Ed.* 38, 236–240.

De Savi, C., Bradbury, R. H., Rabow, A. A., Norman, R. A., de Almeida, C., Andrews, D. M., et al. (2015). Optimization of a novel binding motif to (E)-3-(3,5-difluoro-4-((1R,3R)-2-(2-fluoro-2-methylpropyl)-3-methyl-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indol-1-yl)phenyl)acrylic Acid (AZD9496), a potent and orally bioavailable selective estrogen receptor downregulator and antagonist. *J. Med. Chem.* 58, 8128–8140. doi: 10.1021/acs.jmedchem.5b00984

Duan, J., Dixon, S. L., Lowrie, J. F., and Sherman, W. (2010). Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.* 29, 157–170. doi: 10.1016/j.jmgm.2010.05.008

Fanning, S. W., and Greene, G. L. (2019). Next-generation ERalpha inhibitors for endocrine-resistant ER+ breast cancer. *Endocrinology* 160, 759–769. doi: 10.1210/en.2018-01095

Fanning, S. W., Hodges-Gallagher, L., Myles, D. C., Sun, R., Fowler, C. E., Plant, I. N., et al. (2018). Specific stereochemistry of OP-1074 disrupts estrogen receptor alpha helix 12 and confers pure antiestrogenic activity. *Nat. Commun.* 9:2368. doi: 10.1038/s41467-018-04413-3

Fanning, S. W., Mayne, C. G., Dharmarajan, V., Carlson, K. E., Martin, T. A., Novick, S. J., et al. (2016). Estrogen receptor alpha somatic mutations Y537S and D538G confer breast cancer endocrine resistance by stabilizing the activating function-2 binding conformation. *eLife* 5:e12792. doi: 10.7554/eLife.12792

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* 47, 1739–1749. doi: 10.1021/jm0306430

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., et al. (2016). *Gaussian 09, Revision A.02*. Wallingford CT: Gaussian, Inc.

Gelsomino, L., Gu, G., Rechoum, Y., Beyer, A. R., Pejerrey, S. M., Tsimelzon, A., et al. (2016). ESR1 mutations affect anti-proliferative responses to tamoxifen through enhanced cross-talk with IGF signaling. *Breast Cancer Res. Treat.* 157, 253–265. doi: 10.1007/s10549-016-3829-5

Gelsomino, L., Panza, S., Giordano, C., Barone, I., Gu, G., Spina, E., et al. (2018). Mutations in the estrogen receptor alpha hormone binding domain promote stem cell phenotype through notch activation in breast cancer cell lines. *Cancer Lett.* 428, 12–20. doi: 10.1016/j.canlet.2018.04.023

Giordano, C., Chemi, F., Panza, S., Barone, I., Bonofiglio, D., Lanzino, M., et al. (2016). Leptin as a mediator of tumor-stromal interactions promotes breast cancer stem cell activity. *Oncotarget* 7, 1262–1275. doi: 10.18632/oncotarget.6014

Hamilton, E. P., Patel, M. R., Armstrong, A. C., Baird, R. D., Jhaveri, K., Hoch, M., et al. (2018). A first-in-human study of the new oral selective estrogen receptor degrader AZD9496 for ER(+)/HER2(-) advanced breast cancer. *Clin. Cancer. Res.* 24, 3510–3518. doi: 10.1158/1078-0432.CCR-17-3102

Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997). LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* 18, 1463–1472.

Huang, W., Peng, Y., Kiselar, J., Zhao, X., Albaqami, A., Mendez, D., et al. (2018). Multidomain architecture of estrogen receptor reveals interfacial cross-talk between its DNA-binding and ligand-binding domains. *Nat. Commun.* 9:3520. doi: 10.1038/s41467-018-06034-2

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual molecular dynamics. *J. Mol. Graph.* 14, 33–38, 27–38.

Jeselsohn, R., Yelensky, R., Buchwalter, G., Frampton, G., Meric-Bernstam, F., Gonzalez-Angulo, A. M., et al. (2014). Emergence of constitutively active estrogen receptor-alpha mutations in pretreated advanced estrogen receptor-positive breast cancer. *Clin. Cancer Res.* 20, 1757–1767. doi: 10.1158/1078-0432.CCR-13-2332

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79, 926–935. doi: 10.1063/1.445869

Joseph, J. D., Darimont, B., Zhou, W., Arrazate, A., Young, A., Ingalla, E., et al. (2016). The selective estrogen receptor downregulator GDC-0810 is efficacious in diverse models of ER+ breast cancer. *eLife* 5:e15828. doi: 10.7554/eLife.15828

Kahraman, M., Govek, S. P., Nagasawa, J. Y., Lai, A., Bonnefous, C., Douglas, K., et al. (2019). Maximizing ER-alpha degradation maximizes activity in a Tamoxifen-resistant breast cancer model: Identification of GDC-0927. *ACS Med. Chem. Lett.* 10, 50–55. doi: 10.1021/acsmedchemlett.8b00414

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., et al. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 78, 1950–1958. doi: 10.1002/prot.22711

Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.* 46, 3–26. doi: 10.1016/S0169-409X(00)00129-0

Magistrato, A., Sgrignani, J., Krause, R., and Cavalli, A. (2017). Single or multiple access channels to the CYP450s active site? an answer from free energy simulations of the human aromatase enzyme. *J. Phys. Chem. Lett.* 8, 2036–2042. doi: 10.1021/acs.jpclett.7b00697

Merenbakh-Lamin, K., Ben-Baruch, N., Yeheskel, A., Dvir, A., Soussan-Gutman, L., Jeselsohn, R., et al. (2013). D538G mutation in estrogen receptor-alpha: A novel mechanism for acquired endocrine resistance in breast cancer. *Cancer Res.* 73, 6856–6864. doi: 10.1158/0008-5472.CAN-13-1197

Nilsson, S., and Gustafsson, J. A. (2011). Estrogen receptors: Therapies targeted to receptor subtypes. *Clinical Pharmacol. Ther.* 89, 44–55. doi: 10.1038/clpt.2010.226

Palermo, G., Miao, Y., Walker, R. C., Jinek, M., and McCammon, J. A. (2016). Striking plasticity of CRISPR-Cas9 and key role of non-target DNA, as revealed by molecular simulations. *ACS Cent. Sci.* 2, 756–763. doi: 10.1021/acscentsci.6b00218

Pavlin, M., Spinello, A., Pennati, M., Zaffaroni, N., Gobbi, S., Bisi, A., et al. (2018). A Computational assay of estrogen receptor alpha antagonists reveals the key common structural traits of drugs effectively fighting refractory breast cancers. *Sci. Rep.* 8, 649. doi: 10.1038/s41598-017-17364-4

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Razavi, P., Chang, M. T., Xu, G., Bandlamudi, C., Ross, D. S., Vasan, N., et al. (2018). The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer Cell* 34, 427–438 e426. doi: 10.1016/j.ccell.2018.08.008

Rizza, P., Barone, I., Zito, D., Giordano, F., Lanzino, M., De Amicis, F., et al. (2014). Estrogen receptor beta as a novel target of androgen receptor action in breast cancer cell lines. *Br Cancer Res.* 16:R21. doi: 10.1186/bcr3619

Robinson, D. R., Wu, Y. M., Vats, P., Su, F., Lonigro, R. J., Cao, X., et al. (2013). Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nat. Genet.* 45, 1446–1451. doi: 10.1038/ng.2823

Sastry, M., Lowrie, J. F., Dixon, S. L., and Sherman, W. (2010). Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* 50, 771–784. doi: 10.1021/ci100062n

Schrödinger Release 2017-1 (2017). *Small-Molecule Drug Discovery Suite 2017-1*. New York, NY: Schrodinger, LLC.

Scott, J. S., Bailey, A., Buttar, D., Carbajo, R. J., Curwen, J., Davey, P. R. J., et al. (2019). Tricyclic indazoles-a novel class of selective estrogen receptor degrader antagonists. *J. Med. Chem.* 62, 1593–1608. doi: 10.1021/acs.jmedchem.8b01837

Sgrignani, J., and Magistrato, A. (2015). QM/MM MD simulations on the enzymatic pathway of the human flap endonuclease (hFEN1) elucidating common cleavage pathways to RNase H enzymes. *ACS Catal.* 5, 3864–3875. doi: 10.1021/acscatal.5b00178

Sharma, A., Toy, W., Guillen, V. S., Sharma, N., Min, J., Carlson, K. E., et al. (2018). Antagonists for constitutively active mutant estrogen eeceptors: insights into the roles of antiestrogen-core and side-chain. *ACS Chem. Biol.* 13, 3374–3384. doi: 10.1021/acschembio.8b00877

Spinello, A., Martini, S., Berti, F., Pennati, M., Pavlin, M., Sgrignani, J., et al. (2019a). Rational design of allosteric modulators of the aromatase enzyme: an unprecedented therapeutic strategy to fight breast cancer. *Eur. J. Med. Chem.* 168, 253–262. doi: 10.1016/j.ejmech.2019.02.045

Spinello, A., Pavlin, M., Casalino, L., and Magistrato, A. (2018). A dehydrogenase dual hydrogen abstraction mechanism promotes estrogen biosynthesis: can we expand the functional annotation of the aromatase enzyme? *Chem.-Eur. J.* 24, 10840–10849. doi: 10.1002/chem.201802025

Spinello, A., Ritacco, I., and Magistrato, A. (2019b). Recent advances in computational design of potent aromatase inhibitors: open-eye on endocrine-resistant breast cancers. *Expert. Opin. Drug. Discov.* 245:1–12. doi: 10.1080/17460441.2019.1646245

Spinello, A., Vecile, E., Abbate, A., Dobrina, A., and Magistrato, A. (2019c). How can interleukin-1 receptor antagonist modulate distinct cell death pathways? *J. Chem. Inf. Model.* 59, 351–359. doi: 10.1021/acs.jcim.8b00565

Toy, W., Shen, Y., Won, H., Green, B., Sakr, R. A., Will, M., et al. (2013). ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nat. Genet.* 45, 1439–1445. doi: 10.1038/ng.2822

Toy, W., Weir, H., Razavi, P., Lawson, M., Goeppert, A. U., Mazzola, A. M., et al. (2017). Activating ESR1 mutations differentially affect the efficacy of ER antagonists. *Cancer Discov.* 7, 277–287. doi: 10.1158/2159-8290.CD-15-1523

Tribello, G., Bonomi, M., Branduardi, D., Camilloni, C., and Bussi, G. (2014). PLUMED2: New feathers for an old bird. *Comp. Phys. Comm.* 185:606–613. doi: 10.1016/j.cpc.2013.09.018

van Kruchten, M., de Vries, E. G., Glaudemans, A. W., van Lanschot, M. C., van Faassen, M., Kema, I. P., et al. (2015). Measuring residual estrogen receptor availability during fulvestrant therapy in patients with metastatic breast cancer. *Cancer Discov.* 5, 72–81. doi: 10.1158/2159-8290.CD-14-0697

Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W., and Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* 45, 2615–2623. doi: 10.1021/jm020017n

Wang, J., Wang, W., Kollman, P. A., and Case, D. A. (2006). Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* 25, 247–260. doi: 10.1016/j.jmgm.2005.12.005

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *J. Comput. Chem.* 25, 1157–1174. doi: 10.1002/jcc.20035

Weir, H. M., Bradbury, R. H., Lawson, M., Rabow, A. A., Buttar, D., Callis, R. J., et al. (2016). AZD9496: an oral estrogen receptor inhibitor that blocks the growth of ER-positive and ESR1-mutant breast tumors in preclinical models. *Cancer Res.* 76, 3307–3318. doi: 10.1158/0008-5472.CAN-15-2357

Wickstrom, L., Okur, A., and Simmerling, C. (2009). Evaluating the performance of the ff99SB force field based on NMR scalar coupling data. *Biophys. J.* 97, 853–856. doi: 10.1016/j.bpj.2009.04.063

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Pavlin, Gelsomino, Barone, Spinello, Catalano, Andò and Magistrato. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 23 October 2019
doi: 10.3389/fchem.2019.00701

Teaching an Old Dog New Tricks: Strategies That Improve Early Recognition in Similarity-Based Virtual Screening

Ruifeng Liu[1,2]*, *Mohamed Diwan M. AbdulHameed*[1,2] *and Anders Wallqvist*[1]*

[1] Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Development Command, Fort Detrick, MD, United States, [2] The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, MD, United States

OPEN ACCESS

Edited by:
Jose L. Medina-Franco, National Autonomous University of Mexico, Mexico

Reviewed by:
Olivier Sperandio, Institut National de la Santé et de la Recherche Médicale (INSERM), France
Eva Nittinger, Universität Hamburg, Germany

***Correspondence:**
Ruifeng Liu
rliu@bhsai.org
Anders Wallqvist
sven.a.wallqvist.civ@mail.mil

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: 26 July 2019
Accepted: 08 October 2019
Published: 23 October 2019

Citation:
Liu R, AbdulHameed MDM and Wallqvist A (2019) Teaching an Old Dog New Tricks: Strategies That Improve Early Recognition in Similarity-Based Virtual Screening. Front. Chem. 7:701. doi: 10.3389/fchem.2019.00701

High throughput screening (HTS) is an important component of lead discovery, with virtual screening playing an increasingly important role. Both methods typically suffer from lack of sensitivity and specificity against their true biological targets. With ever-increasing screening libraries and virtual compound collections, it is now feasible to conduct follow-up experimental testing on only a small fraction of hits. In this context, advances in virtual screening that achieve enrichment of true actives among top-ranked compounds ("early recognition") and, hence, reduce the number of hits to test, are highly desirable. The standard ligand-based virtual screening method for large compound libraries uses a molecular similarity search method that ranks the likelihood of a compound to be active against a drug target by its highest Tanimoto similarity to known active compounds. This approach assumes that the distributions of Tanimoto similarity values to all active compounds are identical (i.e., same mean and standard deviation)—an assumption shown to be invalid (Baldi and Nasr, 2010). Here, we introduce two methods that improve early recognition of actives by exploiting similarity information of all molecules. The first method ranks a compound by its highest z-score instead of its highest Tanimoto similarity, and the second by an aggregated score calculated from its Tanimoto similarity values to all known actives and inactives (or a large number of structurally diverse molecules when information on inactives is unavailable). Our evaluations, which use datasets of over 20 HTS campaigns downloaded from PubChem, indicate that compared to the conventional approach, both methods achieve a ~10% higher Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) score—a metric of early recognition. Given the increasing use of virtual screening in early lead discovery, these methods provide straightforward means to enhance early recognition.

Keywords: lead discovery, virtual screening, early recognition, Tanimoto similarity, z-score, BEDROC, ROCS

INTRODUCTION

Lead discovery by high throughput screening (HTS) is often described as a process akin to finding a needle in a haystack (Aherne et al., 2002). Given the significant achievements in automation, major pharmaceutical companies now routinely screen hundreds of thousands of samples to identify compounds that are active against specific drug targets. However, the number of chemicals available for bioactivity testing has increased exponentially over the past decade. For instance, as of 2015, the number of structurally unique chemicals registered in PubChem was more than 60 million (Kim et al., 2016), and in 2018, the total number of organic and inorganic substances disclosed in the literature was estimated to be 154 million[1]. Thus, despite the increased screening capacity, it remains impractical to assay a significant fraction of all available chemicals. Consequently, virtual screening is becoming increasingly important to prioritize and select compounds (Kar and Roy, 2013). The most widely used virtual screening methods are based on molecular similarity searches (Kristensen et al., 2013). These approaches typically rank molecules in a chemical library based on their structural similarity to a set of molecules known to be active against a desired target. Chemicals ranked high on the list can then be acquired and tested for the desired activity or property.

The most commonly used metric to compare the performance of different virtual screening methods is the area under the receiver operating characteristic curve (ROC_AUC) (Triballeau et al., 2005). This is useful for comparing overall performance of methods for ranking a database (Truchon and Bayly, 2007; Zhao et al., 2009). However, the ROC_AUC is inappropriate for virtual screening when the goal is to create a smaller subset enriched with the maximum number of actives (Truchon and Bayly, 2007). The distinction is critical, especially when the chemical libraries are large and only a small fraction of compounds can be tested. Truchon and Bayly (2007) illustrated the difference using three basic cases: (1) half of the actives ranked at the top of a rank-ordered list and the other half at the bottom; (2) all actives randomly distributed across ranks; and (3) all actives ranked in the middle of the list. In all three cases, the ROC_AUC value is 0.5 and, therefore, according to this metric, all three virtual screening methods that generated the three rank-ordered lists perform equally. However, because only a small fraction of chemicals in a large library can be tested, "early recognition" of actives is practically important. That is, case 1 is preferable to case 2 or 3, and case 2 could be considered more desirable than case 3.

Many metrics have been proposed to address early recognition. Examples include the partial area under the ROC curve (McClish, 1989), enrichment factor (Halgren et al., 2004), area under the accumulation curve (Kairys et al., 2006), robust initial enhancement (Sheridan et al., 2001), Boltzmann-enhanced discrimination of the receiver operating characteristic (BEDROC) (Truchon and Bayly, 2007), and predictiveness curve (Empereur-Mot et al., 2015). Although no metric is perfect, perhaps the most frequently adopted is BEDROC, which employs an adjustable parameter, α, to define "early detection." Truchon and Bayly suggest setting this parameter to 20.0, which dictates that 80% of the maximum contribution to BEDROC comes from the top 8% of the ranked list. A comparatively higher BEDROC score between two virtual screening methods indicates an enhanced ability to enrich the list of top-ranking compounds with active molecules.

Using both AUC_ROC and BEDROC, Nasr et al. (2009) carried out a large-scale study of the performance of 14 similarity search methods, including eight parameter-free methods (no parameters to be learned from training data) and six with one or two parameters to be learned from training data. Consistent with previous results, they found that the best parameter-free method is the Max-Sim method, which ranks molecules based on their maximum Tanimoto coefficient (TC, also commonly referred to as Tanimoto similarity) to the active query molecules. Among the six methods that require parameters to be fit to the data, the exponential Tanimoto discriminant (ETD) method was the best performer overall. This method is defined by the following equations.

$$S(B) = \frac{\sum_{i=1}^{m} S(A_i, B)}{\sum_{j=1}^{n} S(I_j, B)} \quad (1)$$

$$S(A,B) = \left[\lambda^{TC_{AB}}(1-\lambda)^{1-TC_{AB}}\right]^{\frac{1}{k}} \quad (2)$$

$$TC_{AB} = \frac{A \cap B}{A \cup B} \quad (3)$$

Here, $S(B)$ denotes the aggregated score for molecule B, m, and n, respectively denote the numbers of active and inactive query molecules, A_i denotes the ith active query molecule, I_j denotes the jth inactive molecule, TC_{AB} denotes the TC between molecules A and B, λ, and k denote parameters to be learned from the data. The higher the aggregated score, the more likely it is that molecule B is active. Nasr et al. (2009) provided neither recommended default parameter values for λ and k, nor values learned from any of their datasets.

In this article, we introduce two parameter-free similarity search methods that improve the early recognition of actives over the Max-Sim method. Using HTS data, we demonstrate that, on average, the BEDROC values derived from both methods are about 10% higher than those derived from the Max-Sim scoring method.

METHODS AND MATERIALS

Rank by Z-Scores

In a Max-Sim similarity search, we first calculate all TCs between the compounds in a chemical library and active query molecules. The library compounds are then ranked based on their highest TCs. The underlying assumption is that the higher the TC, the more likely a compound is to be active. This assumption is valid for searches with a single active query molecule, and for searches with multiple active query molecules if the distributions of TCs are identical (i.e., have the same mean and standard deviation irrespective of the query molecules). Although it has been standard practice for many years to conduct Max-Sim similarity searches, no study had examined the statistical distribution of

[1]https://www.cas.org/about/cas-content (accessed July 5, 2019).

TCs until 2010, when Baldi and Nasr (2010) investigated in detail the significance of Tanimoto similarity. They showed that the statistical distribution of TCs is not invariant, but depends on the number of fingerprint features present in a query molecule. This finding and its implications, however, are largely overlooked by the cheminformatics community, perhaps due to the reasonably good performance of the Max-Sim method and the extremely low mean TCs for any query molecule. As an example, **Figure 1** shows the means and standard deviations of the TCs of 10,000 chemicals randomly selected from the U.S. National Cancer Institute (NCI) chemical library calculated with respect to each of three drugs approved by the U.S. Food and Drug Administration. All of the means and standard deviations are close to zero, suggesting that most NCI compounds do not have the same activity as that of the approved drugs. The small mean TCs may obscure an important fact—that the values are not identical and could be significantly different. For instance, the mean TC of scopolamine is 43% higher than that of pemirolast. To appreciate the implications of the difference, let us assume that a molecule has TCs of 0.80 and 0.70 calculated with respect to scopolamine and pemirolast, respectively. Based on the Max-Sim method, one would expect the molecule to have activities more similar to those of scopolamine. However, because of the difference in the means and standard deviations, the z-scores of the molecule calculated with respect to scopolamine and pemirolast are 17.3 and 22.8, respectively, suggesting that the molecule is more likely to have activities similar to those of pemirolast than to those of scopolamine. If we consider that there are differences in mean TCs and standard deviations, then ranking molecules by the maximum z-score is statistically preferable in a similarity search. Accordingly, we designate this approach as the maxZ method.

Rank by Aggregated Similarity

In the past two decades, HTS has contributed to the discovery of numerous structurally novel active compounds against many important drug targets. As these compounds are identified from large experimentally tested screening libraries, they are classified as either active or inactive based on predefined activity criteria. In follow-up studies based on virtual screening by similarity search methods, only active compounds are used as queries. As noted in the Introduction, Nasr et al. (2009) developed the ETD method, which exploits information of both active and inactive compounds. They found that ETD performed best among 14 parameterized and non-parameterized TC-based similarity search methods. An undesirable feature of this method, however, is that it requires two parameters that may not be universally applicable, but still need to be fit for each individual dataset. Here, we propose an aggregated similarity (AS) method that does not require any parameter fitting based on individual datasets. We define the AS method by the following equation:

$$AS\,(X) = \frac{\sum_{i=1}^{m} e^{-\frac{1-TC(A_i,X)}{\alpha+TC(A_i,X)}}}{\sum_{j=1}^{n} e^{-\frac{1-TC(I_j,X)}{\alpha+TC(I_j,X)}}} \tag{4}$$

where, X denotes a compound in a chemical library, m and n, respectively denote the number of active and inactive molecules, $TC(A_i, X)$ denotes the TC between the ith active molecule and X, $TC(I_j, X)$ denotes the TC between the jth inactive molecule and X, and α is set to 10^{-6}—a small number to avoid division by zero when TC equals zero. Possible *AS(X)* values range from zero to infinity, where zero indicates that molecule X shares no fingerprint features with any of the active query molecules, i.e., all $TC\,(A_i, X) = 0$, and infinity indicates that molecule X shares no fingerprint features with any of the inactive query molecules. In reality, because the number of inactive molecules is large (i.e., a positive is like a needle in a haystack and, therefore, most molecules can be classified as inactives), the probability of X sharing no fingerprint features with any of the inactive query molecules is zero, unless a very small number of inactive query molecules is used (even though a large number of them should be available).

One problem with using information on inactive compounds is that the results of large-scale screening campaigns are not equally reliable for active and inactive compounds. This is because such campaigns are typically executed in multiple confirmatory steps focusing on active compounds. The first step involves an initial primary screening of a large number of samples at a single concentration with few or no replicates. Samples deemed to meet the primary activity criteria are then selected and retested in multiple replicates, usually with counter-assays to affirm activity. Samples that satisfy the retesting criteria may be further tested at multiple concentrations to determine potency. One consequence of this screening protocol is that the activities of a positive compound are more reliable because they are reassessed in multiple tests, whereas compounds fail to meet primary active criteria are not retested to confirm inactivity. As a result, the set of inactive molecules is likely to contain false negatives. A more obvious problem with the AS method is that it cannot be applied to cases where information on inactives is unavailable. As a means to overcome this challenge, we suggest that a set of structurally diverse compounds can be used as putative inactive compounds. This is because compounds that are truly active against the most valuable drug targets are rare (i.e., needles in a haystack). Therefore, within a structurally diverse set of compounds, the number of compounds that are active against a drug target should be small. Here, we tested the validity of this hypothesis by using 10,000 structurally diverse compounds as putative inactives. We selected these compounds by clustering ∼275,000 compounds of the NCI virtual screening library (Shiryaev et al., 2011) into 10,000 clusters based on the TC (a measure of molecular similarity), and selecting the cluster centers as structurally diverse compounds to represent coverage of the chemical space of the full dataset. In doing so, we considered a singleton as a cluster of size one.

Datasets

We evaluated the performance of the similarity search methods using HTS data generated from the National Center for Advancing Translational Sciences of the National Institutes of Health. We downloaded the data in two batches. The first batch consisted of the results of ∼8,000 samples screened against 10 toxicity-related targets using 12 different assays, with two different assays deployed for two of the 10 targets. Thus, roughly

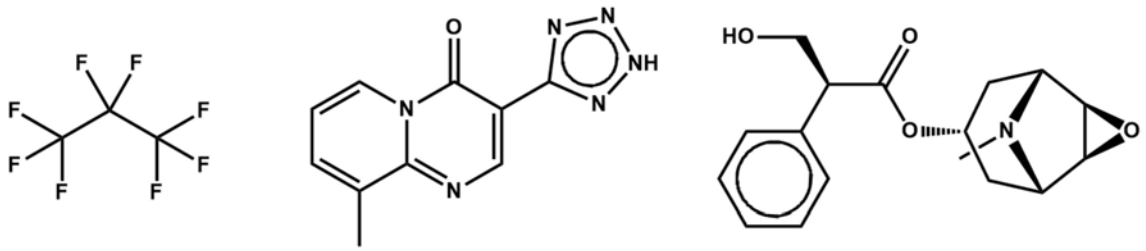


FIGURE 1 | Examples of means and standard deviations (STD) of the Tanimoto coefficients (TCs) of 10,000 compounds randomly selected from the National Cancer Institute's virtual screening library calculated with respect to three drugs approved by the Food and Drug Administration.

the same 8,000 samples were tested in 12 assays, generating 12 molecular activity datasets. As these datasets were used in the Tox21 Data Challenge for molecular activity predictions (Huang and Xia, 2017), we downloaded them from Tox21 Data Challenge web site[2]. Because the datasets were relatively small (consisting of ∼7,000 structurally unique compounds), we used them to evaluate maxZ scoring methods based on two-dimensional (2-D) molecular fingerprints and three-dimensional (3-D) molecular shapes.

A library consisting of 8,000 samples can hardly be considered a "large" library for HTS. Therefore, we used a second batch of data that consisted of results for a few thousand to a few hundred thousand samples screened against 12 different molecular targets. We downloaded these data from the PubChem web site (https://pubchem.ncbi.nlm.nih.gov/) using their assay IDs as queries. **Table 1** shows the assay IDs of these datasets together with the Tox21 Challenge datasets. Details of the datasets, including the molecular targets, specific assays, number of samples screened, and number of samples deemed active, can be found from PubChem using the respective assay IDs as queries.

Because some samples were prepared from the same parent chemicals, we first cleaned the data before using them to evaluate the performance of the similarity search methods. We first removed counter-ions in salts and retained the largest component in samples consisting of non-bonded (i.e., disconnected) components. We then standardized the structures by neutralizing acids and bases (protonating acids and de-protonating bases) and generating a canonical SMILES from the standardized structure for each sample. For the results of each dataset, we applied a first-pass filter on canonical SMILES and retained only the first sample entry of a structurally unique parent compound. **Table 1** summarizes the resulting number of structurally unique parent compounds tested and the number of structurally unique actives from each assay.

In addition to the 24 HTS datasets, we also evaluated performance of the methods on 40 datasets in the Directory of Useful Decoys (DUD) (Huang et al., 2006) and an enhanced version of DUD consisting of 102 datasets called DUDE datasets (Mysinger et al., 2012). Each of these datasets consists of compounds known active on a protein target and many compounds of similar physicochemical properties as the actives but of very different molecular structures as the actives. These datasets are designed for evaluating the performance of docking-based virtual screening methods. We expect them to be less challenging than the HTS datasets for 2-D molecular similarity search methods, because in these datasets the actives and decoys are well-separated in molecular structure space and, therefore, any fingerprint-based similarity search methods are expected to perform well on these datasets.

Evaluation Protocol

To evaluate the performance of the methods, for each dataset we randomly selected 100 actives as the queries, and combined the other actives with the other compounds tested. We then calculated the maximum TC for each of these compounds with respect to the queries, as well as the maximum z-score and AS score. For these calculations, we used the extended connectivity fingerprint (Rogers and Hahn, 2010) with a maximum diameter of four chemical bonds (ECFP_4) and a fixed fingerprint length of 2,048 bits. We calculated ROC_AUC and BEDROC values for the Max-Sim, maxZ, and AS methods. For all BEDROC calculations, we used the default parameter setting of $\alpha = 20.0$, i.e., corresponding to 80% of the maximum contribution to BEDROC coming from the top 8% of the list of ranked molecules. To ensure statistical significance of the findings, we repeated the calculations nine times, using 100 randomly selected actives as queries each time. We compared the performance of the methods based on the resulting mean ROC_AUC and BEDROC values.

RESULTS

Performance of the maxZ Method

Table 2 shows a summary of the mean ROC_AUC and BEDROC values derived from the Max-Sim and maxZ methods for the 24

[2]https://tripod.nih.gov/tox21/challenge/data.jsp (accessed July 5, 2019).

TABLE 1 | PubChem datasets used in this study to evaluate performance of similarity search methods.

Dataset[a]	PubChem Assay ID[b]	Samples[c]	Unique structures[d]	Unique actives[e]
AHR	743122	8169	6318	723
AR	743040	9362	7009	290
ARE	743219	7167	5643	889
AR-lbd	743053	8599	6524	233
Aromatase	743139	7226	5601	273
ATAD5	720516	9091	6825	253
ER	743079	7697	5993	716
ER-lbd	743077	8753	6727	324
HSE	743228	8150	6253	337
MMP	720637	7320	5625	885
P53	720552	8634	6544	410
PPARg	743140	8184	6232	174
4-MU	589	59070	58199	6146
ALDH1A1	1030	220365	215450	15847
BRCA1	624202	377534	373883	3938
DNApb	485314	337903	333082	4466
ERK	1454	133383	130623	532
GCN5L2	504327	387577	379179	741
hERG	588834	5363	4568	553
Lucif	411	72335	70939	1558
MiRNAs	2289	336623	332205	3265
Mitoch	485298	322909	320471	734
NPC1	485313	321376	319001	7532
PR901	1347036	9523	7177	111

[a] *All datasets are derived from quantitative high throughput screening conducted at the National Center for Advancing Translational Sciences to ascertain chemical activities against different molecular targets. The first 12 datasets were used in the 2014 Tox21 Data Challenge.*
[b] *The datasets can be accessed from the PubChem website using the assay IDs as queries.*
[c] *Total number of samples screened in each dataset.*
[d] *Number of structurally unique parent molecules (non-salts, non-mixtures) derived from retaining the largest chemical structure in each sample and performing structure standardization.*
[e] *Number of structurally unique active parent molecules.*
Dataset names: AHR, activators of aryl hydrocarbon receptor; AR, activators of androgen receptor; AR-lbd, activators of androgen receptor ligand binding domain; Aromatase, aromatase inhibitors; ER, estrogen receptor activators; ER-lbd, activators of estrogen receptor ligand binding domain; PPARg, activators of peroxisome proliferator-activated receptor gamma; ARE, activators of antioxidant response element; ATAD5, ATPase family AAA domain-containing protein 5; HSE, activators of heat shock response signaling pathway; MMP, disruptors of mitochondrial membrane potential; p53, activators of p53 signaling pathway; hERG, blockers of hERG potassium channel; PR901, agonists of progesterone receptor; 4-MU, spectroscopic response at the 4-methylumbelliferone region as a counter assay for fluorescence detection; Lucif, inhibitors of Luciferase; ERK, inhibitors of mitogen-activated protein kinase 1; ALDH1A1, inhibitors of aldehyde dehydrogenase 1 family, member A1; NPC1, promoters of Niemann-Pick C1 protein precursor; Mitoch, inhibitors of mitochondrial division; MiRNAs, modulators of miRNAs; DNApb, inhibitors of DNA polymerase beta; BRCA1, activators of BRCA1 expression; GCN5L2, inhibitors of histone acetyltransferase KAT2A.

datasets. The mean ROC_AUC values derived from the Max-Sim method and those derived from the maxZ method were similar, with the latter only 3.7% higher than the former. In contrast, the mean difference in BEDROC values between the maxZ and Max-Sim methods was as high as 15%. However, the result for one dataset, NPC1, was an outlier, as the difference was as high as 170%, and the mean difference in BEDROC values decreased to 8.7% when it was excluded. Nonetheless, the maxZ method still outperformed the Max-Sim method, as the ROC_AUC and BEDROC values derived from maxZ were smaller than those derived from Max-Sim in only two of the 24 datasets. Although the differences between the maxZ and Max-Sim results were small for some datasets, for those showing a considerable difference, maxZ performed significantly better. For instance, the ROC_AUC values derived from maxZ were at least 5% higher than those derived from Max-Sim in 8 of the 24 datasets, whereas Max-Sim performed better than maxZ by 5% or more in only two datasets. This difference was even more pronounced for BEDROC values, as maxZ outperformed Max-Sim by 5% or more in 15 of the 24 datasets, whereas the opposite was true in only one dataset. Overall, the ROC_AUC values show that the maxZ method performs only slightly better than the Max-Sim method for ranking all samples in the dataset, whereas the BEDROC values indicate that the maxZ method performs markedly better than the Max-Sim method in the early recognition of active compounds.

A popular 3-D equivalent of 2-D fingerprint-based molecular similarity search is the Rapid Overlay of Chemical Structures (ROCS) method (OpenEye Scientific Software, Santa Fe, NM) (Fontaine et al., 2007), which calculates the Tanimoto similarity between 3-D molecular shapes and pharmacophore features. Because of the similarity between 2-D fingerprint-based and

TABLE 2 | Mean and standard deviation of ROC_AUC and BEDROC values derived from a similarity search using the rank by maximum similarity (Max-Sim) and maximum z-score (maxZ) approaches over 10 runs, each with 100 randomly selected actives as queries.

	ROC_AUC					BEDROC				
	Max-Sim		maxZ			Max-Sim		maxZ		
Dataset	Mean	Std	Mean	Std	%Diff[a]	Mean	Std	Mean	Std	%Diff[b]
AHR	0.754	0.011	0.759	0.012	0.7	0.402	0.016	0.365	0.019	−9.3
AR	0.740	0.009	0.748	0.009	1.0	0.482	0.019	0.509	0.014	5.6
ARE	0.539	0.008	0.544	0.009	1.0	0.225	0.014	0.266	0.018	17.9
AR-lbd	0.815	0.011	0.811	0.011	−0.5	0.607	0.023	0.610	0.028	0.6
Aromatase	0.662	0.017	0.686	0.016	3.6	0.263	0.026	0.291	0.019	10.7
ATAD5	0.713	0.022	0.724	0.018	1.6	0.303	0.025	0.313	0.020	3.4
ER	0.665	0.005	0.667	0.009	0.4	0.380	0.009	0.395	0.012	3.8
ER-lbd	0.715	0.011	0.726	0.010	1.5	0.381	0.026	0.382	0.025	0.0
HSE	0.579	0.027	0.595	0.027	2.8	0.186	0.020	0.205	0.023	10.3
MMP	0.694	0.009	0.704	0.018	1.4	0.414	0.024	0.469	0.025	13.2
P53	0.611	0.013	0.649	0.013	6.1	0.251	0.016	0.265	0.012	5.8
PPARg	0.681	0.018	0.686	0.027	0.7	0.274	0.031	0.277	0.030	1.0
4-MU	0.565	0.007	0.604	0.007	7.0	0.272	0.010	0.316	0.009	16.1
ALDH1A1	0.506	0.004	0.513	0.009	1.3	0.104	0.007	0.111	0.007	6.8
BRCA1	0.667	0.006	0.694	0.004	4.2	0.147	0.006	0.155	0.006	5.5
DNApb	0.591	0.011	0.633	0.012	7.1	0.137	0.006	0.163	0.012	18.9
ERK	0.647	0.021	0.698	0.017	8.0	0.235	0.017	0.250	0.016	6.3
GCN5L2	0.541	0.014	0.651	0.016	20.5	0.179	0.015	0.245	0.015	36.9
hERG	0.745	0.009	0.732	0.013	−1.8	0.460	0.009	0.447	0.012	−2.7
Lucif	0.707	0.008	0.737	0.010	4.1	0.255	0.010	0.268	0.012	4.9
MiRNAs	0.574	0.004	0.609	0.010	6.1	0.128	0.004	0.144	0.008	12.0
Mitoch	0.510	0.006	0.546	0.007	7.0	0.079	0.007	0.101	0.006	28.9
NPC1	0.653	0.007	0.696	0.007	6.5	0.079	0.007	0.213	0.006	170.0
PR901	0.931	0.013	0.945	0.013	1.5	0.800	0.025	0.822	0.022	2.8
Mean					**3.8**					**15.4**

[a] *Percent difference between mean ROC_AUC values for Max-Sim and maxZ methods.*
[b] *Percent difference between mean BEDROC values for Max-Sim and maxZ methods.*

3-D ROCS-based similarity searches, we hypothesized that the maxZ method would also improve early recognition for ROCS-based 3-D similarity searches. To test this hypothesis, we generated up to 15 low-energy conformers for each molecule in the 12 datasets used in the 2014 Tox21 Data Challenge, using Omega version 3.0.1.2 (OpenEye Scientific Software) with default parameters (Hawkins et al., 2010). We then conducted ROCS-based similarity searches for each dataset using randomly selected 10% actives as active queries. Each query molecule was represented by up to 15 of its lowest-energy conformers. We calculated the Tanimoto combo similarity (commonly called the combo score, which is the sum of the shape TC and color force field TCs) pairwise between the conformers of each active query and each conformer of the other compounds using ROCS version 3.2.2.2 with default parameters. The maximum Tanimoto combo score between a query molecule and a non-query molecule is designated as the Tanimoto combo score of the non-query molecule. We then calculated the ROC_AUC and BEDROC values using the maximum Tanimoto combo scores and the maximum z-scores calculated from the Tanimoto combo scores. We repeated this calculation nine more times, each with 10% of the actives randomly selected as active query molecules. **Table 3** shows the means and standard deviations of the ROC_AUC and BEDROC values. The results were similar to those of 2-D fingerprint-based similarity searches, i.e., the ROC_AUC values derived from the maxZ method were a few percentage points higher than those derived from the Max-Sim method, but the difference between BEDROC values was 12,6% on average. Thus, sorting the samples by the maximum z-values of the combo scores led to a substantial improvement in early recognition.

We evaluated the maxZ method for 3-D similarity search of the Tox21 Challenge datasets only, because the datasets were small (7,009 structurally unique compounds in the largest) and the computations could be completed within a reasonable amount of time. Most of the other datasets are much larger, with up to a few hundred thousand structurally unique compounds. We did not evaluate the performance of the maxZ method on these datasets, because the ROCS calculations would have required substantially more computing resources.

TABLE 3 | Mean and standard deviation of ROC_AUC and BEDROC values derived from a ROCS-based 3-D molecular similarity search using the rank by maximum similarity (Max-Sim) and maximum z-score (maxZ) methods.

	ROC_AUC					BEDROC				
	Max-Sim		maxZ			Max-Sim		maxZ		
Dataset	Mean	Std	Mean	Std	%Diff[a]	Mean	Std	Mean	Std	%Diff[b]
AHR	0.588	0.004	0.603	0.004	2.5	0.112	0.003	0.126	0.003	12.7
AR	0.729	0.007	0.749	0.015	2.7	0.201	0.010	0.234	0.013	16.2
ARE	0.547	0.004	0.560	0.005	2.5	0.090	0.003	0.101	0.002	12.6
AR-lbd	0.645	0.020	0.661	0.023	2.4	0.168	0.014	0.208	0.012	23.3
Aromatase	0.562	0.011	0.579	0.013	3.1	0.068	0.005	0.080	0.004	17.2
ATAD5	0.663	0.010	0.686	0.012	3.5	0.164	0.009	0.183	0.009	11.5
ER	0.685	0.007	0.703	0.009	2.6	0.168	0.007	0.184	0.005	10.1
ER-lbd	0.679	0.007	0.697	0.008	2.8	0.181	0.006	0.201	0.006	11.0
HSE	0.526	0.011	0.538	0.012	2.3	0.078	0.005	0.085	0.005	8.7
MMP	0.553	0.002	0.568	0.003	2.6	0.119	0.003	0.127	0.001	7.0
P53	0.575	0.009	0.590	0.011	2.5	0.091	0.006	0.101	0.004	11.2
PPARg	0.569	0.016	0.592	0.020	3.9	0.104	0.010	0.114	0.008	9.7
Mean					**2.8**					**12.6**

[a] *Percent difference between mean ROC_AUC values for Max-Sim and maxZ methods.*
[b] *Percent difference between mean BEDROC values for Max-Sim and maxZ methods.*

Performance of the AS Method

Table 4 shows the ROC_AUC and BEDROC values calculated from the Max-Sim and AS methods for the 12 Tox21 Challenges datasets. We calculated the AS score using a negative set (inactives) of 1,000 randomly selected compounds from the set of all screening negatives. We used the rest of the actives and inactives in each dataset as test data to evaluate the performance of each similarity search method. Both ROC_AUC and BEDROC values calculated from AS scores were significantly higher than the corresponding values obtained using the Max-Sim method, confirming that exploiting the available information on inactives improves the performance of both virtual screening methods. Note that the improvement of BEDROC values is significantly greater than that of ROC_AUC values, suggesting that the performance gains are mainly due to early recognition of actives in the AS method.

As noted in section Rank by Aggregated Similarity, because drug discovery involves rigorous confirmation of the activities of actives, but rarely any investments in efforts to confirm inactivity, information on inactive compounds is usually less reliable than that on active compounds. In addition, for some projects, active queries are not derived from screening of chemical libraries and, hence, there is no information on inactive compounds. However, because the number of compounds that are active against any drug target can be assumed to be miniscule compared to the number of all available compounds, we hypothesized that a large number of structurally diverse compounds should be able to serve as putative inactive compounds for the AS method. To test this hypothesis, we repeated the evaluation above, using 10,000 structurally diverse compounds selected from the NCI library. **Table 5** shows that the replacement of inactive compounds by structurally diverse compounds led to a significant performance deterioration of the AS method, especially in terms of ROC_AUC values, which were only 2.6% higher on average than those of the Max-Sim method. However, the overall mean BEDROC value was still 14.6% higher than that of the Max-Sim method, indicating that early recognition improved even when inactive compounds from screening were unavailable.

To assess the validity of the findings on the 12 Tox21 Challenge datasets for a wider range of datasets with the number of chemicals ranging from a few thousand to few hundred thousand, we conducted virtual screening using the AS method and the same 10,000 structurally diverse NCI compounds as putative inactive compounds. The results were comparable to those obtained from the Tox21 Challenge datasets, indicating that the method is applicable to a wide range of HTS datasets (**Table 6**).

Performance of the maxZ and AS Methods on the DUD and DUDE Datasets

Because most DUD and DUDE datasets contain <100 actives, we performed evaluations on these datasets by randomly selecting 10% of the actives as queries. We used the remaining actives and all decoys as test sets to evaluate the performance of the methods. As these datasets do not contain any experimentally determined inactives, we used the same set of 10,000 structurally diverse NCI compounds as putative inactives in testing the performance of the AS method. **Table S1** shows detailed results obtained from the 40 DUD and 102 DUDE datasets and **Table 7** summarizes these results.

The most obvious difference between the summary results in **Table 7** and the results in **Table 2** was that the ROC_AUC and BEDROC values of the HTS datasets derived from the Max-Sim method were significantly lower than the corresponding values of the DUD and DUDE datasets. The mean ROC_AUC

TABLE 4 | Mean and standard deviation of ROC_AUC and BEDROC values derived from a fingerprint-based similarity search using the rank by maximum similarity (Max-Sim) and rank by aggregated score (AS) methods.

	ROC_AUC					BEDROC				
	Max-Sim		AS			Max-Sim		AS		
Dataset	**Mean**	**Std**	**Mean**	**Std**	**%Diff[a]**	**Mean**	**Std**	**Mean**	**Std**	**%Diff[b]**
AHR	0.734	0.010	0.829	0.010	13.0	0.385	0.016	0.534	0.018	38.6
AR	0.749	0.013	0.809	0.013	8.0	0.443	0.020	0.618	0.019	39.7
ARE	0.537	0.008	0.650	0.011	21.0	0.264	0.021	0.404	0.015	53.2
AR-lbd	0.795	0.017	0.823	0.014	3.5	0.574	0.026	0.605	0.022	5.5
Aromatase	0.615	0.016	0.727	0.021	18.2	0.219	0.029	0.323	0.019	47.3
ATAD5	0.652	0.018	0.717	0.018	9.9	0.273	0.023	0.303	0.024	11.0
ER	0.602	0.024	0.683	0.024	13.6	0.234	0.016	0.439	0.039	87.7
ER-lbd	0.689	0.014	0.713	0.011	3.6	0.374	0.027	0.386	0.017	3.3
HSE	0.557	0.009	0.650	0.012	16.8	0.122	0.010	0.210	0.024	71.8
MMP	0.663	0.011	0.760	0.007	14.6	0.398	0.030	0.565	0.026	41.9
P53	0.588	0.020	0.723	0.013	23.0	0.224	0.017	0.262	0.020	16.9
PPARg	0.678	0.029	0.744	0.025	9.7	0.262	0.036	0.279	0.036	6.3
Mean					**12.9**					**35.3**

[a] *Percent difference between mean ROC_AUC values for Max-Sim and AS methods.*
[b] *Percent difference between mean BEDROC values for Max-Sim and AS methods.*

TABLE 5 | Mean and standard deviation of ROC_AUC and BEDROC values derived from a fingerprint-based similarity search using the rank by maximum similarity (Max-Sim) and rank by aggregated score (AS) methods, using 10,000 structurally diverse compounds as inactive compounds.

	ROC_AUC					BEDROC				
	Max-Sim		AS			Max-Sim		AS		
Dataset	**Mean**	**Std**	**Mean**	**Std**	**%Diff[a]**	**Mean**	**Std**	**Mean**	**Std**	**%Diff[b]**
AHR	0.730	0.010	0.758	0.009	3.8	0.337	0.023	0.407	0.017	20.8
AR	0.754	0.012	0.768	0.011	1.9	0.436	0.024	0.596	0.017	36.8
ARE	0.535	0.007	0.549	0.009	2.6	0.224	0.021	0.258	0.024	15.2
AR-lbd	0.790	0.018	0.807	0.021	2.2	0.550	0.020	0.614	0.030	11.6
Aromatase	0.621	0.024	0.663	0.023	6.7	0.202	0.040	0.289	0.036	43.2
ATAD5	0.653	0.024	0.666	0.025	2.0	0.268	0.021	0.281	0.023	4.9
ER	0.600	0.009	0.567	0.014	−5.5	0.197	0.010	0.136	0.021	−30.8
ER-lbd	0.683	0.018	0.712	0.017	4.2	0.366	0.030	0.461	0.019	25.8
HSE	0.553	0.013	0.576	0.016	4.2	0.118	0.018	0.147	0.026	24.4
MMP	0.651	0.016	0.689	0.012	5.9	0.336	0.029	0.454	0.019	35.0
P53	0.573	0.018	0.594	0.020	3.6	0.203	0.026	0.202	0.020	−0.2
PPARg	0.689	0.023	0.689	0.022	0.0	0.278	0.024	0.245	0.043	−12.0
Mean					**2.6**					**14.6**

[a] *Percent difference between mean ROC_AUC values for the Max-Sim and AS methods.*
[b] *Percent difference between mean BEDROC values for the Max-Sim and AS methods.*

values of the DUD and DUDE datasets were 0.90 and 0.96, respectively, and the corresponding mean BEDROC values were 0.76 and 0.90. These values were significantly higher than the corresponding mean ROC_AUC and BEDROC values of 0.66 and 0.29 for the 24 HTS datasets. These results corroborate our expectation that, because the actives and decoys are well-separated in molecular structure space, the DUD and DUDE datasets present much less of a challenge than do the HTS datasets for similarity search methods. Because the Max-Sim method achieved near perfect performance for these datasets, as indicated by an average ROC_AUC value of 0.96 and an average BEDROC value of 0.90 for the DUDE datasets, any improvement beyond the Max-Sim results will necessarily be small given the little room left for improvement. Indeed, **Table 7** shows that on average, the ROC_AUC or BEDROC values derived from the maxZ or AS method were only about 1%

TABLE 6 | Mean and standard deviation of ROC_AUC and BEDROC values derived from a fingerprint-based similarity search using the rank by maximum similarity (Max-Sim) and rank by aggregated score (AS) methods, using 10,000 structurally diverse compounds as inactive compounds.

	ROC_AUC					BEDROC				
	Max-Sim		AS			Max-Sim		AS		
Dataset	Mean	Std	Mean	Std	%Diff[a]	Mean	Std	Mean	Std	%Diff[b]
4-MU	0.513	0.007	0.525	0.016	2.3	0.143	0.012	0.175	0.014	22.1
ALDH1A1	0.510	0.004	0.519	0.004	1.8	0.142	0.004	0.144	0.003	1.0
BRCA1	0.643	0.008	0.657	0.008	2.3	0.143	0.008	0.146	0.009	2.1
DNApb	0.527	0.010	0.588	0.009	11.7	0.111	0.007	0.173	0.007	55.5
ERK	0.624	0.011	0.672	0.008	7.6	0.247	0.014	0.299	0.011	21.3
GCN5L2	0.542	0.015	0.542	0.017	−0.1	0.132	0.009	0.133	0.012	0.9
hERG	0.725	0.011	0.794	0.006	9.6	0.411	0.032	0.522	0.024	27.2
Lucif	0.735	0.009	0.782	0.006	6.3	0.285	0.015	0.335	0.010	17.5
MiRNAs	0.613	0.006	0.635	0.006	3.6	0.143	0.006	0.153	0.007	7.0
Mitoch	0.512	0.008	0.513	0.007	0.1	0.082	0.008	0.097	0.008	18.4
NPC1	0.681	0.005	0.705	0.007	3.6	0.222	0.005	0.242	0.008	9.0
PR901	0.896	0.019	0.901	0.020	0.5	0.718	0.030	0.757	0.037	5.5
Mean					**4.1**					**15.6**

[a] *Percent difference between mean ROC_AUC values for the Max-Sim and AS methods.*
[b] *Percent difference between mean BEDROC values for the Max-Sim and AS methods.*

TABLE 7 | Summary of the performance of similarity search methods on 40 DUD[a] and 102 DUDE[b] datasets.

Metric	Mean Max-Sim value[c]	Mean%diff[d]	Diff ≥ 0%[e]	Diff < 0%[f]	Diff ≥ 1%[g]	Diff < −1%[h]
MaxZ vs. Max-Sim method on 40 DUD datasets						
ROC_AUC	0.91	1.0	30	10	13	0
BEDROC	0.79	1.7	30	10	16	1
AS vs. Max-Sim method on 40 DUD datasets						
ROC_AUC	0.90	1.2	28	12	15	3
BEDROC	0.76	1.1	25	15	16	10
MaxZ vs. Max-Sim method on 102 DUDE datasets						
ROC_AUC	0.96	0.5	90	12	18	0
BEDROC	0.90	0.7	87	15	28	2
AS vs. Max-Sim method on 102 DUD datasets						
ROC_AUC	0.96	0.2	55	47	13	3
BEDROC	0.90	0.2	58	44	27	20

[a] *DUD: Directory of Useful Decoys, http://dud.docking.org/.*
[b] *DUDE: Database of Useful Decoys: Enhanced, http://dude.docking.org/.*
[c] *Mean ROC_AUC or BEDROC value calculated from the Max-Sim method over 40 DUD or 102 DUDE datasets.*
[d] *Mean percentage difference between ROC_AUC or BEDROC values derived from the maxZ or AS methods and the Max-Sim method.*
[e] *Number of datasets for which ROC_AUC or BEDROC values calculated from the maxZ or AS methods were higher than or equal to the corresponding values calculated from the Max-Sim method, i.e., the number of datasets on which the maxZ or AS method performed comparable to or better than the Max-Sim method did.*
[f] *Number of datasets for which ROC_AUC or BEDROC values calculated from the maxZ or AS methods were lower than the corresponding values calculated from the Max-Sim method, i.e., the number of datasets on which the maxZ or AS method performed worse than the Max-Sim method.*
[g] *Number of datasets for which ROC_AUC or BEDROC values calculated from the maxZ or AS methods were at least 1% higher than the corresponding values calculated from the Max-Sim method.*
[h] *Number of datasets for which ROC_AUC or BEDROC values calculated from the Max-Sim method was more than 1% higher than the corresponding values calculated from the maxZ or AS methods.*

higher and <1% higher than the corresponding values of the Max-Sim method for the DUD and DUDE datasets, respectively. Nevertheless, **Table 7** shows that the number of datasets for which the maxZ and AS methods performed better than the Max-Sim method by more than 1% was significantly higher than that for which the Max-Sim method performed better by more than 1%. Thus, for these datasets, the maxS and AS methods still outperformed the Max-Sim method (albeit with a smaller effect size) even though the Max-Sim method already achieved near perfect performance.

DISCUSSION

Fingerprint-based molecular similarity search is one of the most important tools for virtual screening of large chemical libraries. Over the years, many similarity search methods have

been investigated, but the simple, parameter-free, rank-by-maximum Tanimoto similarity approach remains a popular method. It achieves robust performance based on the Tanimoto similarity of each compound in a compound library to its closest query molecule and disregarding its similarity to all other query molecules. In addition, it compares the values of Tanimoto similarity to different query molecules directly. This is theoretically correct only if the distribution of similarity values to all other query molecules is identical, an assumption that has been shown to be invalid (Baldi and Nasr, 2010).

In this study, we proposed and evaluated two parameter-free similarity search methods. The AS method considers information on the similarity to all query molecules, whereas the maxZ method converts the Tanimoto similarity into a z-score for a statistically sound, direct comparison. The results of our evaluations using over 20 HTS datasets indicated that neither method achieved significantly higher ROC_AUC values over the standard Max-Sim method. However, BEDROC values derived from both methods were ∼10% higher than those of the Max-Sim method. Thus, although our methods perform comparably to the standard similarity search method when judged by ranking all compounds in a screening library, they perform better on early recognition by placing more actives at the top of a ranked list. This is an important trait for virtual screening of large chemical libraries, considering that follow-up experimental testing is feasible for only a small fraction of chemicals.

A conventional similarity search calculates TCs between all query molecules and all library molecules, and these values are sufficient for converting TCs to z-scores. As such, the additional computational cost to perform a similarity search using the maxZ method is minimal. Conversely, the AS method is notably slower than the Max-Sim method. However, with the ever-increasing power and decreasing cost of computing hardware, the method can become competitive based on its performance. Thus, when early recognition is among the objectives of virtual screening, the two methods provide alternatives to the standard Max-Sim method.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception of the ideas and planned computational evaluations. RL performed the evaluations and wrote the first draft of the manuscript. AW supervised the project and contributed to the final manuscript.

FUNDING

The research was supported by the U.S. Army Medical Research and Development Command (Ft. Detrick, MD) as part of the U.S. Army's Network Science Initiative, and by the Defense Threat Reduction Agency (Grant No. CBCall14-CBS-05-2-0007).

ACKNOWLEDGMENTS

The authors gratefully acknowledge the assistance of Dr. Tatsuya Oyama in editing the manuscript. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army, the U.S. Department of Defense, or The Henry M. Jackson Foundation for Advancement of Military Medicine, Inc.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00701/full#supplementary-material

REFERENCES

Aherne, G. W., McDonald, E., and Workman, P. (2002). Finding the needle in the haystack: why high-throughput screening is good for your health. *Breast Cancer Res.* 4, 148–154. doi: 10.1186/bcr440

Baldi, P., and Nasr, R. (2010). When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values. *J. Chem. Inf. Model.* 50, 1205–1222. doi: 10.1021/ci100010v

Empereur-Mot, C., Guillemain, H., Latouche, A., Zagury, J. F., Viallon, V., and Montes, M. (2015). Predictiveness curves in virtual screening. *J. Cheminform.* 7:52. doi: 10.1186/s13321-015-0100-8

Fontaine, F., Bolton, E., Borodina, Y., and Bryant, S. H. (2007). Fast 3D shape screening of large chemical databases through alignment-recycling. *Chem. Cent. J.* 1:12. doi: 10.1186/1752-153X-1-12

Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* 47, 1750–1759. doi: 10.1021/jm030644s

Hawkins, P. C., Skillman, A. G., Warren, G. L., Ellingson, B. A., and Stahl, M. T. (2010). Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* 50, 572–584. doi: 10.1021/ci100031x

Huang, N., Shoichet, B. K., and Irwin, J. J. (2006). Benchmarking sets for molecular docking. *J. Med. Chem.* 49, 6789–6801. doi: 10.1021/jm0608356

Huang, R., and Xia, M. (2017). Editorial: Tox21 Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental toxicants and drugs. *Front. Environ. Sci.* 5:3. doi: 10.3389/fenvs.2017.00003

Kairys, V., Fernandes, M. X., and Gilson, M. K. (2006). Screening drug-like compounds by docking to homology models: a systematic study. *J. Chem. Inf. Model.* 46, 365–379. doi: 10.1021/ci050238c

Kar, S., and Roy, K. (2013). How far can virtual screening take us in drug discovery? *Expert Opin. Drug Discov.* 8, 245–261. doi: 10.1517/17460441.2013.761204

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., et al. (2016). PubChem substance and compound databases. *Nucleic Acids Res.* 44, D1202–D1213. doi: 10.1093/nar/gkv951

Kristensen, T. G., Nielsen, J., and Pedersen, C. N. (2013). Methods for similarity-based virtual screening. *Comput. Struct. Biotechnol. J.* 5:e201302009. doi: 10.5936/csbj.201302009

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Med. Decis. Making* 9, 190–195. doi: 10.1177/0272989X8900900307

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55, 6582–6594. doi: 10.1021/jm300687e

Nasr, R. J., Swamidass, S. J., and Baldi, P. F. (2009). Large scale study of multiple-molecule queries. *J. Cheminform.* 1:7. doi: 10.1186/1758-2946-1-7

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi: 10.1021/ci100050t

Sheridan, R. P., Singh, S. B., Fluder, E. M., and Kearsley, S. K. (2001). Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J. Chem. Inf. Comput. Sci.* 41, 1395–1406. doi: 10.1021/ci0100144

Shiryaev, S. A., Cheltsov, A. V., Gawlik, K., Ratnikov, B. I., and Strongin, A. Y. (2011). Virtual ligand screening of the National Cancer Institute (NCI) compound library leads to the allosteric inhibitory scaffolds of the West Nile Virus NS3 proteinase. *Assay Drug Dev. Technol.* 9, 69–78. doi: 10.1089/adt.2010.0309

Triballeau, N., Acher, F., Brabet, I., Pin, J. P., and Bertrand, H. O. (2005). Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* 48, 2534–2547. doi: 10.1021/jm049092j

Truchon, J. F., and Bayly, C. I. (2007). Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* 47, 488–508. doi: 10.1021/ci600426e

Zhao, W., Hevener, K. E., White, S. W., Lee, R. E., and Boyett, J. M. (2009). A statistical framework to evaluate virtual screening. *BMC Bioinformatics* 10:225. doi: 10.1186/1471-2105-10-225

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Liu, AbdulHameed and Wallqvist. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 29 October 2019
doi: 10.3389/fchem.2019.00709

Fast Screening of Inhibitor Binding/Unbinding Using Novel Software Tool CaverDock

Gaspar P. Pinto[1,2†], *Ondrej Vavra*[1,2†], *Jiri Filipovic*[3], *Jan Stourac*[1,2], *David Bednar*[1,2*] and *Jiri Damborsky*[1,2*]

[1] Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno, Czechia, [2] International Centre for Clinical Research, St. Anne's University Hospital Brno, Brno, Czechia, [3] Institute of Computer Science, Masaryk University, Brno, Czechia

Protein tunnels and channels are attractive targets for drug design. Drug molecules that block the access of substrates or release of products can be efficient modulators of biological activity. Here, we demonstrate the applicability of a newly developed software tool CaverDock for screening databases of drugs against pharmacologically relevant targets. First, we evaluated the effect of rigid and flexible side chains on sets of substrates and inhibitors of seven different proteins. In order to assess the accuracy of our software, we compared the results obtained from CaverDock calculation with experimental data previously collected with heat shock protein 90α. Finally, we tested the virtual screening capabilities of CaverDock with a set of oncological and anti-inflammatory FDA-approved drugs with two molecular targets—cytochrome P450 17A1 and leukotriene A4 hydrolase/aminopeptidase. Calculation of rigid trajectories using four processors took on average 53 min per molecule with 90% successfully calculated cases. The screening identified functional tunnels based on the profile of potential energies of binding and unbinding trajectories. We concluded that CaverDock is a sufficiently fast, robust, and accurate tool for screening binding/unbinding processes of pharmacologically important targets with buried functional sites. The standalone version of CaverDock is available freely at https://loschmidt.chemi.muni.cz/caverdock/ and the web version at https://loschmidt.chemi.muni.cz/caverweb/.

Keywords: binding, docking, channel, unbinding, virtual screening, inhibitors, substrates, tunnel

OPEN ACCESS

Edited by:
Kamil Kuca,
University of Hradec Králové, Czechia

Reviewed by:
Ariel Fernandez,
National Council for Scientific and Technical Research (CONICET), Argentina
Dr. Rajeev K. Singla,
K.R. Mangalam University, India

****Correspondence:***
David Bednar
222755@mail.muni.cz
Jiri Damborsky
jiri@chemi.muni.cz

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *16 June 2019*
Accepted: *09 October 2019*
Published: *29 October 2019*

Citation:
Pinto GP, Vavra O, Filipovic J, Stourac J, Bednar D and Damborsky J (2019) Fast Screening of Inhibitor Binding/Unbinding Using Novel Software Tool CaverDock. Front. Chem. 7:709. doi: 10.3389/fchem.2019.00709

INTRODUCTION

Until the beginning of the new millennium, drug design mostly relied on experimental high-throughput screening (Kansy et al., 1998; Zhang et al., 1999; Bleicher et al., 2003). These techniques evolved rapidly up to the beginning of the nineties. However, although at that time they seemed promising and the best techniques for drug design and discovery, they were expensive in both time and labor (Bajorath, 2002; Bielska et al., 2014). More cost-effective methods emerged with the introduction of docking algorithms and thorough analysis of protein-ligand interactions. This boom in docking approaches led to the development of over 60 software tools for docking (Sousa et al., 2010; Pagadala et al., 2017). At the beginning of the new millennium, a new technique for drug design called "virtual screening" started to gain recognition (Clark, 2008; Ripphausen et al., 2010).

Virtual screening is now a well-established technique for drug design (Bottegoni et al., 2016), both in academic research and the pharmaceutical industry (Mangoni et al., 1999; Clark, 2008; Huang et al., 2008; Totrov and Abagyan, 2008; Cheng et al., 2012; Kaczor et al., 2016). Many docking programs are available for virtual screening and several comparisons and benchmarks have been published (Cummings et al., 2005; Cross et al., 2009; Lavecchia and Di Giovanni, 2013; Bielska et al., 2014; Chaput et al., 2016; Kim et al., 2016). These programs help in the first step of the drug design process and follow a general protocol of screening a large database of small compounds on a chosen target (receptor). After selecting a target, a library of ligands is chosen. The ligands can be taken from many publically available or commercial libraries. Of these, ZINC (http://zinc15.docking.org/) (Sterling and Irwin, 2015), ChEMBL (https://www.ebi.ac.uk/chembl/) and PubChem (https://pubchem.ncbi.nlm.nih.gov/) are among the most widely used and largest ligand libraries. However, other databases with fewer compounds may be useful when searching for compounds with specific characteristics. An example is Drugbank (https://www.drugbank.ca/) (Law et al., 2014), which is a database of drugs approved by the FDA and Canadian Agency for Drugs and Technologies in Health. It also enables selection of experimental, investigational and illicit drugs.

The success of virtual screening boosted the development of techniques used for drug design and in recent years, binding kinetics has gained increased momentum in the drug design community. A research program supported by the European Innovative Medicine Initiative has, for the last 6 years, focused on understanding target binding kinetics (Laverty et al., 2012; Goldman et al., 2013; Kush and Goldman, 2014). Although there has been a steep rise in the development of methods for drug design, there is still space for further improvements.

The binding of a substrate and release of the products of an enzymatic reaction have been studied using different computational approaches (Straatsma and McCammon, 1992; Kollman, 1993; Lamb and Jorgensen, 1997). Classical and accelerated molecular dynamics simulations have been used to calculate ligand binding affinities. These methods use the free energy perturbation approach to calculate the relative binding free energy between a receptor and two ligands based on the thermodynamic cycle (Kruse et al., 2012; Tomić et al., 2015). However, such methods are computationally demanding and not suitable for screening large libraries. Development of new approaches for analysis of ligand binding and unbinding is clearly needed.

Several computational tools have been developed for searching the best binding positions in the active site pocket of a target molecule and then binding positions with increasing distance from the active site. PELE is a web server that incorporates a wide range of different types of calculation, including protein local motions (Lucas and Guallar, 2012). PELE also enables ligand binding refinement, binding site searches and ligand migration. The latter three scripts yield multiple-pose docking results through all protein free space, which cannot be achieved with simple docking algorithms (Guallar et al., 2009; Hernández-Ortega et al., 2011; Espona-Fiedler et al., 2012; Madadkar-Sobhani and Guallar, 2013). MOMA-LigPath (Devaurs et al., 2013) has a robotics algorithm for space search, not only in the active site pocket but also along an unbinding trajectory. However, as the tool does not output information on the energy of conformations, it is not possible to prioritize individual pathways. SLITHER (Lee et al., 2009) is a web server built to generate conformations of substrates while traveling through membrane channels. It is based on both the AUTODOCK (Morris et al., 2009) and MEDock (Chang et al., 2005) docking algorithms. Energetic information is available from these calculations. However, docking trajectories are often sparse.

We have developed a fast method based on analysis of protein tunnels (Marques et al., 2017) combined with molecular docking in a single implementation—called CaverDock—and used it to address important biochemical problems. Protein tunnels are structural features connecting the buried active site cavities with the protein surface. First, tunnels in proteins are identified using the specialized software Caver (Chovancova et al., 2012). Then, an extensively optimised version of AutoDock Vina (Trott and Olson, 2010) is used to dock a ligand along the tunnel to produce a continuous trajectory. Algorithms implemented in CaverDock (Filipovic et al., 2019; Vavra et al., 2019) can be used to run a virtual screening protocol for binding a library of ligands into and from the active site. This procedure identifies energetically favorable binding sites located outside the active site, providing a profile of potential energies. The goal of CaverDock, in current implementation, is not the calculation of the free energy of binding. Instead of obtaining several trajectories to calculate the free energy (Jarzynski, 1997; Fernández, 2014), CaverDock calculates the binding energy along the several, predetermined, points along the tunnel.

We have utilized the new CaverDock tool in three applications. The first examined differences between substrates and inhibitors and selection of flexible side chains along tunnels bottlenecks, which serve as potential hot spots for mutagenesis. The datasets used for testing of the flexible simulations consisted of seven proteins with six tunnels and one channel: (i) cytochrome P450 17A1, (ii) leukotriene A4 hydrolase/aminopeptidase, (iii) acetylcholinesterase (AChE), (iv) human plasma cholesteryl ester transfer protein (CETP), (v) inducible nitric oxide synthase (iNOS), (vi) UDP-3-O-N-acetylglucosamine deacetylase (LpxC), and (vii) matrix metalloproteinase-13 (MMP-13). Trajectories were calculated for both the natural substrates and inhibitors. The second application was the study of human N-terminal domain of heat shock protein 90α (N-HSP90), an important pharmaceutical cancer target, with a diverse set of inhibitors. The dataset was obtained from previously published study (Kokh et al., 2018). We compared the resulting conformations from CaverDock with positions of inhibitor molecules found in the crystal structures. Furthermore, we analyzed the correlations between CaverDock energies and measured experimental values (Kokh et al., 2018). The third application was the screening of potential inhibitors and identification of the access pathways through simulation of binding processes. The applicability of CaverDock for virtual screening pharmaceutically important molecules was validated with cytochrome P450 17A1 and a dataset of oncological drugs from the NIH.gov website and with leukotriene A4 hydrolase/aminopeptidase and a dataset of anti-inflammatory

drugs from the drugbank.ca website. The presented results demonstrate that CaverDock is a ready-to-use tool that should be of broad interest to biochemists, protein engineers, and medicinal chemists.

METHODS

Protein Targets

Cytochrome P450 17A1 and leukotriene A4 hydrolase/aminopeptidase were selected for flexibility testing as well as the model systems to validate the applicability of CaverDock for the virtual screening of ligand libraries. Seven protein targets were considered, as described below. The description of the structure and function of Acetylcholinesterase (AChE), Cholesteryl ester transfer protein (CETP), Nitric oxide synthase (iNOS), Metal-dependent deacetylase (LpxC), and Matrix metalloproteinase-13 (MMP-13) is provided in the **Supplementary Information**.

Cytochrome P450 17A1 functions as a drug-processing enzyme and was selected as the target protein for both application studies. The starting structure for this work was the crystal structure taken from the Protein Data Bank (Berman et al., 2000) with PDB-ID 3RUK (DeVore and Scott, 2012). The structure comprised an agglomerate of 4 cytochromes P450 17A1, from which we only used chain A. The structure also contained the inhibitor Abiraterone, which blocked access to the active site and was deleted prior to CaverDock screening.

Leukotriene A4 hydrolase/aminopeptidase, with crystal structure PDB-ID 4L2L (Stsiapanava et al., 2014), was selected as the second target for both application studies. Leukotrienes are a family of lipid mediators that play important roles in a variety of allergic and inflammatory reactions (Haeggström et al., 1990; Funk, 2001; Haeggström, 2004; Szul et al., 2016). Leukotriene A4 hydrolase/aminopeptidase (EC 3.3.2.6) is a bifunctional zinc metalloenzyme that catalyzes formation of the chemotactic agent LTB4, a key lipid mediator in the immune response. This enzyme, had an inhibitor, 4-(4-benzylphenyl) thiazol-2-amine, bound to the crystal structure, which had to be removed prior to screening. LTA4H possesses two known activities, both of which are exerted via distinct but overlapping active sites and depend on a catalytic zinc atom. The catalytic zinc atom is bound to the signature HEXXH, known also for other M1 metallopeptidases (Gomis-Rüth et al., 2012; Zhang et al., 2015).

Heat shock protein 90α (HSP90) is a chaperone protein that assists the folding of client proteins. The HSP90 consists of three domains. The highly conserved N-terminal domain with ATP-binding cleft which is responsible for the catalytic activity. The middle domain contains a large hydrophobic surface needed for the folding of client proteins. The C-terminal domain is involved in the dimerization of HSP90 (Li et al., 2012). The function of HSP90 is linked to hydrolysis of ATP and the dimerization. A number of the HSP90 client proteins are part of cancer cell-associated signaling pathways, therefore the HSP90 is an important target in drug design. The function of HSP90 can be blocked by small molecules. This inhibition leads to degradation of the client proteins and impacts tumor growth (Kabakov et al., 2010). In this study, we analyzed the bound (HOLO) crystal structures with several small inhibitors inside the ATP-binding pocket. Furthermore, we conducted CaverDock simulations with a larger set of inhibitors using the unbound (APO) crystal structure of N-HSP90 (Kokh et al., 2018).

Structural Analysis of N-HSP90 HOLO Complexes

We studied the ability of CaverDock to find protein-ligand conformations similar to the crystal structures using the set of previously published complexes (Kokh et al., 2018). We analyzed the 34 crystal complexes of the N-HSP90 with different co-crystallized inhibitors. The list of the PDB IDs is in **Supplementary Table S1**. The crystal structures were aligned by DeepAlign (Wang et al., 2013) to simplify the following analyses. The tunnels for CaverDock runs were calculated by Caver 3.02 (Chovancova et al., 2012) in each inhibitor-free structure starting from the catalytic residues 93 and 138 with the probe radius, shell radius and shell depth set to 1.5, 20, and 20 Å, respectively. The tunnel leading through the main opening of the ATP-binding cleft to the active site was selected, discretized with 0.3 Å steps and extended by 20 Å to ensure complete unbinding of the tested inhibitor molecules. The receptor and ligand PDBQT files for CaverDock were prepared by MGLtools (Morris et al., 2009).

Energy Analysis of N-HSP90 HOLO and APO Complexes

Based on the previously published kinetic data (Kokh et al., 2018), we prepared two datasets. The first dataset consists of a subset of 32 inhibitors and HOLO structures from the HOLO structure analysis dataset described above. The kinetic data for two inhibitors (compound_01 and compound_04) was not complete in the original publication. The second dataset was created to check the findings from the HOLO dataset. It consists of 68 inhibitors. In this case, we ran the CaverDock calculations with the APO structure of N-HSP90 (PDB ID 3T0H). The CaverDock calculations were carried out in the same manner as described above for the structural analysis.

Libraries of Small Ligands

Several approaches can be used to choose libraries for virtual screening. For instance, screening as many ligands as possible from a broad dataset of molecules, such as the ZINC database. Another approach is to screen for drug-like compounds with specific biological activities. Virtual screening may also be performed using cognate ligands belonging to a group of compounds that the enzyme naturally binds and/or catalyzes. In the present study, we conducted a virtual screening campaign on a group of drug-like molecules possessing predefined biological activities. The chosen ligands were converted to the AutoDock Vina-compatible PDBQT format using MGLTools v1-5-7rc1 (Morris et al., 2009). We used the inhibitors complexed in the structures for validation of flexible side chains (inhibitor dataset). We built the substrates in Avogadro and minimized them with the UFF forcefield (Hanwell et al., 2012) for the natural substrates (substrate dataset).

A dataset of *133 cancer drugs* was downloaded from the NIH.gov website for the cytochrome P450 17A1. The drugs were

all FDA-approved and have been used against different types of cancer. Of the 133 drugs, 105 were used for the screening and 28 were excluded due to being salts or having unconventional atoms that could not be properly handled by AutoDock Vina. Among the 28 excluded drugs, 22 had two ligand molecules in the same file. The other six drugs had some atoms (one with arsenic, three with platinum and two with boron) for which there were no parameters available in the force field of AutoDock Vina. A dataset of *56 anti-inflammatory drugs* was downloaded from the drugbank.ca website for leukotriene A4 hydrolase/aminopeptidase. Out of these 56 drugs, 54 were used and 2 were excluded from the screening. One excluded drug contained a gold atom, for which AutoDock Vina had no defined parameters. The other drug was a silicate mineral with two molecules in the same file.

CaverDock Calculation

The software tool CaverDock is available free of charge at the website https://loschmidt.chemi.muni.cz/caverdock/. The CaverDock protocol (**Figure 1**) starts with finding the tunnels by using Caver (Chovancova et al., 2012). Caver can be used as a standalone program or as a Pymol plugin. The active site is selected as a starting point for the Caver calculation. For all seven proteins, the location of ligand binding in the active site was known (Gerber et al., 1997; Funk, 2001; Thunnissen et al., 2001; Epps and Vosters, 2002; Rudberg et al., 2002; Haeggström et al., 2007; Gattis et al., 2010; Singh and Konwar, 2012; Clayton et al., 2013; Khatri et al., 2014; Yoshimoto and Auchus, 2015). All the other settings were fixed at default values. This step of the protocol yields several tunnels for each protein. The numbering of the tunnels is given by a parameter called priority, which is the ratio between (i) bottleneck radius, (ii) tunnel length and (iii) curvature of the tunnel. The tunnels were represented as a series of sequential spheres.

We only used tunnels with priority 1 for the flexible simulations. These tunnels had the inhibitor already inside, although we removed it before the Caver calculation since we did not relax the protein in any way. Hence, there was an implicit bias for these tunnels. The flexibility in CaverDock arises from the already implemented flexibility capabilities of Autodock Vina. Flexibility on side chains was introduced in three iterations. In the default mode, CaverDock adds flexibility to two residues in each iteration, up to three iterations. These values may be changed by the user to better fit their needs. For each tunnel in the substrate and inhibitor datasets, two flexible bottleneck residues were added in each iteration. These flexible residues were not necessarily the same for the substrate and inhibitor simulations. Since substrates and inhibitors may differ in length and volume, the bottlenecks that they encounter along a tunnel may also differ.

For cytochrome P450 17A1, we used three tunnels for our virtual screening study. The first two tunnels found by Caver were also described in the literature, whereas the third tunnel was ranked as #5 by Caver. By individually inspecting every tunnel, we noted that tunnels ranked #3 and #4 by Caver were too long and narrow to be feasible as a ligand access pathway. For leukotriene A4 hydrolase/aminopeptidase, we used two tunnels ranked #1 and #2 by Caver. The results obtained were consistent with the literature (Cui et al., 2015), confirming that these two tunnels were used by the protein to transport ligands/drugs in and products out. Since the active site in leukotriene A4 hydrolase/aminopeptidase is inserted deeper into the protein and the protein itself is packed closer together than in cytochrome P450 17A1, only six tunnels were described for this protein vs. 15 tunnels described for cytochrome P450 17A1. A literature search showed that the tunnels ranked highest by Caver were indeed tunnels used by the natural substrate and inhibitors (Yu et al., 2013; Stsiapanava et al., 2014).

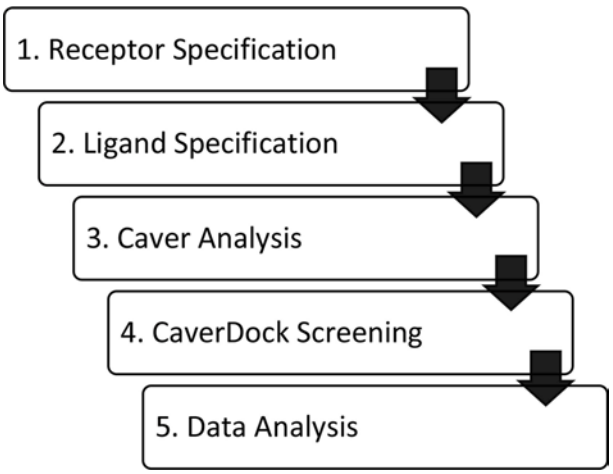


FIGURE 1 | Workflow of virtual screening using CaverDock. (1) Receptor and (2) ligand specification follow established protocols of the software tool AutoDock (Morris et al., 2009). (3) Identification of protein tunnels using Caver (Chovancova et al., 2012). (4) Tunnel discretization and sequential ligand binding study using CaverDock (Filipovic et al., 2019; Vavra et al., 2019). (5) Analysis of docking trajectories and energy profiles, extraction of energy barriers and protein-ligand complexes possessing the lowest energies.

After selecting the tunnels to study, the next step in a CaverDock protocol is to discretize the tunnels. Tunnel discretization divides each tunnel into a set of discs. The ligand is *glued* to a disc by one of its atoms and as the disc moves through the tunnel, the software defines a ligand path coordinate. After discretization, we extended the tunnels by two Ångströms. This step ensured that the tunnels were long enough to enable identification of the local binding minima at the tunnel mouth. Having prepared the tunnels, we used MGL tools to set the AutoDock atom types and Gasteiger charges for the receptor and ligands. MGL tools provide scripts that convert pdb and mol2 files into pdbqt file format. Having prepared the receptors and ligands, we next prepared a CaverDock file to run the docking step. This file was equivalent to the one used by AutoDock Vina but with the path to the file containing tunnel information instead of the receptor (Trott and Olson, 2010). We then added information about the studied tunnel from Caver to the configuration file. This new information allowed the "docking" conformation to be searched along the tunnel on each disc. One configuration file needs to be created for each ligand. **Figure 2** shows a representation of a ligand bound along a tunnel taken from several snapshots of a CaverDock calculation.

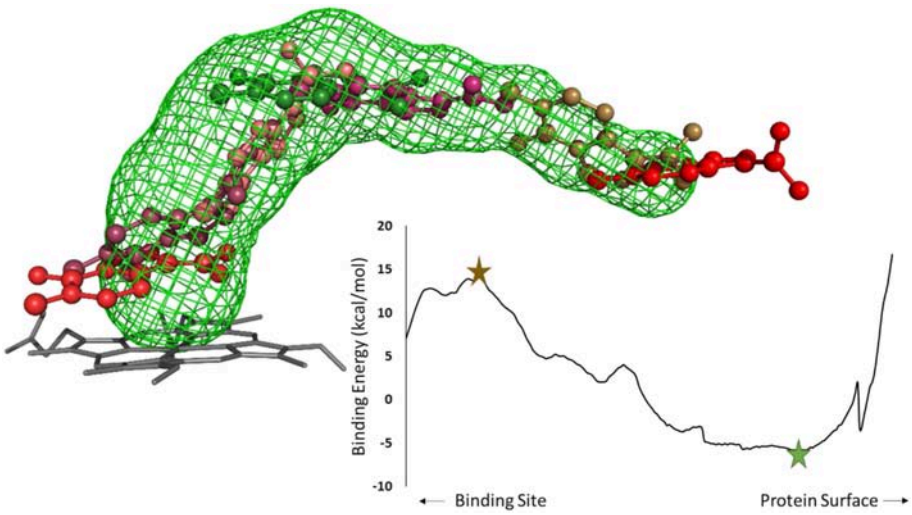


FIGURE 2 | Analysis of an inhibitor passing through a tunnel using CaverDock. Left: structure of cytochrome P450 17A1 and tunnel #3 with seven (7/254) snapshots of a ligand bound along the tunnel. The minimum binding energy conformation is shown in green, whereas the maximum binding energy is shown in brown, in a ball-and-stick representation. Right: energy profile obtained from the CaverDock calculation for cytochrome P450 17A1 and its tunnel #3 with the ligand.

An AutoDock Vina virtual screening was also performed with the same targets as the CaverDock virtual screening. To ensure that AutoDock Vina yielded the best result possible within a reasonable time, we used an exhaustiveness setting of 30. The center of the matrix grid was the same as that used for the CaverDock calculation and the box was 27 Å on each side.

RESULTS

Simulations With Flexible Side Chains

CaverDock allows flexibility of residue side chains along a tunnel. We tested the intrinsic flexibility of AutoDock Vina implemented in CaverDock with the substrate and inhibitor datasets. We introduced flexibility in three iterations by adding two flexible bottleneck residues in each iteration. Thus, iteration 1, 2, and 3 had two, four and six flexible side chains, respectively. The energetic barrier for the substrate and inhibitor to travel from the inside to outside of cytochrome P450 17A1 or leukotriene A4 hydrolase/aminopeptidase was lowered when we added flexible side chains (**Figure 3**). For cytochrome P450 17A1, the binding energy was lowered with each iteration for both the inhibitor and substrate. In the case of the substrate of leukotriene A4 hydrolase/aminopeptidase, the energetic barrier was stabilized with only two flexible side chains and addition of further flexible side chains gave no apparent change in the energetic barrier along the trajectory of the substrate. On the other hand, the inhibitor of leukotriene A4 hydrolase/aminopeptidase had a lower energetic barrier with each iteration. As expected, the inhibitor had a similar or more stable binding energy when compared to the substrate.

We showed that the flexible simulations were able to open parts of the tunnel with high barriers with the substrate and inhibitor datasets (electronic SI). Significant energetic barriers were lost in the iteration with six flexible residues. In this scenario, the ligand was able to leave without any spatial or energetic hindrance. The flexible side chains moved out of the way to let the ligand escape, but the new conformations of side chains were close to the rest of the protein structure. Adding flexible residues did not affect the energetic barrier in iNOS, which showed a similar profile through all iterations in the ligand simulations (electronic SI). In these cases, the tunnel radius was already large enough for unrestricted ligand exchange with no obvious bottleneck.

The usage of the intrinsic AutoDock Vina flexibility in CaverDock is still under development and new algorithms are being tested for obtaining better results. With the current version, users are advised to use both rigid and flexible simulations with four or less flexible side chains. We can get more information about the tunnel with the flexible side chains, e.g., to identify which residues need to be flexible to open the tunnel for ligand passage since these residues are natural hot-spots for potential mutagenesis. However, there is an obvious computational price to pay when using flexible simulations, as shown in **Table 1**. In particular, adding flexible residues leads to longer simulation times. We advise running CaverDock simulations with lower bound trajectories only when running in a rigid mode because rigid trajectories may yield unrealistic high barriers when running upper bound simulations (discussed below).

Comparison of Calculated and Experimental Results

Structural Analysis of HOLO Structures

We calculated the RMSD between the positions of bound inhibitors and the lower-bound CaverDock snapshots. We report the lowest RMSDs and the RMSDs for the lowest energy conformations in **Supplementary Table S1**. Validation of CaverDock in terms of reproducibility of experimental structures of enzyme-inhibitor complexes revealed that the tool identified

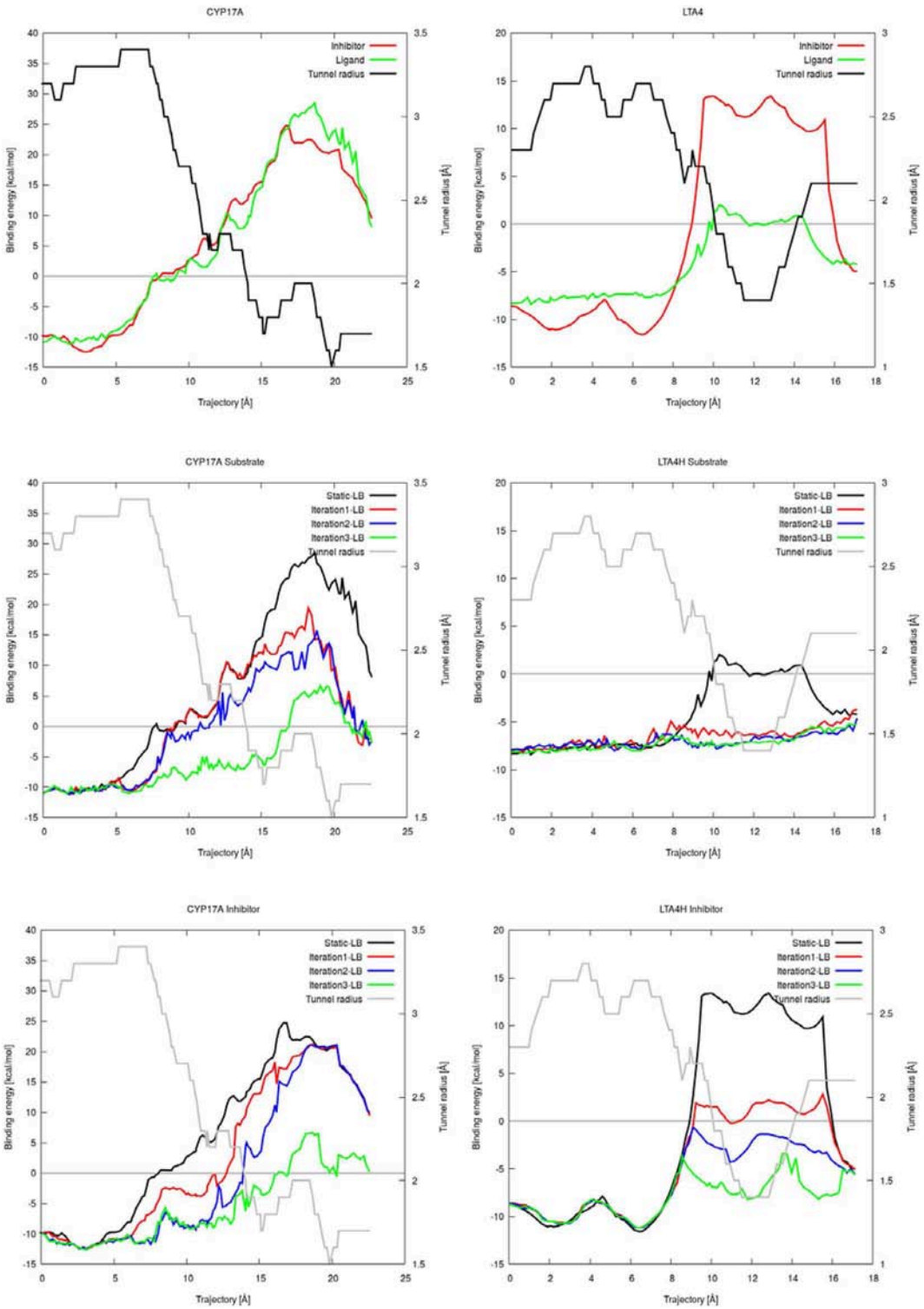


FIGURE 3 | Plots of the binding energies of cytochrome P450 17A1 (left) and leukotriene A4 hydrolase/aminopeptidase (right) obtained from CaverDock with and without flexibility. Binding energies between substrate and inhibitor with tunnel radius present (top). Binding energies from substrate simulations with flexibility, rigid simulation, and tunnel radius on the background (middle). Binding energies from all inhibitor simulations with flexibility, rigid simulation and tunnel radius on the background (bottom).

TABLE 1 | Summary of calculation times for lower bound calculations with rigid and flexible simulations and indication of the flexible residues added on each iteration.

		Cytochrome P450 17A1		Leukotriene A4 hydrolase/aminopeptidase	
		Inhibitors	**Substrates**	**Inhibitors**	**Substrates**
Rigid	Time	47 min 48 s	48 min 13 s	205 min 32 s	24 min 39 s
1st Iteration	Time	110 min 35 s	66 min 18 s	338 min 41 s	50 min 32 s
	Flexible Residues	Ile205, Tyr201	Ile205, Ile246	Val381, Val367	Val367, Phe362
2nd Iteration	Time	252 min 28 s	168 min 3 s	854 min 20 s	112 min 56 s
	Flexible Residues	Arg239, His235	Tyr201, His235	Lys364, Phe362	His360, Lys364
3rd Iteration	Time	436 min 20 s	332 min 22 s	1,994 min 1 s	156 min 11 s
	Flexible Residues	Ile198, Leu242	Leu243, Arg239	His360, Lys385	Gln136, Ile372

Residues are chosen automatically to lower potential energy at bottlenecks.

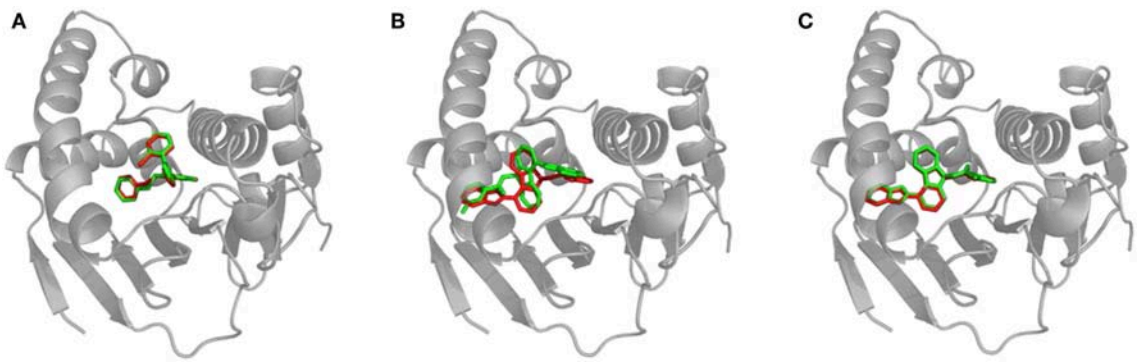


FIGURE 4 | Structural comparison of CaverDock binding poses and X-ray structures. **(A)** Compound_14 correctly predicted by CaverDock (green) with low RMSD compared to the original crystal structure (red). **(B)** Compound_70 wrongly fitted in the pocket by CaverDock (green) with the original size of the tunnel. The overall position is similar, but orientation of the compound is incorrect. **(C)** Compound_70 correctly fitted in the pocket with the increased size of the tunnel. CaverDock (green) predicted a pose in equivalent position to the original crystal structure (red).

proper location and configuration in a vast majority (29 out of 34) cases. We show the example of correct fit in **Figure 4A**. In the case of compound_11, compound_19, compound_38, and compound_70 the correct pose was found by CaverDock but was not correctly identified. A different pose with the lowest energy was picked. In the case of the compound_18, CaverDock failed to find the correct conformation both for the closest and the lowest energy case. The high RMSDs may be caused by incorrect orientation of the ligand and also by the location of the inhibitor. The conformations of inhibitors which are deeper in the protein structure and are out of the tunnel may become unreachable for CaverDock since the ligand is always spatially constrained to the disks.

We experimented with the settings of CaverDock and recalculated the trajectories for the five problematic cases. We found out that by extending the radius of discretized tunnel discs by 10 Å, CaverDock is able to explore deeper parts of the cavity since the ligand has more freedom for movement. The resulting changes of RMSD are shown in **Supplementary Table S2**. The RMSDs were lowered and the binding poses were improved tremendously in case of compound_18, compound_19, and compound_70. This improvement in geometry for the compound_70 is shown in **Figures 4B,C**. The lowest energy pose for the compound_38 was still not identified correctly. Based on these findings, we decided to implement the tunnel extension for our future CaverDock calculations since the improvements were substantially beneficial.

Energy Analysis of N-HSP90 HOLO and APO Forms

CaverDock was used to analyze the unbinding of inhibitors from corresponding HOLO structures. We studied 32 cases with available kinetic data (Kokh et al., 2018). Selected energy values were extracted from the energy profiles: the energy minimum close to the start of the trajectory corresponding with the ligand-bound in the active site (E_{Bound}) and the energy at the tunnel mouth—the last disk of the original tunnel—related with the surface-bound ligand ($E_{Surface}$). In this specific case, there were no visible barriers as shown in the energy profiles **Supplementary Figure S1**. Therefore, we had to use the difference between bound and surface state, the ΔE_{BS} as possible energy barrier which needs to be overcome during the process of unbinding. We calculated the correlation between ΔE_{BS} and the experimentally measured values for k_{on}, K_D, and k_{off}. We found a significant correlation of 0.53 for ΔE_{BS} with $\log(k_{off})$. Comparison of our results with

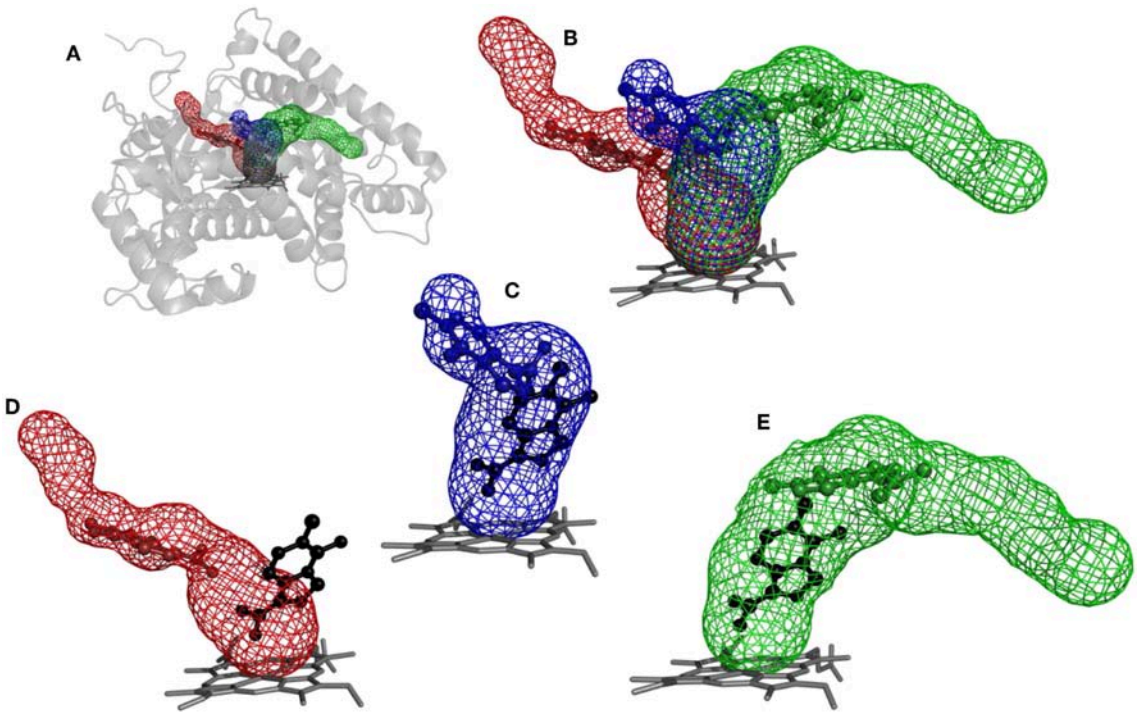


FIGURE 5 | Structure of cytochrome P450 17A1 with tunnels calculated using Caver and binding poses obtained by CaverDock. **(A)** Global view of tunnels—tunnel 1 (blue), tunnel 2A (red), tunnel 3 (green). **(B)** View of tunnels with no protein visualized. Each tunnel has the ligand bound in the minimum obtained from the CaverDock continuous trajectory calculation. **(C)** Tunnel 1 and minimum binding energy pose obtained from CaverDock (blue) and AutoDock Vina (black). **(D)** Tunnel 2A and minimum binding energy pose obtained by CaverDock (red) and AutoDock Vina (black). **(E)** Tunnel 3 and minimum binding energy pose obtained by CaverDock (green) and AutoDock Vina (black).

the previously published correlation 0.63 for the computed relative residence times from molecular dynamics simulations, tcomp, with the measured residence times t_{expt} ($t_{expt} = 1/k_{off}$), we confirmed that CaverDock is able to predict koff rates when HOLO structures are used with only a fraction of the computational effort.

We checked the previous findings from the HOLO dataset by simulating the complete set of inhibitors with the APO structure. We did not find any correlation in this case. This, together with no visible barriers in the CaverDock profiles and slow kinetic rates suggests conformational changes in the protein during the binding and unbinding of the inhibitor molecules. Essential conformational change is missing in the APO structure forcing the molecules to bind differently when simulated by CaverDock. Development of the new version of CaverDock that will be taking into account protein backbone dynamics is currently on-going in our laboratory.

Screening of Inhibitors

The purpose of this analysis was to test whether CaverDock (Filipovic et al., 2019; Vavra et al., 2019) could be used for virtual screening. After deciding on the targets and libraries of compounds to use, we analyzed the tunnels for both targets. First, we choose the tunnels according to their ranking given by Caver and information from the literature and then used CaverDock to move the ligands from the outside of the proteins to the active site. Next, we performed virtual screening with the same libraries and targets using AutoDock Vina. It is worth noting that there was a large difference in the exhaustiveness used between the two programs: an exhaustiveness of thirty was used with Autodock Vina, whereas an exhaustiveness of one was used with CaverDock to keep the run time as short as possible. We showed that CaverDock provided new insights into the receptor ligand affinity. We also showed that CaverDock was a computationally cheap method with low run times. We studied 5 tunnels in the two proteins: three tunnels in cytochrome P450 17A1 (**Figure 5**) and two tunnels in leukotriene A4 hydrolase/aminopeptidase (**Figure 6**). Tunnel 1 was much shorter than the other two tunnels studied in cytochrome P450 17A1 (**Table 2**). It also had a narrow mouth when compared to the rest of the tunnel, but it was still wider than tunnel 3. Tunnel 2A was the most sinuous tunnel of the three, with more twists than the other two tunnels. However, they were not as sharp as the turn in tunnel 3. Tunnel 3 had a sharp turn halfway through the tunnel. It was narrow at the entrance of the protein, but after the turn widened sufficiently to allow a bulky inhibitor like Abiraterone to bind to the heme-group, as in the crystal structure.

The tunnels are modeled with the drug Temozolomide in both the CaverDock (**Supplementary Video**) and AutoDock

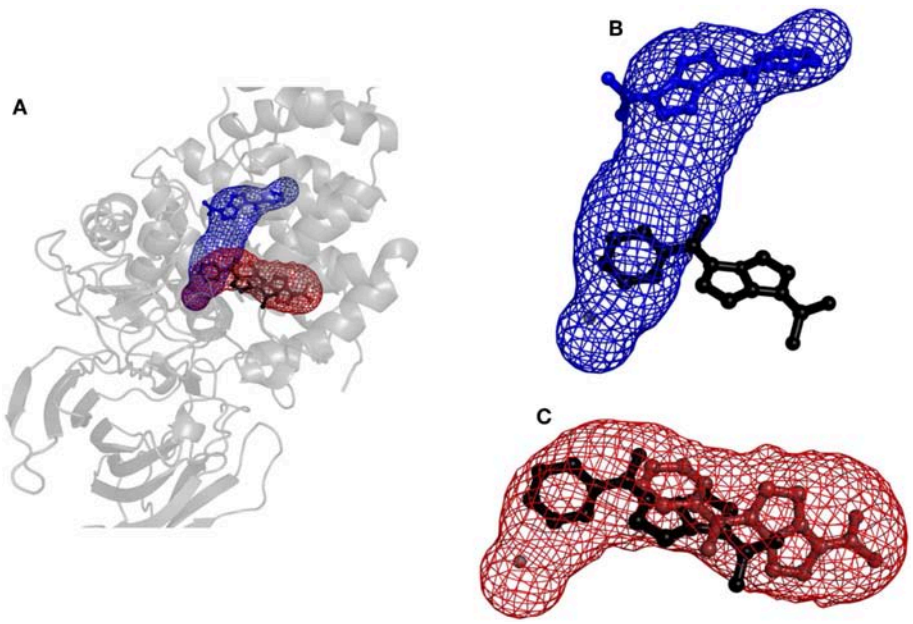


FIGURE 6 | Structure of protein leukotriene A4 hydrolase/aminopeptidase with tunnels calculated using Caver and binding poses obtained by CaverDock. **(A)** Global view with tunnels LTA4 (blue) and PGP (red). **(B)** Tunnel LTA4 and minimum binding energy pose represented by balls and sticks obtained from CaverDock (blue) and AutoDock Vina (black). **(C)** Tunnel PGP and minimum binding energy pose represented by balls and sticks obtained from CaverDock (red) and AutoDock Vina (black).

TABLE 2 | Summary of data for tunnels in the target proteins.

	Cytochrome P450 17A1			Leukotriene A4 hydrolase/aminopeptidase	
	Tunnel 1	**Tunnel 2A**	**Tunnel 3**	**Tunnel LTA4**	**Tunnel PGP**
Size of library	105	105	105	54	54
Continuous	41 (39.0%)	42 (40.0%)	42 (40.0%)	20 (37.0%)	21 (38.9%)
Lower bound	100 (95.2%)	91 (86.7%)	93 (88.6%)	48 (88.8%)	50 (92.6%)
Only lower bound	49 (46.7%)	59 (56.2%)	51 (48.6%)	28 (51.8%)	29 (53.7%)
Stopped at bottleneck	5 (4.8%)	14 (13.3%)	11 (10.5%)	6 (11.1%)	4 (7.4%)
Time average	41 min 50 s	68 min 59 s	52 min 11 s	36 min 17 s	22 min 23 s
Highest time	272 min 42 s	136 min 4 s	269 min 38 s	73 min 13 s	47 min 6 s
Lowest time	4 min 1 s	14 min 10 s	5 min 22 s	2 min 59 s	1 min 20 s
Length (Å)	15.1	24.9	28.2	20.4	25.4
Curvature (Å)	1.2	1.4	1.4	1.3	1.2
Maximum bottleneck (Å)	1.4	1.3	1.3	1.9	1.7
Average ligand RMSD lower bound docking (Å)	3.0	5.3	1. 9	5.4	2.8
Average ligand RMSD continuous docking (Å)	6.8	10.8	10.5	11.2	6.8

Vina virtual screening (**Figure 5**). CaverDock yielded a minimum binding energy for a conformation inside the tunnel, rather than close to the heme group indicated by the AutoDock Vina calculation. The distance to the heme group was 10.3 Å with CaverDock for tunnel 3 and 2.6 Å with AutoDock Vina. Similar trends were observed for the other two tunnels. For tunnels 1 and 2A, CaverDock gave a minimum binding energy at 8.1 and 7.5 Å from the heme group, respectively. It should be noted that CaverDock was still able to bind the ligand to the heme group but at a higher energy than the conformations presented here. This result clearly demonstrates the value of the analysis of ligand binding and unbinding using CaverDock. Whereas, AutoDock Vina performs docking in a matrix box set by the user, CaverDock considers a continuous motion from the entrance of the tunnel to the active site, restrained to the tunnel found by Caver. It is also apparent in **Figure 5**, that tunnel 2A was deprecated by AutoDock Vina. Whereas, the closest nitrogen was bound to the heme group, the rest of the molecule was in a common area overlapping both tunnel 1 and tunnel 3. At the same time, the ligand was positioned away from tunnel 2A, with only a few

atoms in the common space where all three tunnels overlapped. Despite these differences in the docking calculations, the minima binding energy obtained from CaverDock and binding affinity obtained from AutoDock Vina showed no significant differences for the case presented here. A complete data table comparing the AutoDock Vina and CaverDock virtual screening is presented in the **Supporting Information**.

Results for leukotriene A4 hydrolase/aminopeptidase are shown in **Figure 6**. Comparing tunnel LTA4 (blue) with tunnel PGP (red), only slight differences were discerned in the sizes of the tunnels. Tunnel PGP had a sharp turn, whereas tunnel LTA4 did not. Tunnel LTA4 had a higher overall curvature than tunnel PGP. Both tunnels are presented with the minimum binding energy pose obtained from CaverDock with the drug Ketorolac. Ketorolac was not bound to the zinc atom in the active site for both studied cases. AutoDock Vina yielded a conformation with the drug molecule at a distance of 4.8 Å from the zinc atom and clearly docked in the tunnel PGP, with only one ring in the common area overlapping both tunnels (**Figure 6**). Using CaverDock, the minima were even farther away from the zinc in the active site: the distance in tunnel PGP was 8.8 Å and in tunnel LTA4 11.8 Å. When the drug molecule is bound in tunnel LTA4, higher energetic barriers were obtained. It is known that pro-inflammatory mediator biosynthesis occurs through tunnel LTA4 and that the inhibitor Pro-Gly-Pro enters and exerts its effects through a different tunnel (Sanson et al., 2011; Čolović et al., 2013).

One of the main objectives of this study was to assess the computational costs of a project with the novel computational tool. In total, 105 drugs were docked to the three tunnels in cytochrome P450 17A1 (**Table 2**). From these, CaverDock was able to finish a continuous (upper bound) trajectory calculation for 39.7% of the drugs and a discontinuous (lower bound) trajectory for 90.2%. On average, the drugs were not able to overcome the bottlenecks 9.8% of the time. This result does not mean that the calculation failed but that the ligand was not able to pass through the rigid receptor. Values were also similar for leukotriene A4 hydrolase/aminopeptidase. The length of the five tunnels studied ranged from 15.1 to 28.2 Å, the curvature of the tunnels ranged from 1.2 to 1.4 Å and bottlenecks ranged from 1.3 to 1.9 Å. These differences in length, curvature and bottlenecks yielded very different tunnels and tunnel shapes, as evident in **Figures 5**, **6**. The approach presented here constitutes a computationally low-cost method for virtual screening with a run time average of 2,660 s (∼44 min). Moreover, when the upper bound calculation was turned off, the lower bound results could be completed within several minutes using a computer with 4 processors. Note that each calculation runs independently, allowing users with sufficient computing power to perform a virtual screening protocol on a full library in a parallel manner.

Using data obtained from a virtual screening campaign, it is possible to analyze a functionally important tunnel for a given target and set of drugs. Although it is not always easy to select a preferred tunnel, it may be possible to identify tunnels that are not favored. We found that in the case of cytochrome P450 17A1, the jobs finished successfully with a continuous trajectory and tunnel 2A had higher barriers than the other two tunnels

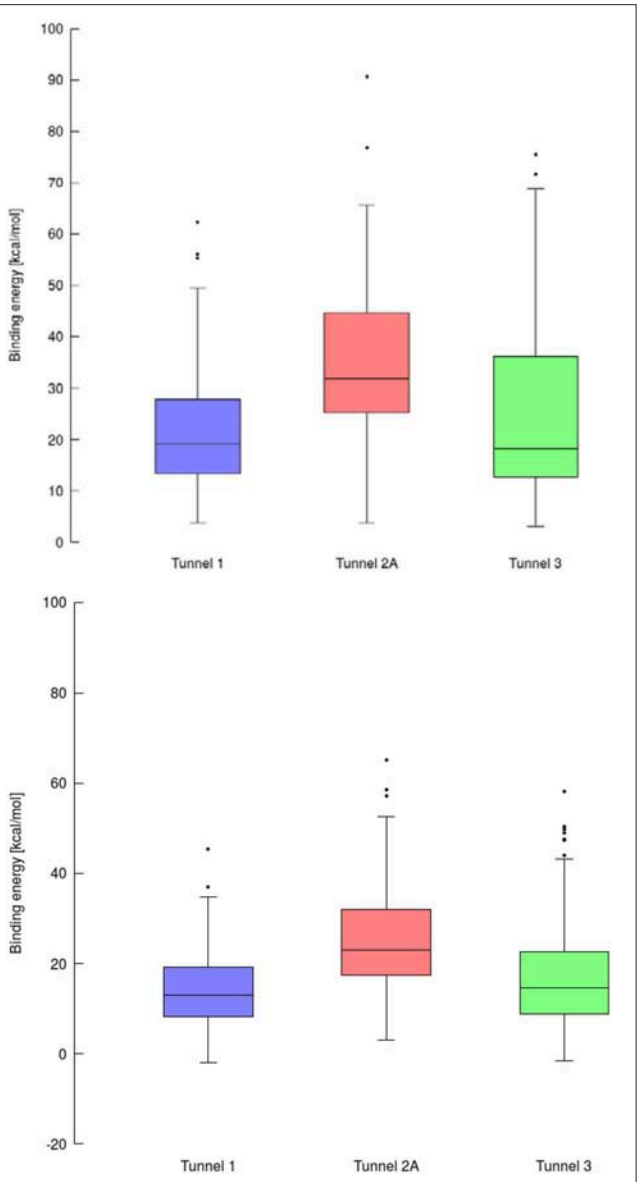


FIGURE 7 | Box and Whiskers plot representing the maxima (energy barriers) for each continuous (up) and lower bound (down) trajectories obtained for cytochrome P450 17A1. Outlying values are indicated by circles.

(**Figure 7**), therefore it is not preferred. However, we could not determine which of the two remaining tunnels 1 and 3 would be better to consider in a drug design project since there was no statistically significant difference in the energy barriers. Possibly both tunnels can be explored by ligands during their (un)binding. The results were more conclusive for leukotriene A4 hydrolase/aminopeptidase, as shown by the differences in energy barriers (**Figures 7**, **8**). In the case of the continuous (upper-bound) calculation, the drug molecule was taken through one smooth trajectory with the possibility of backtracking if it encountered a bottleneck. Backtracking allowed the drug to find a more favorable conformation to overcome the bottleneck. In the

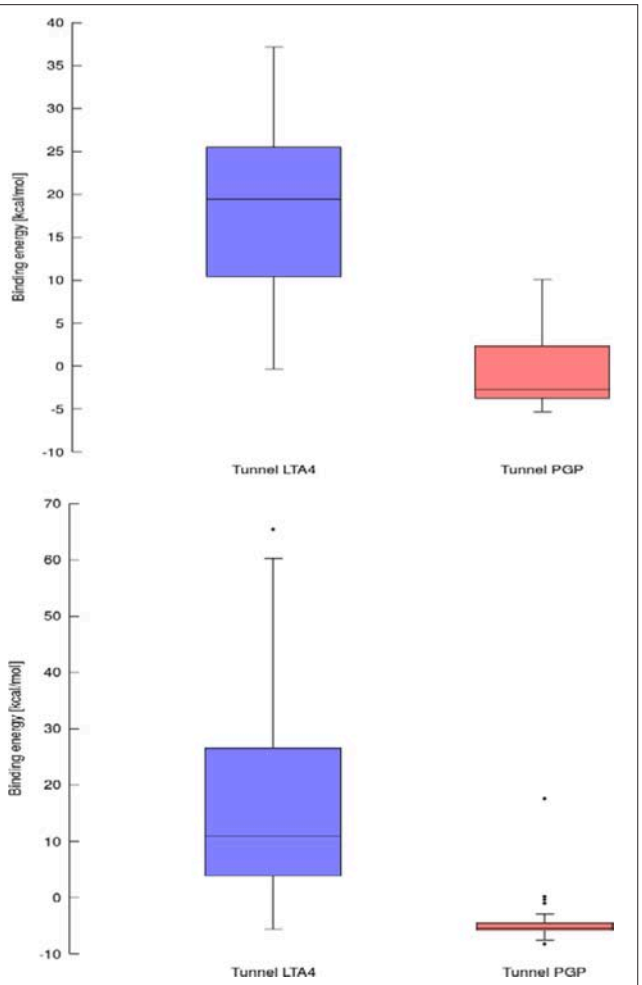


FIGURE 8 | Box and Whiskers plot representing the maxima (energy barriers) for continuous (up) and lower bound (down) trajectories obtained for leukotriene A4 hydrolase/aminopeptidase. Outlying values are indicated by circles.

case of the lower-bound calculation, once the drug encountered a bottleneck, it was allowed to flip in order to find a more suitable conformation on the other side of the bottleneck, while the point being dragged through the discs of the tunnel was kept constant (**Figure 8**). This trajectory always yielded a lower energy value for the barrier because, by definition, the bottleneck was easier to overcome. On the other hand, lower time demands and similar results make the lower bound calculation very powerful for virtual screening.

CONCLUSIONS

Our results demonstrate that CaverDock is applicable for screening of large libraries of potential inhibitors. It provides information on binding and unbinding processes. The tool estimates a profile of potential energies and calculates respective trajectories without the need for time-demanding molecular dynamics simulations. Setting up a calculation using CaverDock is simple and comprises five steps: (i) definition of a receptor, (ii) definition of the ligands, (iii) calculation of tunnels using Caver, (iv) screening of un/binding trajectories, and (v) data analysis. The tool is accompanied by a user manual that explains the setting up of calculations as well as troubleshooting. A standalone version of CaverDock with detailed documentation is available at https://loschmidt.chemi.muni.cz/caverdock/. The automated version of CaverDock is available via the web https://loschmidt.chemi.muni.cz/caverweb/.

The dynamics of side chains lining the protein tunnels and channels can be described to a certain level with the current implementation of CaverDock. Making residue side chains flexible increases calculation times but ultimately considers protein dynamics. We concluded that simulations employing a large number (>4) of flexible amino acid residues may cause undesirable steric clashes. Thus, we advise that results obtained with flexible residues should be interpreted carefully using biochemical intuition when analyzing calculated trajectories and energy profiles. Implementation of a more thorough protocol to address protein flexibility is on-going in our laboratory. CaverDock calculations can be extended to ensembles of protein structures. Particularly challenging is the trade-off between rigorous description of flexible systems and time demands connected with such calculations. Structural comparison of complexes obtained by CaverDock with those determined by crystallographic analysis revealed that we were able to predict the correct poses for a vast majority of inhibitors. The comparison of our profile of potential energies with the rates obtained by kinetic results yields a correlation of 0.53 whereas the more computational expensive molecular dynamics simulation had a correlation of 0.63. Prediction accuracy can be potentially improved by proper treatment of backbone flexibility.

Our study demonstrates that CaverDock is sufficiently fast to screen even large libraries of ligands. Calculation of rigid trajectories using 4 processors took on average 53 min per molecule with 90% successfully calculated cases. Bulky or very flexible ligands take more time, but some of these large ligands may not be able to access the active site via the studied access tunnels. Although it takes longer to perform a CaverDock calculation than a pure virtual screening of ligand binding to the active site with AutoDock Vina, CaverDock provides more data, which may be useful in rational drug design projects. Information on the bottlenecks and energy required for ligands to pass through these narrowed parts of the access tunnel could be useful for medicinal chemists. CaverDock was able to correctly identify tunnels in the proteins explored by the inhibitors included in our screening campaigns.

In summary, we have shown that CaverDock is a robust and ready-to-use software that can be employed in screening campaigns of important pharmacological targets. CaverDock analysis may be a useful complement

to virtual screening campaigns carried out using traditional docking tools.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the https://loschmidt.chemi.muni.cz/data/caverdock/pinto_2019_suppl/.

AUTHOR CONTRIBUTIONS

GP and OV carried out the computational work and wrote the manuscript. JF developed the software. DB and JD designed the study. All authors contributed to interpretation of the data, revision of the manuscript, and have given approval of its final version.

FUNDING

This work was supported by the Ministry of Education, Youth, and Sports of the Czech Republic (LQ1605, LM2015047, LM2015051, LM2015055, CZ.02.1.01/0.0/0.0/16_013/0001761) and the European Union (720776). Computational resources were provided by CESNET (LM2015042) and the CERIT Scientific Cloud (LM2015085).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00709/full#supplementary-material

Supplementary Video | Lower bound calculation of Temozolomide through Tunnel 3 of cytochrome P450 17A1.

REFERENCES

Bajorath, J. (2002). Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* 1:882. doi: 10.1038/nrd941

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Bielska, E., Lucas, X., Czerwoniec, A., Kasprzak, J. M., Kaminska, K. H., and Bujnicki, J. M. (2014). Virtual screening strategies in drug design – methods and applications. *BioTechnologia* 92, 249–264. doi: 10.5114/bta.2011.46542

Bleicher, K. H., Böhm, H.-J., Müller, K., and Alanine, A. I. (2003). Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* 2, 369–378. doi: 10.1038/nrd1086

Bottegoni, G., Veronesi, M., Bisignano, P., Kacker, P., Favia, A. D., and Cavalli, A. (2016). Development and application of a virtual screening protocol for the identification of multitarget fragments. *ChemMedChem* 11, 1259–1263. doi: 10.1002/cmdc.201500521

Chang, D. T.-H., Oyang, Y.-J., and Lin, J.-H. (2005). MEDock: a web server for efficient prediction of ligand binding sites based on a novel optimization algorithm. *Nucleic Acids Res.* 33, W233–W238. doi: 10.1093/nar/gki586

Chaput, L., Martinez-Sanz, J., Saettel, N., and Mouawad, L. (2016). Benchmark of four popular virtual screening programs: construction of the active/decoy dataset remains a major determinant of measured performance. *J. Cheminform.* 8:56. doi: 10.1186/s13321-016-0167-x

Cheng, T., Li, Q., Zhou, Z., Wang, Y., and Bryant, S. H. (2012). Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J.* 14, 133–141. doi: 10.1208/s12248-012-9322-0

Chovancova, E., Pavelka, A., Benes, P., Strnad, O., Brezovsky, J., Kozlikova, B., et al. (2012). CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput. Biol.* 8:e1002708. doi: 10.1371/journal.pcbi.1002708

Clark, D. E. (2008). What has virtual screening ever done for drug discovery? *Expert Opin. Drug Discov.* 3, 841–851. doi: 10.1517/17460441.3.8.841

Clayton, G. M., Klein, D. J., Rickert, K. W., Patel, S. B., Kornienko, M., Zugay-Murphy, J., et al. (2013). Structure of the bacterial deacetylase LpxC bound to the nucleotide reaction product reveals mechanisms of oxyanion stabilization and proton transfer. *J. Biol. Chem.* 288, 34073–34080. doi: 10.1074/jbc.M113.513028

Čolović, M. B., Krstić, D. Z., Lazarević-Pašti, T. D., BondŽić, A. M., and Vasić, V. M. (2013). Acetylcholinesterase inhibitors: pharmacology and toxicology. *Curr. Neuropharmacol.* 11, 315–335. doi: 10.2174/1570159X11311030006

Cross, J. B., Thompson, D. C., Rai, B. K., Baber, J. C., Fan, K. Y., Hu, Y., et al. (2009). Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* 49, 1455–1474. doi: 10.1021/ci900056c

Cui, Y.-L., Xue, Q., Zheng, Q.-C., Zhang, J.-L., Kong, C.-P., Fan, J.-R., et al. (2015). Structural features and dynamic investigations of the membrane-bound cytochrome P450 17A1. *Biochim. Biophys. Acta* 1848, 2013–2021. doi: 10.1016/j.bbamem.2015.05.017

Cummings, M. D., DesJarlais, R. L., Gibbs, A. C., Mohan, V., and Jaeger, E. P. (2005). Comparison of automated docking programs as virtual screening tools. *J. Med. Chem.* 48, 962–976. doi: 10.1021/jm049798d

Devaurs, D., Bouard, L., Vaisset, M., Zanon, C., Al-Bluwi, I., Iehl, R., et al. (2013). MoMA-LigPath: a web server to simulate protein-ligand unbinding. *Nucleic Acids Res.* 41, W297–W302. doi: 10.1093/nar/gkt380

DeVore, N. M., and Scott, E. E. (2012). Structures of cytochrome P450 17A1 with prostate cancer drugs abiraterone and TOK-001. *Nature* 482, 116–119. doi: 10.1038/nature10743

Epps, D. E., and Vosters, A. F. (2002). The essential role of a free sulfhydryl group in blocking the cholesteryl site of cholesteryl ester transfer protein (CETP). *Chem. Phys. Lipids* 114, 113–122. doi: 10.1016/S0009-3084(01)00187-6

Espona-Fiedler, M., Soto-Cerrato, V., Hosseini, A., Lizcano, J. M., Guallar, V., Quesada, R., et al. (2012). Identification of dual mTORC1 and mTORC2 inhibitors in melanoma cells: prodigiosin vs. obatoclax. *Biochem. Pharmacol.* 83, 489–496. doi: 10.1016/j.bcp.2011.11.027

Fernández, A. (2014). Communication: chemical functionality of interfacial water enveloping nanoscale structural defects in proteins. *J. Chem. Phys.* 140:221102. doi: 10.1063/1.4882895

Filipovic, J., Vávra, O., Plhák, J., Bednar, D., Marques, S. M., Brezovsky, J., et al. (2019). CaverDock: a novel method for the fast analysis of ligand transport. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2019.2907492. [Epub ahead of print].

Funk, C. D. (2001). Prostaglandins and leukotrienes: advances in eicosanoid biology. *Science* 294, 1871–1875. doi: 10.1126/science.294.5548.1871

Gattis, S. G., Hernick, M., and Fierke, C. A. (2010). Active site metal ion in UDP-3-O-((R)-3-hydroxymyristoyl)-N-acetylglucosamine deacetylase (LpxC) switches between Fe(II) and Zn(II) depending on cellular conditions. *J. Biol. Chem.* 285, 33788–33796. doi: 10.1074/jbc.M110.147173

Gerber, N. C., Rodriguez-Crespo, I., Nishida, C. R., and Ortiz de Montellano, P. R. (1997). Active site topologies and cofactor-mediated conformational changes of nitric-oxide synthases. *J. Biol. Chem.* 272, 6285–6290. doi: 10.1074/jbc.272.10.6285

Goldman, M., Wittelsberger, A., and De Magistris, M.-T. (2013). The innovative medicines initiative moves translational immunology forward. *Eur. J. Immunol.* 43, 298–302. doi: 10.1002/eji.201370024

Gomis-Rüth, F. X., Botelho, T. O., and Bode, W. (2012). A standard orientation for metallopeptidases. *Biochim. Biophys. Acta* 1824, 157–163. doi: 10.1016/j.bbapap.2011.04.014

Guallar, V., Lu, C., Borrelli, K., Egawa, T., and Yeh, S.-R. (2009). Ligand migration in the truncated hemoglobin-II from *Mycobacterium tuberculosis*: the role of G8 tryptophan. *J. Biol. Chem.* 284, 3106–3116. doi: 10.1074/jbc.M806183200

Haeggström, J. Z. (2004). Leukotriene A4 hydrolase/aminopeptidase, the gatekeeper of chemotactic leukotriene B4 biosynthesis. *J. Biol. Chem.* 279, 50639–50642. doi: 10.1074/jbc.R400027200

Haeggström, J. Z., Tholander, F., and Wetterholm, A. (2007). Structure and catalytic mechanisms of leukotriene A4 hydrolase. *Prostaglandins Other Lipid Mediat.* 83, 198–202. doi: 10.1016/j.prostaglandins.2007.01.006

Haeggström, J. Z., Wetterholm, A., Vallee, B. L., and Samuelsson, B. (1990). Leukotriene A4 hydrolase: an epoxide hydrolase with peptidase activity. *Biochem. Biophys. Res. Commun.* 173, 431–437. doi: 10.1016/S0006-291X(05)81076-9

Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E., and Hutchison, G. R. (2012). Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminform.* 4:17. doi: 10.1186/1758-2946-4-17

Hernández-Ortega, A., Borrelli, K., Ferreira, P., Medina, M., Martínez, A. T., and Guallar, V. (2011). Substrate diffusion and oxidation in GMC oxidoreductases: an experimental and computational study on fungal aryl-alcohol oxidase. *Biochem. J.* 436, 341–350. doi: 10.1042/BJ20102090

Huang, Z., Wong, C. F., and Wheeler, R. A. (2008). Flexible protein-flexible ligand docking with disrupted velocity simulated annealing. *Proteins* 71, 440–454. doi: 10.1002/prot.21781

Jarzynski, 1997###Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* 78, 2690–2693. doi: 10.1103/PhysRevLett.78.2690

Kabakov, A. E., Kudryavtsev, V. A., and Gabai, V. L. (2010). Hsp90 inhibitors as promising agents for radiotherapy. *J. Mol. Med.* 88, 241–247. doi: 10.1007/s00109-009-0562-0

Kaczor, A. A., Silva, A. G., Loza, M. I., Kolb, P., Castro, M., and Poso, A. (2016). Structure-based virtual screening for dopamine D2 receptor ligands as potential antipsychotics. *ChemMedChem* 11, 718–729. doi: 10.1002/cmdc.201500599

Kansy, M., Senner, F., and Gubernator, K. (1998). Physicochemical high throughput screening: parallel artificial membrane permeation assay in the description of passive absorption processes. *J. Med. Chem.* 41, 1007–1010. doi: 10.1021/jm970530e

Khatri, Y., Gregory, M. C., Grinkova, Y. V., Denisov, I. G., and Sligar, S. G. (2014). Active site proton delivery and the lyase activity of human CYP17A1. *Biochem. Biophys. Res. Commun.* 443, 179–184. doi: 10.1016/j.bbrc.2013.11.094

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., et al. (2016). PubChem substance and compound databases. *Nucleic Acids Res.* 44, D1202–D1213. doi: 10.1093/nar/gkv951

Kokh, D. B., Amaral, M., Bomke, J., Grädler, U., Musil, D., Buchstaller, H.-P., et al. (2018). Estimation of drug-target residence times by τ-random acceleration molecular dynamics simulations. *J. Chem. Theory Comput.* 14, 3859–3869. doi: 10.1021/acs.jctc.8b00230

Kollman, P. (1993). Free energy calculations: applications to chemical and biochemical phenomena. *Chem. Rev.* 93, 2395–2417. doi: 10.1021/cr00023a004

Kruse, A. C., Hu, J., Pan, A. C., Arlow, D. H., Rosenbaum, D. M., Rosemond, E., et al. (2012). Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* 482, 552–556. doi: 10.1038/nature10867

Kush, R., and Goldman, M. (2014). Fostering responsible data sharing through standards. *N. Engl. J. Med.* 370, 2163–2165. doi: 10.1056/NEJMp1401444

Lamb, M. L., and Jorgensen, W. L. (1997). Computational approaches to molecular recognition. *Curr. Opin. Chem. Biol.* 1, 449–457. doi: 10.1016/S1367-5931(97)80038-5

Lavecchia, A., and Di Giovanni, C. (2013). Virtual screening strategies in drug discovery: a critical review. *Curr. Med. Chem.* 20, 2839–2860. doi: 10.2174/09298673113209990001

Laverty, H., Gunn, M., and Goldman, M. (2012). Improving RandD productivity of pharmaceutical companies through public-private partnership: experiences from the Innovative Medicines Initiative. *Expert Rev. Pharmacoecon. Outcomes Res.* 12, 545–548. doi: 10.1586/erp.12.59

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 42, D1091–D1097. doi: 10.1093/nar/gkt1068

Lee, P.-H., Kuo, K.-L., Chu, P.-Y., Liu, E. M., and Lin, J.-H. (2009). SLITHER: a web server for generating contiguous conformations of substrate molecules entering into deep active sites of proteins or migrating through channels in membrane transporters. *Nucleic Acids Res.* 37, W559–W564. doi: 10.1093/nar/gkp359

Li, J., Sun, L., Xu, C., Yu, F., Zhou, H., Zhao, Y., et al. (2012). Structure insights into mechanisms of ATP hydrolysis and the activation of human heat-shock protein 90. *Acta Biochim. Biophys. Sin.* 44, 300–306. doi: 10.1093/abbs/gms001

Lucas, M. F., and Guallar, V. (2012). An atomistic view on human hemoglobin carbon monoxide migration processes. *Biophys. J.* 102, 887–896. doi: 10.1016/j.bpj.2012.01.011

Madadkar-Sobhani, A., and Guallar, V. (2013). PELE web server: atomistic study of biomolecular systems at your fingertips. *Nucleic Acids Res.* 41, W322–W328. doi: 10.1093/nar/gkt454

Mangoni, M., Roccatano, D., and Di Nola, A. (1999). Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins* 35, 153–162.

Marques, S. M., Daniel, L., Buryska, T., Prokop, Z., Brezovsky, J., and Damborsky, J. (2017). Enzyme tunnels and gates as relevant targets in drug design. *Med. Res. Rev.* 37, 1095–1139. doi: 10.1002/med.21430

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* 30, 2785–2791. doi: 10.1002/jcc.21256

Pagadala, N. S., Syed, K., and Tuszynski, J. (2017). Software for molecular docking: a review. *Biophys. Rev.* 9, 91–102. doi: 10.1007/s12551-016-0247-1

Ripphausen, P., Nisius, B., Peltason, L., and Bajorath, J. (2010). Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* 53, 8461–8467. doi: 10.1021/jm101020z

Rudberg, P. C., Tholander, F., Thunnissen, M. M. G. M., and Haeggström, J. Z. (2002). Leukotriene A4 hydrolase/aminopeptidase. Glutamate 271 is a catalytic residue with specific roles in two distinct enzyme mechanisms. *J. Biol. Chem.* 277, 1398–1404. doi: 10.1074/jbc.M106577200

Sanson, B., Colletier, J.-P., Xu, Y., Lang, P. T., Jiang, H., Silman, I., et al. (2011). Backdoor opening mechanism in acetylcholinesterase based on X-ray crystallography and molecular dynamics simulations. *Protein Sci.* 20, 1114–1118. doi: 10.1002/pro.661

Singh, S. P., and Konwar, B. K. (2012). Molecular docking studies of quercetin and its analogues against human inducible nitric oxide synthase. *Springerplus* 1:69. doi: 10.1186/2193-1801-1-69

Sousa, S. F., Cerqueira, N. M., Fernandes, P. A., and Ramos, M. J. (2010). Virtual screening in drug design and development. *Comb. Chem. High Throughput Screen.* 13, 442–453. doi: 10.2174/138620710791293001

Sterling, T., and Irwin, J. J. (2015). ZINC 15 – ligand discovery for everyone. *J. Chem. Inf. Model.* 55, 2324–2337. doi: 10.1021/acs.jcim.5b00559

Straatsma, T. P., and McCammon, J. A. (1992). Computational alchemy. *Annu. Rev. Phys. Chem.* 43, 407–435. doi: 10.1146/annurev.pc.43.100192.002203

Stsiapanava, A., Olsson, U., Wan, M., Kleinschmidt, T., Rutishauser, D., Zubarev, R. A., et al. (2014). Binding of Pro-Gly-Pro at the active site of leukotriene A4 hydrolase/aminopeptidase and development of an epoxide hydrolase selective inhibitor. *PNAS* 111, 4227–4232. doi: 10.1073/pnas.1402136111

Szul, T., Castaldi, P., Cho, M. H., Blalock, J. E., and Gaggar, A. (2016). Genetic regulation of expression of leukotriene A4 hydrolase. *ERJ Open Res.* 2:00058-2015. doi: 10.1183/23120541.00058-2015

Thunnissen, M. M., Nordlund, P., and Haeggström, J. Z. (2001). Crystal structure of human leukotriene A(4) hydrolase, a bifunctional enzyme in inflammation. *Nat. Struct. Biol.* 8, 131–135. doi: 10.1038/84117

Tomić, A., Berynskyy, M., Wade, R. C., and Tomić, S. (2015). Molecular simulations reveal that the long range fluctuations of human DPP III change upon ligand binding. *Mol. Biosyst.* 11, 3068–3080. doi: 10.1039/C5MB00465A

Totrov, M., and Abagyan, R. (2008). Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.* 18, 178–184. doi: 10.1016/j.sbi.2008.01.004

Trott, O., and Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comput. Chem.* 31, 455–461. doi: 10.1002/jcc.21334

Vavra, O., Filipovic, J., Plhak, J., Bednar, D., Marques, S. M., Brezovsky, J., et al. (2019). CaverDock: a molecular docking-based tool to analyse ligand transport through protein tunnels and channels. *Bioinformatics* btz386. doi: 10.1093/bioinformatics/btz386. [Epub ahead of print].

Wang, S., Ma, J., Peng, J., and Xu, J. (2013). Protein structure alignment beyond spatial proximity. *Sci. Rep.* 3:1448. doi: 10.1038/srep01448

Yoshimoto, F. K., and Auchus, R. J. (2015). The diverse chemistry of cytochrome P450 17A1 (P450c17, CYP17A1). *J. Steroid Biochem. Mol. Biol.* 151, 52–65. doi: 10.1016/j.jsbmb.2014.11.026

Yu, X., Cojocaru, V., and Wade, R. C. (2013). Conformational diversity and ligand tunnels of mammalian cytochrome P450s. *Biotechnol. Appl. Biochem.* 60, 134–145. doi: 10.1002/bab.1074

Zhang, J. H., Chung, T. D., and Oldenburg, K. R. (1999). A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.* 4, 67–73. doi: 10.1177/108705719900400206

Zhang, T., Ozbil, M., Barman, A., Paul, T. J., Bora, R. P., and Prabhakar, R. (2015). Theoretical insights into the functioning of metallopeptidases and their synthetic analogues. *Acc. Chem. Res.* 48, 192–200. doi: 10.1021/ar500301y

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Pinto, Vavra, Filipovic, Stourac, Bednar and Damborsky. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 12 November 2019
doi: 10.3389/fchem.2019.00764

Computational Insight Into the Small Molecule Intervening PD-L1 Dimerization and the Potential Structure-Activity Relationship

Danfeng Shi[1], Xiaoli An[1], Qifeng Bai[2], Zhitong Bing[2,3], Shuangyan Zhou[1,4], Huanxiang Liu[4] and Xiaojun Yao[1,5]**

[1] State Key Laboratory of Applied Organic Chemistry, Department of Chemistry, Lanzhou University, Lanzhou, China, [2] School of Basic Medical Science, Lanzhou University, Lanzhou, China, [3] Institute of Modern Physics of Chinese Academy of Sciences, Lanzhou, China, [4] School of Pharmacy, Lanzhou University, Lanzhou, China, [5] State Key Laboratory of Quality Research in Chinese Medicine, Macau Institute for Applied Research in Medicine and Health, Macau University of Science and Technology, Macau, China

OPEN ACCESS

Edited by:
Simone Brogi,
University of Pisa, Italy

Reviewed by:
Eduardo De Faria Franca,
Federal University of Uberlandia, Brazil
Chanin Nantasenamat,
Mahidol University, Thailand

***Correspondence:**
Huanxiang Liu
hxliu@lzu.edu.cn
Xiaojun Yao
xjyao@lzu.edu.cn

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: 02 May 2019
Accepted: 24 October 2019
Published: 12 November 2019

Citation:
Shi D, An X, Bai Q, Bing Z, Zhou S, Liu H and Yao X (2019) Computational Insight Into the Small Molecule Intervening PD-L1 Dimerization and the Potential Structure-Activity Relationship. Front. Chem. 7:764. doi: 10.3389/fchem.2019.00764

Recently, small-molecule compounds have been reported to block the PD-1/PD-L1 interaction by inducing the dimerization of PD-L1. All these inhibitors had a common scaffold and interacted with the cavity formed by two PD-L1 monomers. This special interactive mode provided clues for the structure-based drug design, however, also showed limitations for the discovery of small-molecule inhibitors with new scaffolds. In this study, we revealed the structure-activity relationship of the current small-molecule inhibitors targeting dimerization of PD-L1 by predicting their binding and unbinding mechanism via conventional molecular dynamics and metadynamics simulation. During the binding process, the representative inhibitors (BMS-8 and BMS-1166) tended to have a more stable binding mode with one PD-L1 monomer than the other and the small-molecule inducing PD-L1 dimerization was further stabilized by the non-polar interaction of Ile54, Tyr56, Met115, Ala121, and Tyr123 on both monomers and the water bridges involved in $_{A}$Lys124. The unbinding process prediction showed that the PD-L1 dimerization kept stable upon the dissociation of ligands. It's indicated that the formation and stability of the small-molecule inducing PD-L1 dimerization was the key factor for the inhibitory activities of these ligands. The contact analysis, R-group based quantitative structure-activity relationship (QSAR) analysis and molecular docking further suggested that each attachment point on the core scaffold of ligands had a specific preference for pharmacophore elements when improving the inhibitory activities by structural modifications. Taken together, the results in this study could guide the structural optimization and the further discovery of novel small-molecule inhibitors targeting PD-L1.

Keywords: PD-L1, small-molecule inhibitors, molecular dynamics simulation, metadynamics simulation, R-group QSAR

INTRODUCTION

The blockage of the protein-protein interaction (PPI) between programmed cell death protein 1 (PD-1) and programmed cell death 1 ligand 1 (PD-L1) can reactivate the effector functions of T cell and eliminate tumor phenotypes with significant PD-L1 expression (Gatalica et al., 2014; Patel and Kurzrock, 2015; Sharma and Allison, 2015a,b). The crystal structures of PD-1/PD-L1 complex revealed the interface and hot-spot domains for both proteins (Zak et al., 2015; Pascolutti et al., 2016), which provided the structural basis for drug design. Ligands such as monoclonal antibodies (mAbs) (Lee et al., 2016, 2017; Liu K. et al., 2016; Tan et al., 2017; Zhang et al., 2017a,b), peptides (Chang et al., 2015; Magiera-Mularz et al., 2017), and small-molecule compounds (Abdel-Magid, 2015; Zak et al., 2016; Skalniak et al., 2017) had been discovered to interact with the PPI interface of PD-1 or PD-L1, showed obvious inhibitory activities against PD-1/PD-L1 signaling pathways. As the small-molecule inhibitors have better characteristic on aspects like production cost, drug-like property, immunogenic side effects, and half-life period (Liu K. et al., 2016) than peptides and monoclonal antibodies, the development of small-molecule inhibitor tended to be more promising. The crystal structures of small-molecule complex provided a good chance for the drug design of anti-cancer immunotherapy targeting on PD-1/PD-L1 immune checkpoint.

According the patents by Bristol-Myers Squibb (BMS) company, the compounds with (2-methyl-3-biphenylyl)methanol scaffold were privileged for inducing the dimerization of PD-L1 and interacted with the hydrophobic tunnel formed by two PD-L1 monomers (Zak et al., 2016; Guzik et al., 2017). Previously, George F. Gao's group resolved a dimeric interface of PD-L1 formed by B, C′′, D, and E strands on each monomer, which was proved to be either a functional unit in immunological synapse formation or a revolution relics of B7 family (Tan et al., 2016). The crystal lattice analysis by Zak et al. also didn't suggest the spontaneous dimerization of PD-L1 (Zak et al., 2015), indicating that the interfacial interaction between two PD-L1 monomers was quite weak for dimerization process. As for the small molecule intervening PD-L1 dimerization, the interacting interface analysis showed that these ligands interacted with the G, F, C, C′ strands of PD-L1 in a competitive manner vs. PD-1 like mAbs or peptide inhibitors (Sharpe et al., 2011; Liu A. et al., 2016). Specially, the dimerized crystal structures tend to be a common pharmacodynamic characteristics for BMS small-molecule analogs despite of inhibitory activity difference from millimole to nanomole level (Abdel-Magid, 2015; Zak et al., 2016; Skalniak et al., 2017; Perry et al., 2019). Considering the potential relationship between the inhibitory activities of BMS small-molecule inhibitors and the stabilities of the dimerized complex systems, the dimerization process and the structure-activity relationship of small-molecule inhibitors need to be further elucidated. Besides, the broad, scattered and hydrophobic interface on PD-L1 makes it difficult for the discovery of novel small molecule ligands and also results in the strong hydrophobicity of BMS small-molecule inhibitors (Zarganes-Tzitzikas et al., 2016). Therefore, an understanding of the inhibitory mechanism of small-molecule ligands targeting PD-L1 such as key residues at the binding site, effect of the solvation and binding or unbinding process of small molecule inhibitors would help in the discovery of novel inhibitors and structural optimization of reported small-molecule inhibitors.

In this study, we aimed to reveal the detailed molecular mechanism of BMS small-molecule inhibitors from the formation and disassociation of PD-L1 dimers by multiple molecular modeling methods. Two representative compounds (BMS-8 and BMS-1166) with known inhibitory activities and complex crystal structures were selected to perform molecular dynamics simulations. During the formation process, both monomer and dimer systems of PD-L1 in complex with small-molecule ligands were applied to evaluate the stabilities of binding modes between ligands and PD-L1. The binding free energy calculation by MM-PBSA and MM-GBSA (Genheden and Ryde, 2015; Chen et al., 2018; Sun et al., 2018) were also used to analyze the energy contribution of the interfacial residues on PD-L1 dimers. During the disassociation process, metadynamics simulations (Bernardi et al., 2015) with specific collective variables (CVs) were performed to explore the key transition states along unbinding pathways. Based on the results of molecular modeling, an interplay mechanism of BMS small-molecule ligands with PD-L1 was proposed. Finally, R-group based QSAR analysis (Holliday et al., 2003; Hirons et al., 2005) and molecular docking were constructed on the reported BMS small-molecule inhibitors. The results of this study would provide a good guidance for the discovery of novel small-molecule inhibitors and structural modification of BMS small-molecule inhibitors targeting PD-L1.

TABLE 1 | The details of conventional molecular dynamics simulations.

	Dimer systems	Monomer systems	
	PD-L1 dimer	PD-L1 (Conformation A)	PD-L1 (Conformation B)
BMS-8	150 ns	150 ns × 2	150 ns × 2
BMS-1166	150 ns	150 ns × 2	150 ns × 2

METHODS AND MATERIALS

The Conventional Molecular Dynamics Simulations

The complex crystal structures of BMS-8 and BMS-1166 were used to perform conventional molecular dynamics simulations. The Cartesian coordinates of the heavy atoms of PD-L1 (sequence 18–132) and small-molecule ligands were derived from the PDB database with accession number of 5J8O (Zak et al., 2016) and 5NIX (Skalniak et al., 2017). In order to eliminate the electrostatic effect of terminal residues, both monomers were capped with ACE and NME at two ends. The simulation details of the monomer systems and dimer systems were shown in **Table 1**. All the complex systems were firstly prepared through

structural inspection and optimization in Schrödinger 2015 software suite (Schrödinger, LLC: New York, NY, 2015). Then, the complex proteins were solvated in a rectangular box of TIP3P waters and neutralized with Na^+ ions. The periodic boundary conditions were setup with all the solvents at least 10 Å away from the complex. Then, the solvated systems were parameterized using the AMBER FF14SB force field (Case et al., 2014). The molecular dynamics simulations were performed in four steps. Firstly, energy minimization was performed to remove the local atomic collision in the systems. The energy minimization was conducted by both descent steepest method and conjugated gradient method with 5,000 steps. Then, the temperature of each system was gradually upgraded from 0 to 300 K in the NVT ensemble with all the solute atoms constrained with a force constant of 2.0 kcal $mol^{-1}\cdot Å^{-2}$. After that, each system was equilibrated with the force constant decreasing from 2.0 to 0 kcal $mol^{-1}\cdot Å^{-2}$ in a period of 1 ns. Finally, a production run of 150 ns was performed for each system in the NPT ensemble at 300 K and 1.0 atm condition. The snapshots for all the trajectories were saved every 2 ps.

The Binding Free Energy Calculation

For dimer systems of BMS-8 and BMS1166, two PD-L1 monomers were selected as the receptor, while small-molecule inhibitors were selected as the ligand. Both MM-PBSA and MM-GBSA methods were performed to calculate the binding free energy of BMS inhibitors according to the equation below:

$$\Delta G = < G_{Complex} - G_{Receptor} - G_{Ligand} > \quad (1)$$

Where $< >$ represents the average value for all the snapshots used for MM-PBSA and MM-GBSA calculation. Different energy

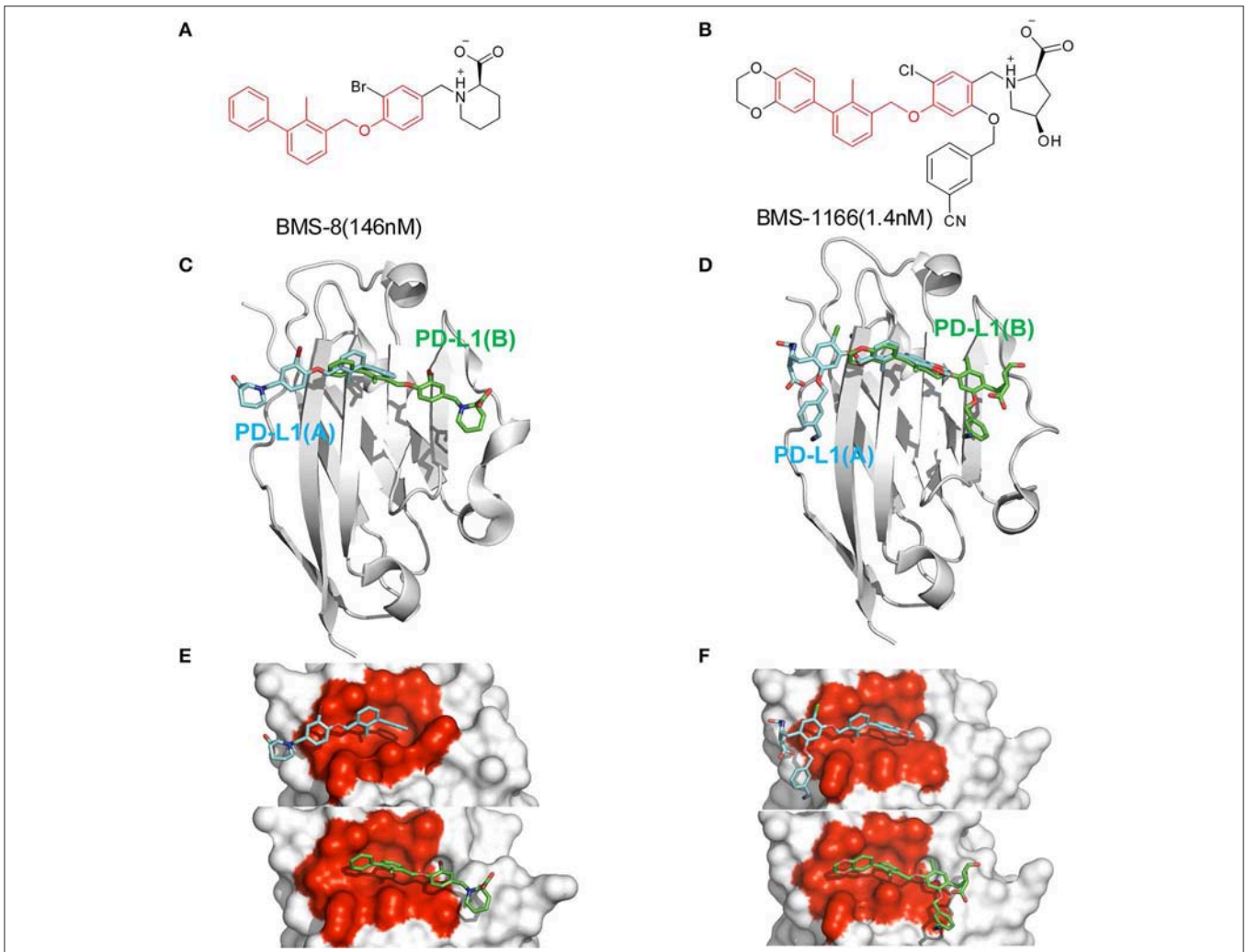


FIGURE 1 | The structural information of BMS-8 and BMS-1166. **(A,B)** The chemical formulas of BMS-8 and BMS-1166. The core scaffold is colored in red. **(C,D)** The conformational superposition of BMS-8 and BMS-1166 interacting with the monomer conformation A, B of PD-L1. **(E,F)** The surface of PD-L1 (A) and PD-L1 (B) interacting with BMS-8 and BMS-1166. The binding pockets formed by I54, V55, Y56, M115, I116, S117, A121, D122, and Y123 were colored in red.

terms can be estimated as follows:

$$\Delta G = \Delta H - T\Delta S \quad (2)$$

$$\Delta H = \Delta E_{gas} + \Delta E_{sol} = \Delta E_{polar} + \Delta E_{nonpolar} \quad (3)$$

$$\Delta E_{gas} = \Delta E_{int} + \Delta E_{ele} + \Delta E_{vdW} \quad (4)$$

$$\Delta H_{sol} = \Delta E_{ele,\ sol} + \Delta E_{nonpl,sol} \quad (5)$$

$$\Delta E_{nonpl,sol} = \gamma * \Delta SASA \quad (6)$$

500 snapshots were extracted from the last 20 ns trajectories and used for MM-PBSA and MM-GBSA calculation. The parameter settings during MM-PBSA and MM-GBSA calculation were referred to the previous works published by our group (Xue et al., 2013). Then, the per-residue based decomposition was performed to identify the key residues in both dimer systems. Finally, the contribution of entropy change ($-T\Delta S$) was calculated by 100 snapshots from the last 20 ns trajectory.

The Calculation of Water Occupancies

The water molecules on the surface affected the conformational stability of proteins (Bellissent-Funel et al., 2016). By calculating the water occupancies on the surface of protein complex, water sites with a higher probability of finding a water molecule could be identified (Gauto et al., 2013). The water molecules at those sites were involved in the water bridges between protein and ligand and could enhance the stability of protein complex thermodynamically (Romero et al., 2016). To evaluate the effects of the water-mediated complex stability upon the binding of BMS inhibitors, the water occupancies and the water bridges were calculated over the last 20 ns trajectories for each dimer system using the "cpptraj" module of the AMBER14. All the

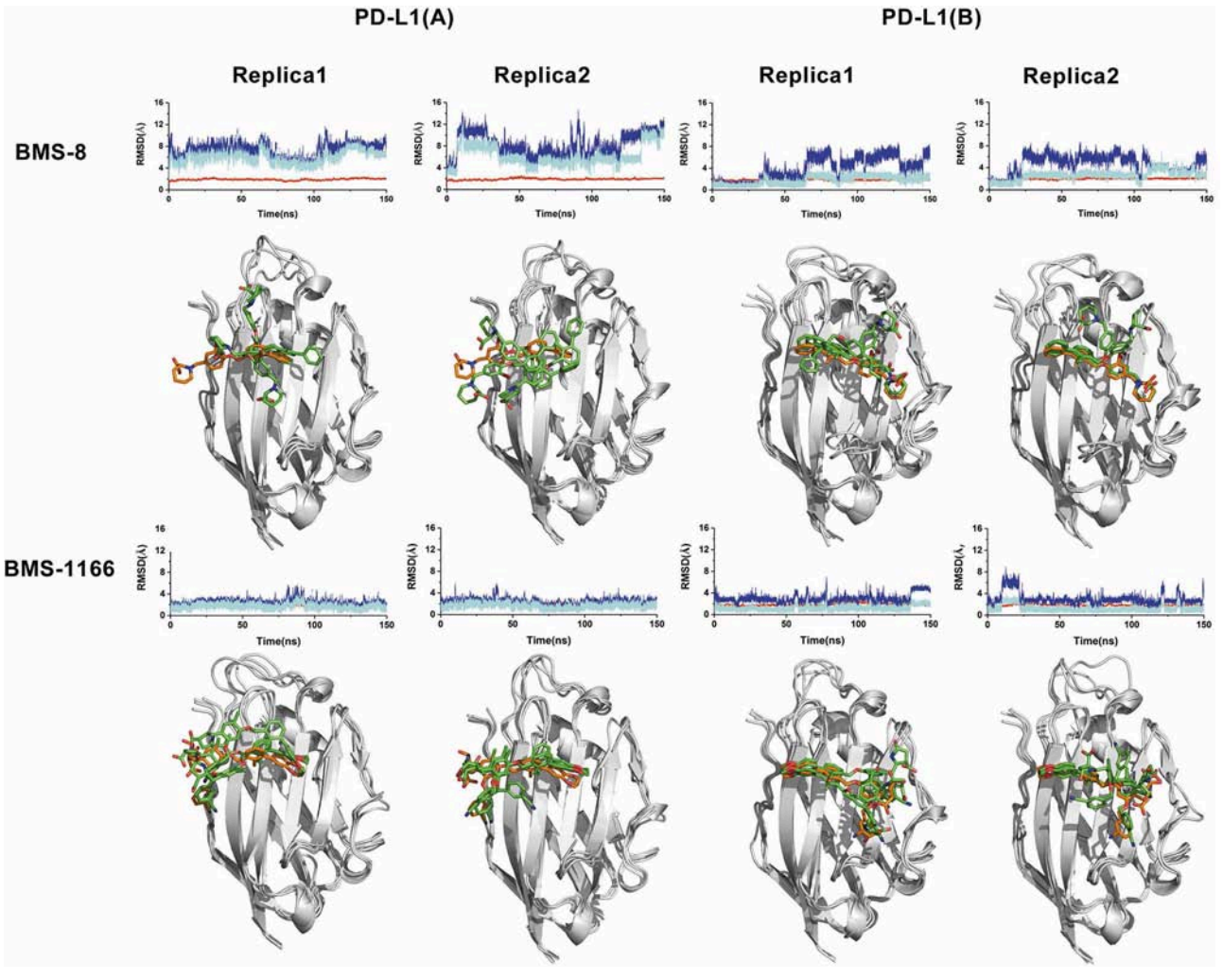


FIGURE 2 | The stability evaluation of the monomer systems. The RMSDs of the heavy atoms of PD-L1 monomer, ligands (BMS-8, BMS-1166) and the core scaffold of the ligand are shown in red, blue, and cyan lines, respectively. The representative conformations of three clusters for every monomer system was shown below the corresponding system. PD-L1 is shown in gray cartoon while the initial conformation of ligand is shown in orange sticks and the dynamics conformations of ligand is shown in green sticks.

trajectories were first imaged and fit to the first frame by the root mean square deviation (RMSD) of the heavy atoms of PD-L1 dimers. Then, the water occupancies were calculated using the "grid" command with a 0.5 Å *0.5 Å * 0.5 Å spacing over the whole box. And the water occupancies for both dimer systems were represented in the Chimera software (Pettersen et al., 2004).

Metadynamics Simulations

Metadynamics simulations have been widely used to predict the unbinding pathways and dissociation energy barrier of ligands for ligand-target systems (Cavalli et al., 2015). The sampling process of metadynamics simulations had an advantage of not requiring an initial estimate of the energy landscape to explore by periodically adding history-dependent biasing potential on selected collective variables (CVs) (Masetti et al., 2009; Barducci et al., 2011; Casasnovas et al., 2017; Sun et al., 2017). In this study, CV1 was selected as the distance between the mass center of the heavy atoms on ligand and the mass center of heavy atoms on key residues including Ile54, Tyr56, Met115, Ala121, Tyr123 in both chains; CV2 was selected as the angle between the C_α atom of Tyr56 and two carbon atoms that were the furthest away from each other on the core scaffold. The metadynamics simulations were performed for both dimer systems. The prepared topology files and coordinate files by AMBER ff14SB force field were further applied in the NAMD2.9 software (Kalé et al., 1999) implemented by PLUMED code (Bonomi et al., 2009). The initial structures were minimized for 5,000 steps with all the atoms on protein and ligand restrained with 5 kcal $mol^{-1}\cdot Å^{-2}$ and all restraints released therewith. Then the temperature of systems were upgraded to 300 K in 30,000 steps. Afterward, all the systems were submitted to two short time NVT simulations (100,000 steps) to equilibrate the systems with restraining force constant of 5 kcal $mol^{-1}\cdot Å^{-2}$ and all restraints released therewith. Finally, the equilibrated structures restarted from the NVT simulation were used for metadynamics simulations.

$$V(s,t) = \sum_{k\tau < t} W(k\tau)\ exp(-\sum_{i=1}^{d} \frac{(s_i - s_i(q(k\tau)))^2}{2\sigma_i^2}) \quad (7)$$

$$W(k\tau) = W_0\ exp(-\frac{V(s(q(k\tau))), k\tau}{K_B \Delta T}) \quad (8)$$

$$\gamma = \frac{T + \Delta T}{T} \quad (9)$$

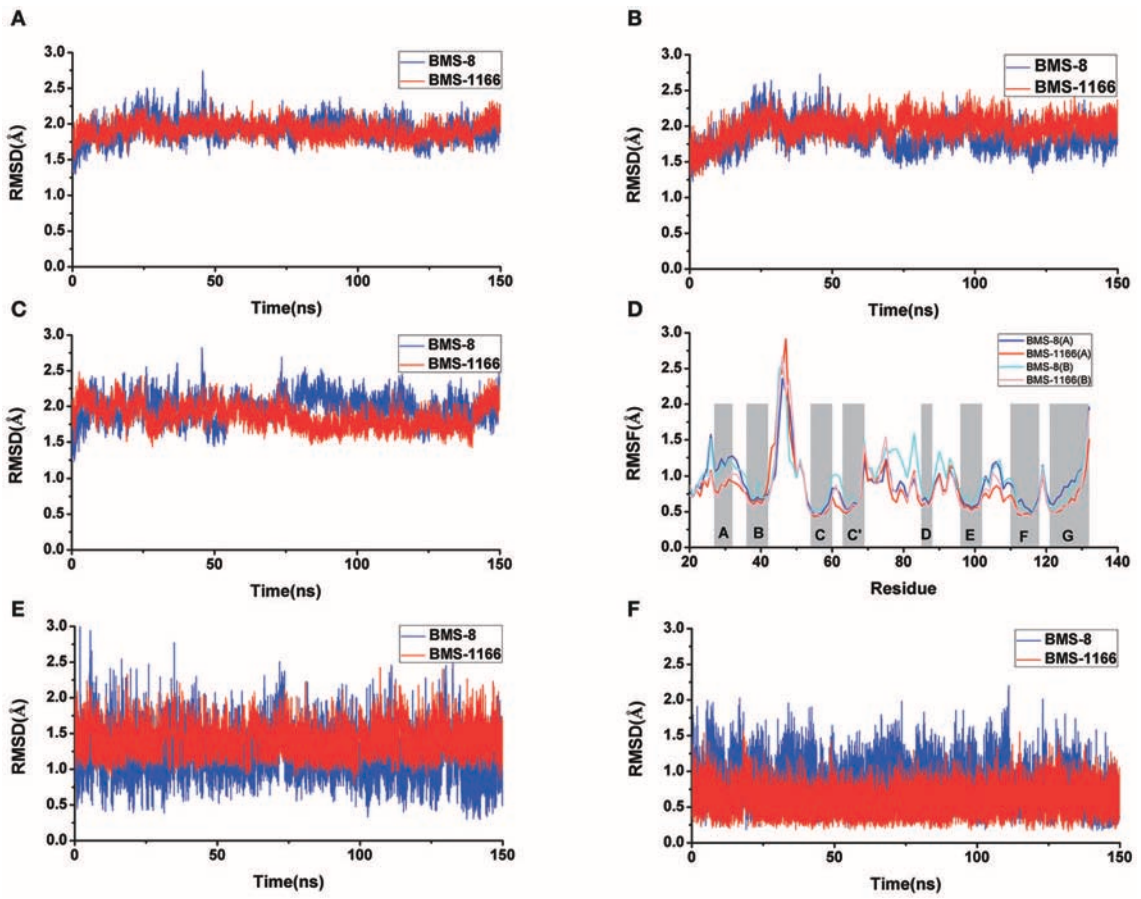


FIGURE 3 | The RMSDs and RMSFs of the heavy atoms in dimer systems. **(A)** RMSD of complex, **(B)** RMSD of PD-L1 at conformation A, **(C)** RMSD of PD-L1 at conformation B, **(D)** RMSF of residues on PD-L1 at conformation A, B, **(E)** RMSD of the ligand, and **(F)** RMSD of the core scaffold.

Metadynamics could reconstruct the free-energy surface as a function of specific collective variables (CVs). The general formalism of history-dependent Gaussian potential was shown as Equation (7). V represents the sum of the history-dependent Gaussian potential along the specific reactive coordinate (s_i) during time span ($k\tau$). In this study, the deposition time (τ) was set as 1 ps to give enough dissociation time for ligands without adding biasing potential on the dissociation boundary. The Gaussian width (σ) of CV1 and CV2 were set to 0.8 Å and 0.02 rad, respectively. As for the well-tempered metadynamics, the height of the Gaussian potential (W) is affected by a parameter ΔT as Equation (8). The initial hill height (W_0) of Gaussian potential was set to 0.6 kcal/mol·ps and the bias-factor (γ) was set to 10 with a temperature (T) of 300 K to control the decrease rate of the biasing potential as Equation (9).

R-Group QSAR Model and Molecular Docking of BMS Small-Molecule Inhibitors

The pharma R-group quantitative structure-activity relationship (RQSAR) models tended to be an effective approach for the SAR evaluation of the congeneric series of compounds (Adhikari et al., 2015). It was more suggestive than other approaches for the structural modification of small-molecule inhibitors by identifying the core scaffold and evaluating the effective pharma element at different attachment points (Kolarevic et al., 2018; Ts Mavrova et al., 2018). A total of 110 BMS small-molecule inhibitors with 2-methyl-3-(phenoxymethyl)-1,1'-biphenyl scaffold were collected from the patents of BMS company (Abdel-Magid, 2015; **Table S1**). All these small molecules had seven attachment points and diverse substitution groups, which were suitable to perform R-group QSAR analysis in the Canvas software of Schrödinger Suite (Duan et al., 2010; Sastry et al., 2010). The linear relationship between the substitutions and the activities (–log IC_{50}) was analyzed and the importance of six key pharmacophore elements including hydrogen bond acceptor (A), hydrogen bond donor (D), hydrophobic group (H), negative ionic group (N), positive ionic group (P), and aromatic ring (A) were evaluated at each attachment point. During the process, the error and the importance were both set as 0.30. Eight representative small-molecule inhibitors (NO. of compound:4, 101, 102, 103, 104, 108, 109, 110) with substitutions on R1, R2, or R3 were selected to perform molecular docking to further study the binding modes. In order the compare the effect of R-groups, the core scaffold atoms with SMILES of "cOCc(c1C)cccc1c" were constrained with RMSD of 0.5 angstrom while other atoms were selected flexible. The standard precision (SP) docking score was used to evaluate the binding poses. The molecular docking was performed in Schrödinger 2015 software suite (Schrödinger, LLC: New York, NY, 2015).

Residue-Ligand Contact Analysis

In this study, we performed residue-ligand contact analysis to detect the surrounding residues around different substituent groups of BMS-8 and BMS-1166. It is assumed that the contacts exist between two groups as long as their distance was below a cutoff of 3.5 Å. The occupancy of each contact was calculated by the existence frequency in the 5,000 snapshots of the last 50 ns trajectories.

TABLE 2 | The binding free energies of BMS-8 and BMS-1166 evaluated by MM-PBSA, MM-GBSA, and metadynamics simulations.

Contribution[a]	BMS-8	BMS-1166	ΔΔE ($\Delta E_{BMS-1166} - \Delta E_{BMS-8}$)
$\Delta E_{ele,gas}$	−29.29 ± 6.14	−72.23 ± 12.05	−42.94 ± 13.52
$\Delta E_{vdw,gas}$	−63.45 ± 2.77	−81.60 ± 3.72	−18.15 ± 4.64
$\Delta E_{nonpl,sol}$	−6.62 ± 0.18	−9.17 ± 0.19	−2.55 ± 0.26
$\Delta E_{polar,sol,PB}$	49.89 ± 7.19	105.28 ± 10.16	55.39 ± 12.45
$\Delta E_{sol,PB}$	43.27 ± 7.07	96.11 ± 10.11	52.84 ± 12.34
$\Delta E_{polar,sol,GB}$	45.04 ± 5.85	93.32 ± 9.91	48.28 ± 11.51
$\Delta E_{sol,GB}$	38.42 ± 5.75	84.15 ± 9.88	45.73 ± 11.43
ΔH_{PB}	−49.49 ± 3.71	−57.72 ± 4.64	−8.23 ± 5.94
ΔH_{GB}	−54.32 ± 3.14	−69.69 ± 3.32	−15.37 ± 4.57
$-T\Delta S$	24.72 ± 5.75	27.02 ± 5.89	2.30 ± 8.23
ΔG_{PB}	−24.77 ± 6.84	−30.70 ± 7.50	−5.93 ± 10.15
ΔG_{GB}	−29.60 ± 6.55	−42.67 ± 6.76	−13.07 ± 9.41
ΔE_{CV1}[b]	−16.23	−28.79	−12.56
ΔE_{CV2}[c]	−15.51	−27.89	−12.38
ΔE_{exp}[d]	−9.32	−12.07	−2.75

[a] *The unit for the free energy contributions are shown in kcal/mol.*

[b,c] *ΔE_{CV1} and ΔE_{CV2} were estimated by the history-dependent free energy surfaces along CV1 and CV2.*

[d] *The experimental affinities for BMS-8 and BMS-1166 were extracted from the reference and calculated by using the equation as follows: $\Delta G = -RT \ln(1/IC_{50})$ at 298.15 K.*

RESULTS AND DISCUSSION

The Conformational Stabilities Between PD-L1 Monomer and BMS Small-Molecule Inhibitors

In order to explore the interactive process of BMS small-molecule inhibitors, we constructed two kinds of ligand-bound PD-L1 monomer systems as shown in **Figure 1** and used molecular dynamics simulations to evaluate the stabilities of both binding modes by two replicas. As shown in **Figure 2**, the stabilities of both binding modes were evaluated by the RMSD of the heavy atoms of receptor, ligand and the core scaffold of ligands in two replicas. The core scaffold of BMS-8 tended to have a more stable contact with the conformation B than conformation A of PD-L1 according to the comparison of RMSD and the representative conformations of both binding modes. The detailed docking interactions diagram in **Figures S1**, **S2** showed that the π-π stacking interaction between the biphenyl moiety and $_A$Tyr56 tended to be easily affected by the conformation of $_A$Tyr56 and unstable among three clusters, while the hydrophobic interaction between biphenyl moiety and residues on conformation B of PD-L1 tended to be stable among all three clusters. As for BMS-1166, both binding modes seemed to be quite stable, which probably accounting for the best inhibitory activities of BMS-1166 among

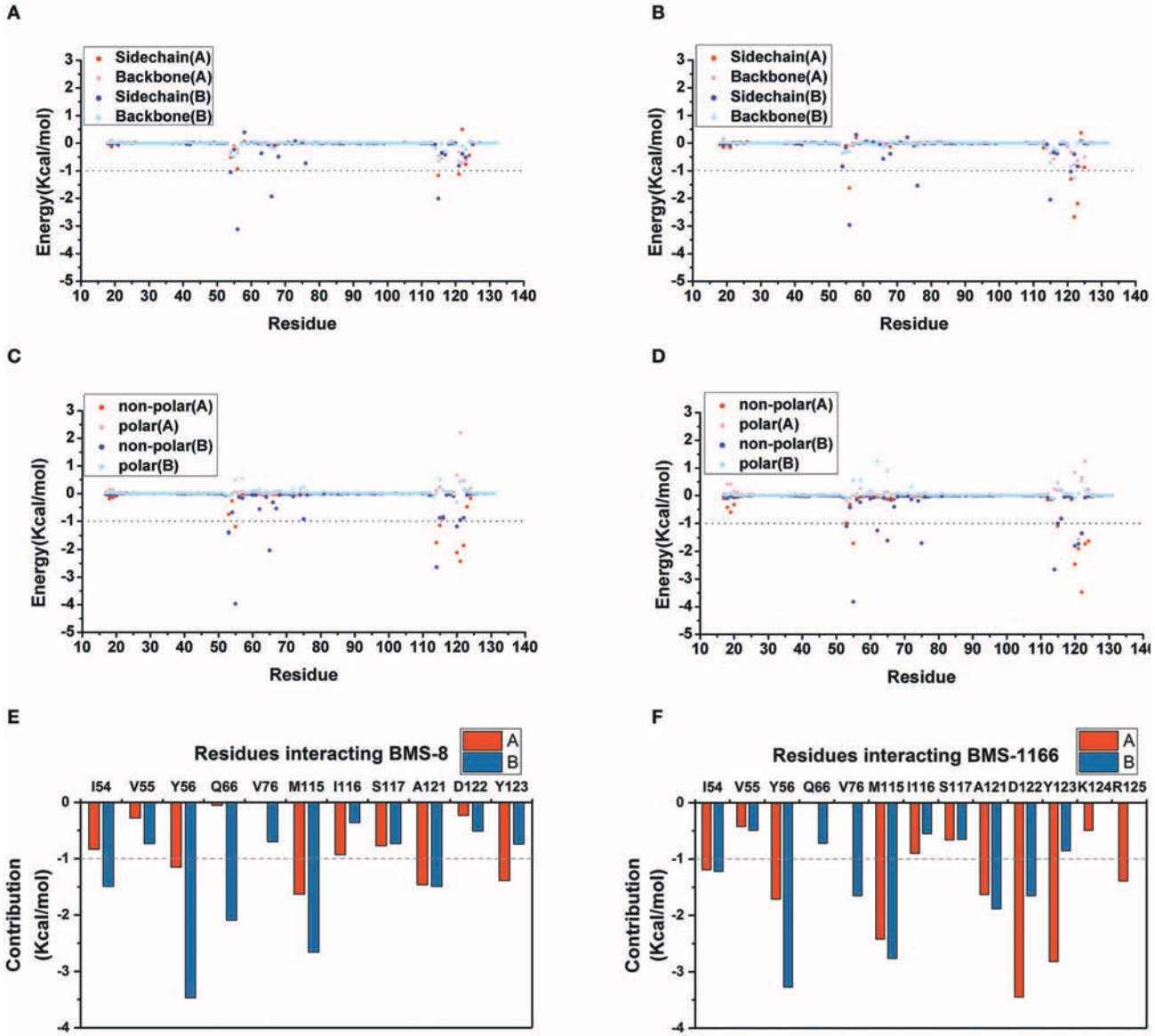


FIGURE 4 | The residue energy decomposition of the key residues for BMS-8 **(A,C,E)** and BMS-1166 **(B,D,F)** dimer systems. **(A,B)** The energy contribution of sidechain and backbone, **(C,D)** the energy contribution of polar and non-polar interaction, **(E,F)** the total energy contribution for each residue. The cutoff value of the energy contribution for key residues were set as −1 kcal/mol.

TABLE 3 | The hydrogen bond analysis of the BMS-8 and BMS-1166 dimer systems.

Acceptor	DonorH	Donor	Occupancy[a] (%)	Distance (Å)[b]	Angle (°)[c]
$_{B}$Gln66@OE1	BMS-8@H1	BMS-8@N1	57.21	2.90	162.74
$_{A}$Asp122@OD2	BMS-1166@H1	BMS-1116@O5	63.54	2.71	161.10
$_{A}$Asp122@OD1	BMS-1166@H1	BMS-1116@O5	47.58	2.93	151.41
$_{A}$Asp122@OD1	BMS-1166@H20	BMS-1116@N2	41.43	2.93	144.20
$_{A}$Asp122@OD2	BMS-1166@H20	BMS-1116@N2	29.74	3.01	140.98
$_{A}$Asp122@OD1	BMS-1166@H19	BMS-1116@N2	21.02	2.90	139.00
BMS-1166@N1	$_{A}$Arg125@H	$_{A}$Arg125@N	81.70	3.08	149.17
BMS-1166@O5	$_{A}$Thr20@HG1	$_{A}$Thr20@OG1	36.64	2.95	154.22

[a]*The occupancy of hydrogen bonds were analyzed through the last 20 ns trajectories and only hydrogen bonds with an occupancy more than 0.20 were extracted and shown.*
[b,c]*The hydrogen bonds were determined by an acceptor-donor atom distance of <3.5 Å and acceptor H-donor angle of >120°.*

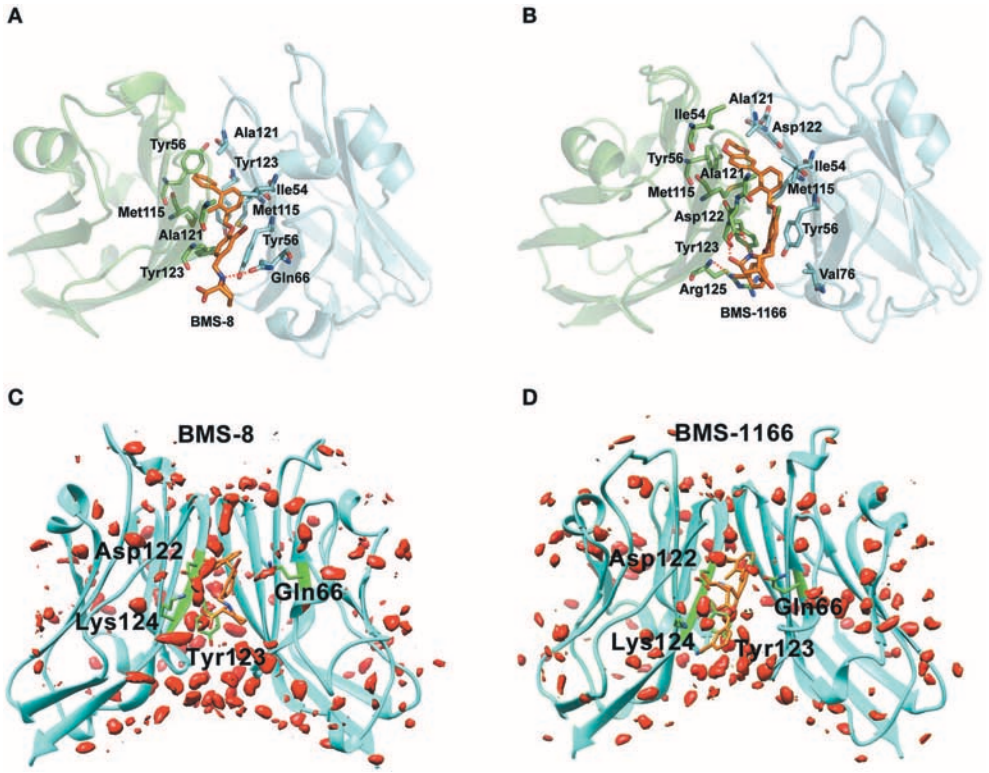


FIGURE 5 | The binding modes and water occupancies in dimer systems. **(A,B)** The binding modes of BMS-8 and BMS-1166. The key residues were shown in green and cyan sticks and the ligand was shown in orange sticks. The hydrogen bonds were shown in red dash. **(C,D)** The water occupancies in BMS-8 and BMS-1166 dimer systems. The PD-L1 dimer is shown in cyan cartoon and the residues involved in water bridges are shown in green sticks. The water distributions are shown in red solid surface and the small molecule ligands are shown in orange sticks.

TABLE 4 | The water bridge with occupancies higher than 20.00% in the BMS-8 and BMS-1166 dimer systems.

Residues involving in water bridge	BMS-8 (%)	BMS-1166 (%)
$_{A}$Thr20	–	32.42
$_{A}$Asp122	25.04	–
$_{A}$Lys124	91.57	94.15
$_{A}$Arg125	–	22.63
$_{B}$Gln66	58.62	32.93
$_{B}$Lys75	–	22.58
$_{B}$Val76	–	50.13
$_{B}$Asp122	–	29.66
$_{A}$Phe19, $_{A}$Ala121, $_{A}$Asp122	–	23.52
$_{A}$Asp122, $_{A}$Tyr123, $_{A}$Lys124	38.90	65.18

the small-molecule inhibitors of BMS. The detailed docking interactions diagram in **Figures S3**, **S4** showed that the biphenyl moiety had less conformational fluctuation and more stable hydrophobic interactions among the clusters of conformation A and B of PD-L1. The stabilities of the monomer complex of PD-L1 and ligand was affected by the hydrophobic interactions and turned out to be associated with the inhibitory activities of BMS small-molecule inhibitors.

The Interaction Stabilities Between PD-L1 Dimer and BMS Small-Molecule Inhibitors

The conformational stabilities of the dimer systems were evaluated by root mean square deviation (RMSD) and mean square root fluctuation (RMSF) as shown in **Figure 3**. The RMSDs of the complex, two PD-L1 monomers in complex systems showed that both dimer and monomers of PD-L1 had strong structural stabilities upon ligand binding. The conformational fluctuation of PD-L1 indicated that PD-L1 showed more flexibilities upon BMS-8 binding than BMS-1166. The comparison of the RMSDs of the core scaffold of two ligands showed that BMS-1166 had a more stable binding modes than BMS-8. The binding free energies were also calculated to evaluate the affinities of dimer systems. As shown in **Table 2**, the energy items of $\Delta\Delta G_{PB}$, $\Delta\Delta G_{GB,}$ and $\Delta\Delta E_{exp}$ by MM-PBSA and MM-GBSA methods could properly evaluate the difference of affinities of BMS-8 and BMS-1166, which showed the fact that the affinities between small-molecule inhibitors and PD-L1 dimer could reflect the inhibitory activities relatively. BMS-1166 had a stronger enthalpy contribution ($\Delta\Delta H_{PB}$, $\Delta\Delta H_{GB}$) and a worse entropy contribution ($-T\Delta\Delta S$) than BMS-8, which were consistent with the stability difference of BMS-8 and BMS-1166 dimer complex.

The key residues on two PD-L1 monomers interacting with ligands were recognized by per-residue energy decomposition.

The energy contribution for each residue were decomposed into the sidechain part and the backbone part, the non-polar part and the polar part as shown in **Figures 4A–D**. It could be seen that BMS-8 and BMS-1166 mainly formed non-polar interactions with the sidechain of the residues on PD-L1. With a cutoff value of −1.0 kcal/mol, the key residues in BMS-8 dimer system included $_{A}$Tyr56, $_{A}$Met115, $_{A}$Ala121, $_{A}$Tyr123 and $_{B}$Ile54, $_{B}$Tyr56, $_{B}$Gln66, $_{B}$Met115, $_{B}$Ala121 as shown in **Figure 4E**, while the key residues in BMS-1166 dimer system included $_{A}$Ile54, $_{A}$Tyr56, $_{A}$Met115, $_{A}$Ala121, $_{A}$Asp122, $_{A}$Tyr123, $_{A}$Arg125 and $_{B}$Ile54, $_{B}$Tyr56, $_{B}$Val76, $_{B}$Met115, $_{B}$Ala121, $_{B}$Asp122 as shown in **Figure 4F**. Taken together, the interaction residues on conformation A and conformation B of PD-L1 were symmetrical both including Ile54, Tyr56, Met115, Ala121, and Tyr123. The hydrogen bond analysis in **Table 3**, **Figures 5A,B** showed that the protonated tertiary ammonium in BMS-8 formed a hydrogen bond with the side-chain oxygen of $_{B}$Gln66 with an occupancy of 57.21%, while the BMS-1166 dimer system also formed hydrogen bond between the ammonium group on BMS-1166 and the carboxyl group of $_{A}$Asp122. The binding mode analysis of substitute groups on BMS-8 and BMS-1166 with the interfacial residues

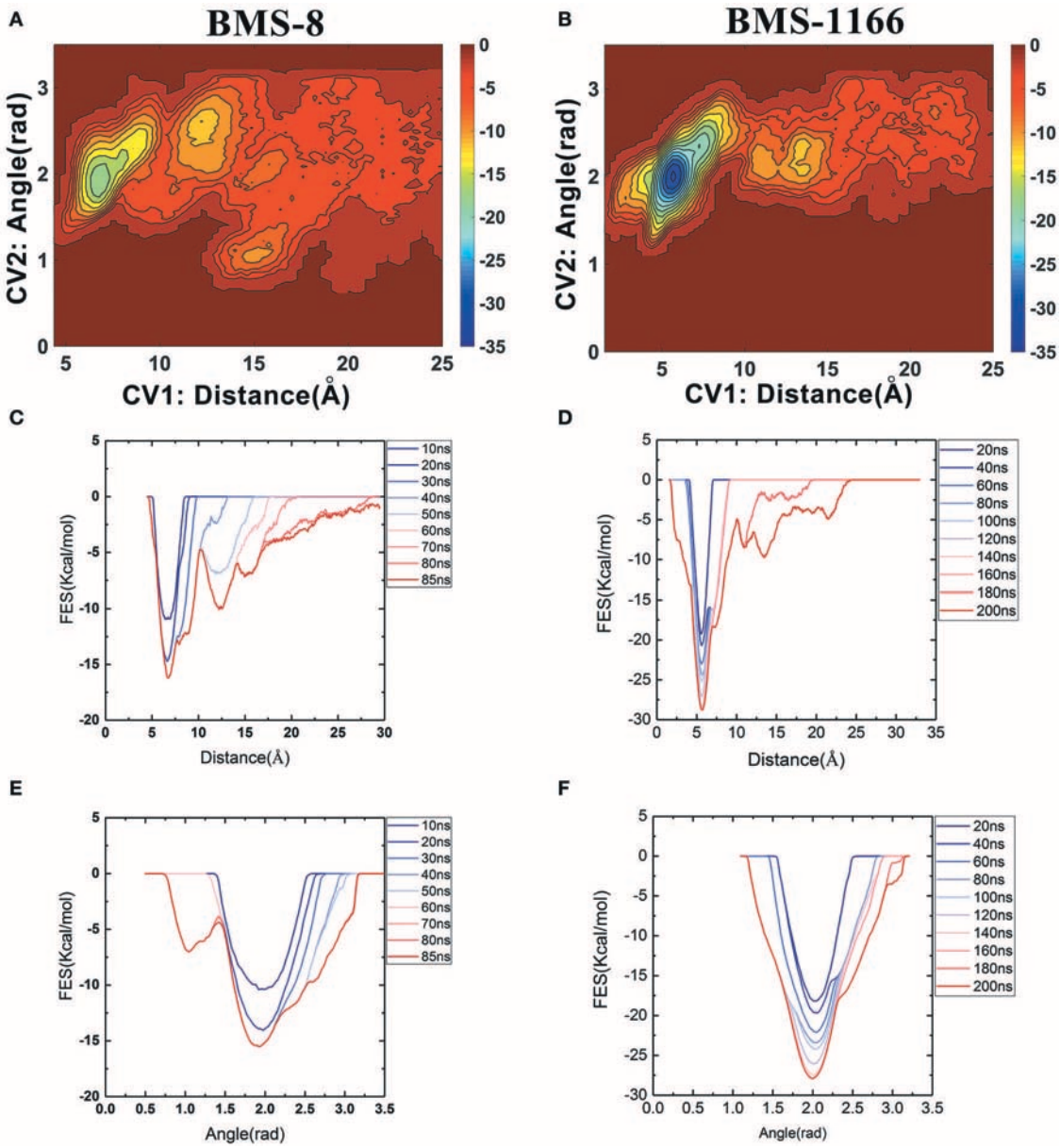


FIGURE 6 | The energy change during the unbinding process. **(A,B)** The free energy landscapes for the unbinding process of BMS-8 and BMS-1166. **(C–F)** The convergence of sampling process during the unbinding process of BMS-8 and BMS-1166. The history-dependent free energy surfaces along CV1 **(C,D)** and CV2 **(E,F)** are estimated by a segmented accumulation of simulation time. The unit for the free energy is kcal/mol.

on PD-L1 indicated that the interaction with the peripheral residues including $_{A}$Ile54, $_{A}$Arg125, $_{B}$Val76, and $_{B}$Asp122 could significantly enhance the inhibitory activities of BMS small-molecule inhibitors.

In order to analyze the effect of solvent on PD-L1 dimer complex, water occupancies and water bridges involved in receptor-ligand interaction were both calculated. As shown in **Table 4**, the residues or residue pairs involved in water bridges with an occupancy higher than 20% were extracted from both dimer systems. It can be seen in **Figure 5** that three water bridges involved in $_{A}$Asp122, $_{A}$Tyr123, $_{A}$Lys124, and $_{B}$Gln66 were stable in both dimer systems. Both ligands formed a strong water bridge with $_{A}$Lys124 with an occupancy higher than 90%, which indicated that $_{A}$Lys124 had a significant effect on the stabilities of the ligand conformations.

The Disassociation Process of BMS Small-Molecule Inhibitors

The free energy landscape of the unbinding processes of both BMS small-molecule inhibitors were constructed by CV1 and CV2. The distribution of minima in the landscapes showed that the most stable conformational state in the unbinding

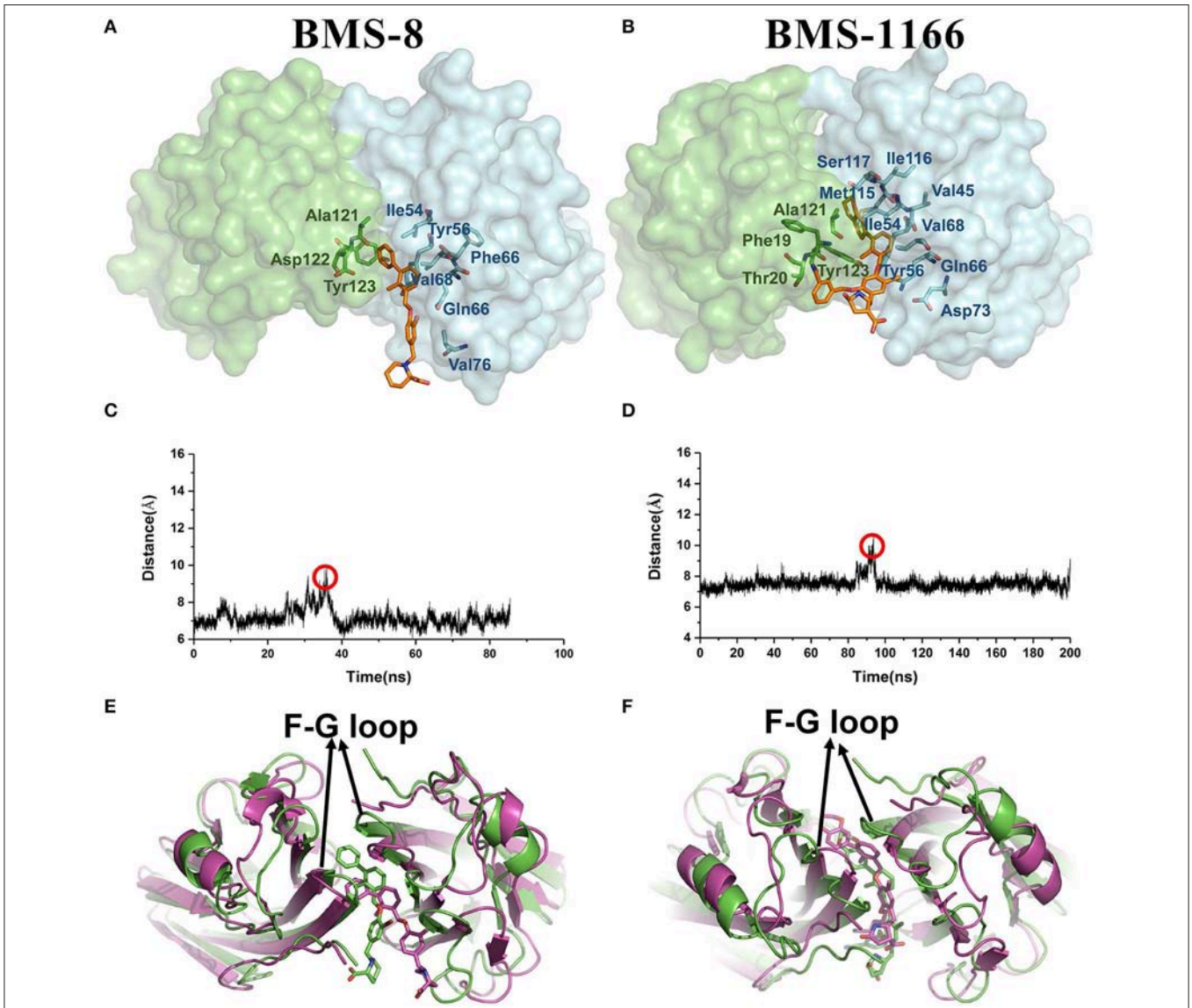


FIGURE 7 | The conformational change during the unbinding process. **(A,B)** The key transition conformational states of BMS-8 and BMS-1166 during the unbinding process with CV1 value of 12.54 Å and 10.95 Å, respectively. The chain A and chain B of PD-L1 dimers are shown in green and cyan surface, respectively. The key residues (green or cyan) and ligands (orange) are shown in sticks. **(C,D)** The distance between Ile54, Tyr56, Met115, Ala121, and Tyr123 on conformation A and conformation B of PD-L1 during the unbinding process of BMS-8 and BMS-1166. **(E,F)** The extracted conformational state (green cartoon) of each complex with the largest distance between two PD-L1 monomers was overlapped with the original crystal structures (magenta cartoon).

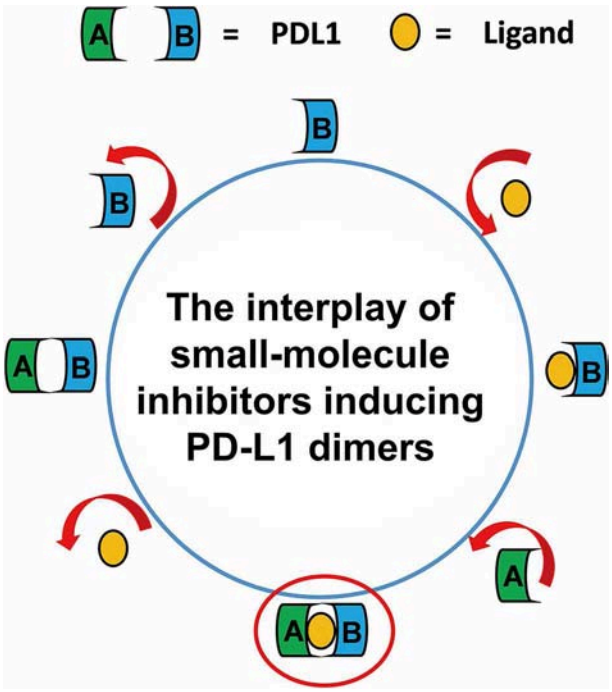


FIGURE 8 | The complete binding and unbinding mechanism of BMS small-molecule inhibitors. The conformation A and conformation B of PD-L1 monomer, the small-molecule ligand are represented by green, light blue, orange objects, respectively.

process was corresponding to the conformational states of the initial crystal structures as shown in **Figures 6A,B**. During the unbinding process, there were four different transition states for BMS-8 and three transition states for BMS-1166. In order to test the convergence of unbinding process, the free energies along both CVs were estimated. It can be seen that the free energy surface of CV1 (**Figures 6C,D**) and CV2 (**Figures 6E,F**) gradually came to a convergence along with the accumulation of time. As CV1 represented the distance between the ligand and the binding site of PD-L1 dimer and depicted the unbinding process better than CV2, the corresponding minimum points along CV1 were extracted from the unbinding trajectories. In BMS-8 complex systems, the minima along CV1 were 6.75 Å (−16.34 kcal/mol), 12.54 Å (−10.12 kcal/mol), and 14.97 Å (−7.31 kcal/mol). In BMS-1166 complex systems, the minima along CV1 were 5.67 Å (−28.79 kcal/mol), 10.95 Å (−8.50 kcal/mol), and 13.46 Å (−9.74 kcal/mol). The ultimate unbinding energy barriers of both small-molecule ligands estimated by CV1 and CV2 were shown in **Table 1** and **Figure 6**, which were in good consistency with the inhibitory activities. Considering the difference between the binding free energies predicted by different methods, MM-PBSA and MM-GBSA calculated the binding free energies using implicit water models while metadynamics simulation considered the explicit water interaction between protein and ligand. Therefore, the results from metadynamics simulation tended to be more approximate to the experimental results.

From the free energy estimation of different conformational states, it can be seen that the conformational states of the crystal structures were much more stable than the other transition conformational states along the unbinding process. Therefore, the dissociation of small-molecule ligands of the initial conformational states tended to be the most important intermediate process for the unbinding of small-molecule ligands, which were corresponding to the minima of CV1 at 12.54 Å in BMS-8 dimer systems and the minima of CV1 at 10.95 Å in BMS-202 dimer systems. The corresponding transition states were extracted from the trajectories as shown in **Figures 7A,B**. The binding poses of BMS small-molecule ligands at the transition states were quite distinct from each other, which was probably owing to the difference of substituent groups. A common feature for both systems was that the ligands at transition states significantly lost the interaction with the chain A while the interaction with chain B were still compact and involved with a series of residues especially in BMS-1166 dimer systems. During the unbinding process, the core scaffold of ligands gradually divorced from the location of $_{A}$Tyr56 and got away from the pocket formed by PD-L1 dimer. In order to monitor the conformational change of the pocket formed by PD-L1 monomers, the distance between chain A and chain B were calculated by the distance between Ile54, Tyr56, Met115, Ala121, Tyr123 on each chain as shown in **Figures 7C–F**. The conformational fluctuation of PD-L1 monomers was reflected by the conformational change of the F-G loops on both PD-L1 monomers. It can be seen that the pockets in BMS-8 and BMS-1166 complex systems were quite stable with occasionally occurring conformational fluctuations. According to unbinding processes of BMS ligands, it can be seen that the dimer of PD-L1 had a large tendency to keep stable although accompanied with subtle conformational fluctuation of PD-L1 dimer.

Taken together, the most possible deduction for the interaction mechanism of BMS small-molecule inhibitors with PD-L1 was depicted as shown in **Figure 8**. Firstly, all BMS small-molecule inhibitors with different activities tended to interacted with a monomer conformation B of PD-L1. As the PD-L1 dimer complex had strong conformational stability, the PD-L1 monomer complex further interacted with the other monomer of PD-L1 to form PD-L1 dimer complex. According to the results of metadynamics simulation, a complete dissociation for BMS inhibitors would probably be like that the small-molecule ligand was firstly unbound from the PD-L1 dimer and the rest receptor part was further depolymerized into monomer.

The R-Group QSAR Model of BMS Small-Molecule Inhibitors

110 BMS small-molecule inhibitors with 2-methyl-3-(phenoxymethyl)-1,1′-biphenyl scaffolds were tested with diverse inhibitory activities with IC_{50} ranging from 9.492 μM to 1.4 nM. As shown in **Figure 9A**, there were 7 different of attachment points from R1 to R7 and the substituent groups of R6 and R7 had a relatively larger proportion than other

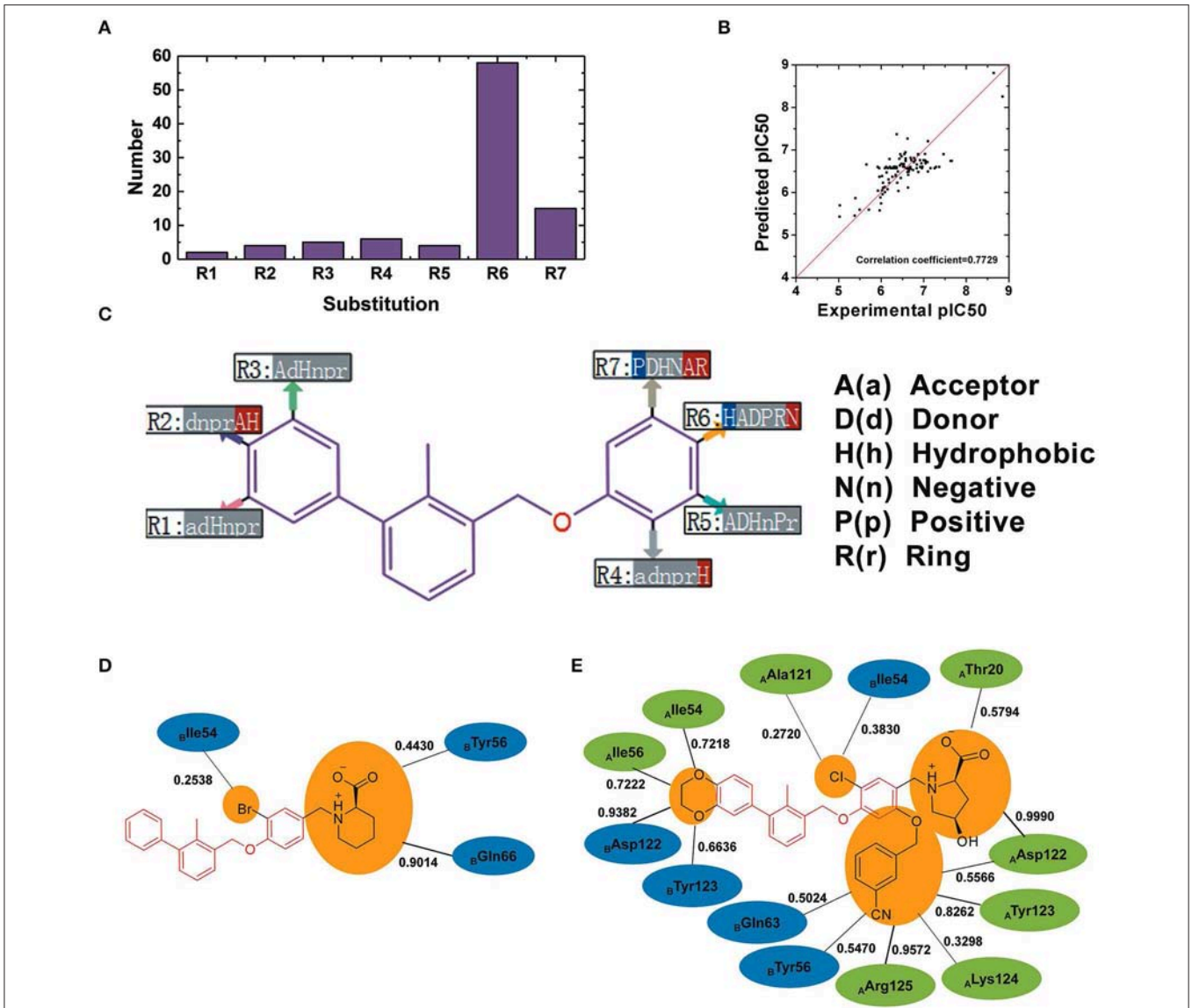


FIGURE 9 | The R-group based QSAR for BMS small-molecule inhibitors. **(A)** The number of substituent groups at seven attachment points of BMS small molecule inhibitors. **(B)** The correlation validation between the predicted pIC50 and experimental pIC50. **(C)** R-group QSAR model for the BMS small molecule inhibitors. The case-insensitive alphabets A(a), D(d), H(h), N(n), P(p), and R(r), respectively represent the hydrogen bond acceptor, the hydrogen bond donor, the hydrophobic, the negative ionic, the positive ionic, aromatic ring. The significantly increase effect is colored in red while the significantly decrease effect is colored in blue. **(D,E)** The interacting residues of different substituent groups of BMS-8 and BMS-1166. The occupancies of the contact between each residues and substituent group were listed along with the black solid.

attachment points. As shown in **Figure 9B**, the correlation coefficient between the predicted pIC_{50} and the experimental pIC_{50} was 0.7729. According to the evaluation of six key pharmacophore elements in **Figure 9C**, the substituent groups at R2, R4, R6, and R7 had obvious effect on the affinity of BMS small-molecule inhibitors. The substituent groups at R2, R4, R6, R7 of BMS-8 and BMS-116 as well as the interaction residues were recognized by the contact analysis as shown in **Figures 9D,E**. The contact analysis of BMS-1166 showed that the 1,4-benzodioxinyl group at R2 mainly was involved in the interaction with $_{A}$Ile54, $_{A}$Tyr56, $_{B}$Asp122, $_{B}$Tyr123. The hydrogen bond acceptor and hydrophobic groups at R2 were favorable for BMS inhibitors such as the 2, 3-dihydro-1, 4-benzodioxinyl group on BMS-114, BMS-200, BMS-1001, and BMS-1166. The analysis of effect of solvent in dimer systems showed that the substituent groups at R4, R5, R6, and R7 were exposed to solvent environment. The hydrophobic groups at R4 were favorable for BMS inhibitors, which corresponded to the fact that the bromine atom on BMS-8 and the chlorine atom on BMS-1166 had a close contact with $_{B}$Ile54. The hydrophobic group at R6 was adverse while the negative ionic group was favorable. The substituent group at R6 of BMS-8 mainly

interacted with $_B$Tyr56 and $_B$Gln66, however, that of BMS-1166 mainly interacted with $_A$Thr20 and $_A$Asp122. Nevertheless, the substituent group at R6 of BMS-8 and BMS-1166 both formed hydrogen bonding with PD-L1. The positive ionic at R7 was adverse while the hydrogen bond acceptor and aromatic ring were favorable. The substituent group at R7 of BMS-1166 formed interaction with $_A$Asp122, $_A$Tyr123, $_A$Lys124, $_A$Arg125, $_B$Tyr56, and $_B$Gln63. The comparison of the contact residues between BMS-8 and BMS-1166 indicated that the substituent group at R2 and R7 strongly strengthened the interactions with the conformation A of PD-L1, which was consistent with the stability of the monomer complex of conformation A and BMS-1166.

The further molecular docking study of eight representative small-molecule inhibitors showed that the docking scores had a good linear correlation with the experimental inhibitory activities (**Figure S5**). The further residue contribution comparison (**Figure S6**) and conformational analysis (**Figure S7**) of the residues within 5 angstroms showed that the residues interacting with R-group substituents had an obvious effect on the docking scores including $_B$Asp122 (interacting with R1 to R3) and $_A$Asp122, $_A$Lys124, $_B$Tyr56, $_B$Gln66 (interacting with R4 to R7). The binding mode analysis of novel series of [1,2,4]triazolo[4,3-a]pyridines designed by Qin et al. also revealed the retaining hydrophobic interaction with Tyr56, Met115, and Ala121 on both chain of PD-L1 and extra π-π stacking with the $_B$Tyr56 and π-anion interactions with $_A$Asp122 (Qin et al., 2019). These interacting modes of [1,2,4]triazolo[4,3-a]pyridines inhibitors were consistent with the binding mode analysis of eight representative small-molecule inhibitors. It's suggested that the structure-activity relationship analysis of BMS small-molecule inhibitors was applicable for the further structure modifications.

CONCLUSIONS

In this study, we used multiple molecular modeling methods to study the detailed molecular mechanism of the interaction between BMS small-molecule inhibitors and PD-L1. A detailed mechanism of the interaction process between small-molecule inhibitors and PD-L1 was proposed and validated by molecular dynamics simulations.

The BMS small-molecule inhibitors tended to interact with one PD-L1 monomer first and further formed dimer with the other monomer for an advantage of stability. The results of binding free energy and water occupancy calculation revealed the key stability factors for ligand-induced PD-L1 dimers including the hydrophobic contribution of Ile54, Tyr56, Met115, Ala121, and Tyr123 on both monomers and the water bridges involved in $_A$Lys124. The unbinding pathway prediction also indicated that the tunnel formed by PD-L1 dimers tended to be stable upon the getting away of BMS-inhibitors. The R-group QSAR model suggested that the substituents at R2, R4, R6, and R7 had a significant effect on the inhibition activities of BMS inhibitors. The structural modification with these substituent positions tended to be an effective way to improve the inhibition activities of BMS inhibitors. Taken together, this study would provide a comprehensive view of the inhibition mechanism for BMS small-molecule inhibitors and guide the further development of more potential small-molecule inhibitors targeting PD-L1.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

DS, HL, and XY designed the research. DS and XY were responsible for the writing and revising of the manuscript. XA, QB, SZ, and ZB were responsible the main data analysis in the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 21475054 and 2175060).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00764/full#supplementary-material

Figure S1 | The interactions diagrams between the monomer conformation A of PD-L1 (the initial crystal structure and three respective dynamics structures) and BMS-8 in the monomer system of replica 2.

Figure S2 | The interaction diagrams between the monomer conformation B of PD-L1 (the initial crystal structure and three respective dynamics structures) and BMS-8 in the monomer system of replica 2.

Figure S3 | The interaction diagrams between the monomer conformation A of PD-L1 (the initial crystal structure and three respective dynamics structures) and BMS-1166 in the monomer system of replica 2.

Figure S4 | The interaction diagrams between the monomer conformation B of PD-L1 (the initial crystal structure and three respective dynamics structures) and BMS-1166 in the monomer system of replica 2.

Figure S5 | The linear correlation between experimental pIC50 and the absolute values of the docking scores.

Figure S6 | The distance and residue contribution analysis of the binding poses of eight representative small-molecule inhibitors. **(A)** The respective and average distance between the small-molecule inhibitor and the residues on PD-L1 dimer. **(B)** The respective energy contribution of residues on PD-L1 dimer when interacting with the small-molecule inhibitor.

Figure S7 | The binding pose analysis of eight representative small-molecule inhibitors. **(A)** The surrounding residues of the substituent groups at R1 to R3 for eight representative small-molecule inhibitors. **(B–I)** The surrounding residues of the substituent groups at R4 to R7 for small-molecule inhibitor with NO. of 4, 101, 102, 103, 104, 108, 119, 110, respectively.

Table S1 | The detailed structural and activity information for 110 BMS small-molecule inhibitors.

REFERENCES

Abdel-Magid, A. F. (2015). Inhibitors of the PD-1/PD-L1 pathway can mobilize the immune system: an innovative potential therapy for cancer and chronic infections. *ACS Med. Chem. Lett.* 6, 489–490. doi: 10.1021/acsmedchemlett.5b00148

Adhikari, N., Halder, A. K., Saha, A., Das Saha, K., and Jha, T. (2015). Structural findings of phenylindoles as cytotoxic antimitotic agents in human breast cancer cell lines through multiple validated QSAR studies. *Toxicol. In Vitro* 29, 1392–1404. doi: 10.1016/j.tiv.2015.05.017

Barducci, A., Bonomi, M., and Parrinello, M. (2011). Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol.* 1, 826–843. doi: 10.1002/wcms.31

Bellissent-Funel, M. C., Hassanali, A., Havenith, M., Henchman, R., Pohl, P., Sterpone, F., et al. (2016). Water determines the structure and dynamics of proteins. *Chem. Rev.* 116, 7673–7697. doi: 10.1021/acs.chemrev.5b00664

Bernardi, R. C., Melo, M. C., and Schulten, K. (2015). Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta. Gen. Sub.* 1850, 872–877. doi: 10.1016/j.bbagen.2014.10.019

Bonomi, M., Branduardi, D., Bussi, G., Camilloni, C., Provasi, D., Raiteri, P., et al. (2009). PLUMED: a portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* 180, 1961–1972. doi: 10.1016/j.cpc.2009.05.011

Casasnovas, R., Limongelli, V., Tiwary, P., Carloni, P., and Parrinello, M. (2017). Unbinding kinetics of a p38 MAP kinase type II inhibitor from metadynamics simulations. *J. Am. Chem. Soc.* 139, 4780–4788. doi: 10.1021/jacs.6b12950

Case, D., Babin, V., Berryman, J., Betz, R., Cai, Q., Cerutti, D., et al. (2014). *AMBER 14.* San Francisco, CA: University of California.

Cavalli, A., Spitaleri, A., Saladino, G., and Gervasio, F. L. (2015). Investigating drug-target association and dissociation mechanisms using metadynamics-based algorithms. *Acc. Chem. Res.* 48, 277–285. doi: 10.1021/ar500356n

Chang, H. N., Liu, B. Y., Qi, Y. K., Zhou, Y., Chen, Y. P., Pan, K. M., et al. (2015). Blocking of the PD-1/PD-L1 interaction by a D-peptide antagonist for cancer immunotherapy. *Angew. Chem. Int. Edit.* 54, 11760–11764. doi: 10.1002/anie.201506225

Chen, F., Sun, H., Wang, J., Zhu, F., Liu, H., Wang, Z., et al. (2018). Assessing the performance of MM/PBSA and MM/GBSA methods. 8. Predicting binding free energies and poses of protein-RNA complexes. *RNA* 24, 1183–1194. doi: 10.1261/rna.065896.118

Duan, J., Dixon, S. L., Lowrie, J. F., and Sherman, W. (2010). Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.* 29, 157–170. doi: 10.1016/j.jmgm.2010.05.008

Gatalica, Z., Snyder, C., Maney, T., Ghazalpour, A., Holterman, D. A., Xiao, N., et al. (2014). Programmed cell death 1 (PD-1) and its ligand (PD-L1) in common cancers and their correlation with molecular cancer type. *Cancer Epidem. Biomar.* 23, 2965–2970. doi: 10.1158/1055-9965.EPI-14-0654

Gauto, D. F., Petruk, A. A., Modenutti, C. P., Blanco, J. I., Di Lella, S., and Marti, M. A. (2013). Solvent structure improves docking prediction in lectin-carbohydrate complexes. *Glycobiology* 23, 241–258. doi: 10.1093/glycob/cws147

Genheden, S., and Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* 10, 449–461. doi: 10.1517/17460441.2015.1032936

Guzik, K., Zak, K. M., Grudnik, P., Magiera, K., Musielak, B., Törner, R., et al. (2017). Small-molecule inhibitors of the Programmed Cell Death-1/Programmed Death-ligand 1 (PD-1/PD-L1) interaction via transiently-induced protein states and dimerization of PD-L1. *J. Med. Chem.* 60, 5857–5867. doi: 10.1021/acs.jmedchem.7b00293

Hirons, L., Holliday, J. D., Jelfs, S. P., Willett, P., and Gedeck, P. (2005). Use of the R-group descriptor for alignment-free QSAR. *QSAR Comb. Sci.* 24, 611–619. doi: 10.1002/qsar.200510102

Holliday, J., Jelfs, S., Willett, P., and Gedeck, P. (2003). Calculation of intersubstituent similarity using R-group descriptors. *J. Chem. Inf. Comput. Sci.* 43, 406–411. doi: 10.1021/ci025589v

Kalé, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., et al. (1999). NAMD2: greater scalability for parallel molecular dynamics. *J. Comput. Phys.* 151, 283–312. doi: 10.1006/jcph.1999.6201

Kolarevic, A., Ilic, B. S., Anastassova, N., Mavrova, A. T., Yancheva, D., Kocic, G., et al. (2018). Benzimidazoles as novel deoxyribonuclease I inhibitors. *J. Cell. Biochem.* 119, 8937–8948. doi: 10.1002/jcb.27147

Lee, H. T., Lee, J. Y., Lim, H., Lee, S. H., Moon, Y. J., Pyo, H. J., et al. (2017). Molecular mechanism of PD-1/PD-L1 blockade via anti-PD-L1 antibodies atezolizumab and durvalumab. *Sci. Rep.* 7:5532. doi: 10.1038/s41598-017-06002-8

Lee, J. Y., Lee, H. T., Shin, W., Chae, J., Choi, J., Kim, S. H., et al. (2016). Structural basis of checkpoint blockade by monoclonal antibodies in cancer immunotherapy. *Nat. Commun.* 7:13354. doi: 10.1038/ncomms13354

Liu, A., Dong, L., Wei, X. L., Yang, X. H., Xiao, J. H., and Liu, Z. Q. (2016). Development of amino- and dimethylcarbamate-substituted resorcinol as programmed cell death-1 (PD-1) inhibitor. *Eur. J. Pharm. Sci.* 88, 50–58. doi: 10.1016/j.ejps.2016.03.023

Liu, K., Tan, S., Chai, Y., Chen, D., Song, H., Zhang, C. W., et al. (2016). Structural basis of anti-PD-L1 monoclonal antibody avelumab for tumor therapy. *Cell Res.* 27, 151–153. doi: 10.1038/cr.2016.102

Magiera-Mularz, K., Skalniak, L., Zak, K. M., Musielak, B., Rudzinska-Szostak, E., Berlicki, Ł., et al. (2017). Bioactive macrocyclic inhibitors of the PD-1/PD-L1 immune checkpoint. *Angew. Chem. Int. Ed.* 129, 13920–13923. doi: 10.1002/ange.201707707

Masetti, M., Cavalli, A., Recanatini, M., and Gervasio, F. L. (2009). Exploring complex protein–ligand recognition mechanisms with coarse metadynamics. *J. Phys. Chem. B* 113, 4807–4816. doi: 10.1021/jp803936q

Pascolutti, R., Sun, X., Kao, J., Maute, R. L., Ring, A. M., Bowman, G. R., et al. (2016). Structure and dynamics of PD-L1 and an ultra-high-affinity PD-1 receptor mutant. *Structure* 24, 1719–1728. doi: 10.1016/j.str.2016.06.026

Patel, S. P., and Kurzrock, R. (2015). PD-L1 expression as a predictive biomarker in cancer immunotherapy. *Mol. Cancer Ther.* 14, 847–856. doi: 10.1158/1535-7163.MCT-14-0983

Perry, E., Mills, J. J., Zhao, B., Wang, F., Sun, Q., Christov, P. P., et al. (2019). Fragment-based screening of programmed death ligand 1 (PD-L1). *Bioorg. Med. Chem. Lett.* 29, 786–790. doi: 10.1016/j.bmcl.2019.01.028

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Qin, M., Cao, Q., Zheng, S., Tian, Y., Zhang, H., Xie, J., et al. (2019). Discovery of [1,2,4]Triazolo[4,3-a]pyridines as potent inhibitors targeting the programmed cell death-1/programmed cell death-ligand 1 interaction. *J. Med. Chem.* 62, 4703–4715. doi: 10.1021/acs.jmedchem.9b00312

Romero, J. M., Trujillo, M., Estrin, D. A., Rabinovich, G. A., and Di Lella, S. (2016). Impact of human galectin-1 binding to saccharide ligands on dimer dissociation kinetics and structure. *Glycobiology* 26, 1317–1327. doi: 10.1093/glycob/cww052

Sastry, M., Lowrie, J. F., Dixon, S. L., and Sherman, W. (2010). Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* 50, 771–784. doi: 10.1021/ci100062n

Sharma, P., and Allison, J. P. (2015a). The future of immune checkpoint therapy. *Science* 348, 56–61. doi: 10.1126/science.aaa8172

Sharma, P., and Allison, J. P. (2015b). Immune checkpoint targeting in cancer therapy: toward combination strategies with curative potential. *Cell* 161, 205–214. doi: 10.1016/j.cell.2015.03.030

Sharpe, A. H., Butte, M. J., and Oyama, S. (2011). *Modulators of Immunoinhibitory Receptor pd-1, and Methods of Use Thereof.* U.S. Patent Application.

Skalniak, L., Zak, K. M., Guzik, K., Magiera, K., Musielak, B., Pachota, M., et al. (2017). Small-molecule inhibitors of PD-1/PD-L1 immune checkpoint alleviate the PD-L1-induced exhaustion of T-cells. *Oncotarget* 8, 72167–72181. doi: 10.18632/oncotarget.20050

Sun, H., Duan, L., Chen, F., Liu, H., Wang, Z., Pan, P., et al. (2018). Assessing the performance of MM/PBSA and MM/GBSA methods. 7. Entropy effects on the performance of end-point binding free energy calculation approaches. *Phys. Chem. Chem. Phys.* 20, 14450-14460. doi: 10.1039/c7cp07623a

Sun, H., Li, Y., Shen, M., Li, D., Kang, Y., and Hou, T. (2017). Characterizing drug-target residence time with metadynamics: how to achieve dissociation rate efficiently without losing accuracy against time-consuming approaches. *J. Chem. Inf. Model.* 57, 1895–1906. doi: 10.1021/acs.jcim.7b00075

Tan, S., Chen, D., Liu, K., He, M., Song, H., Shi, Y., et al. (2016). Crystal clear: visualizing the intervention mechanism of the PD-1/PD-L1 interaction by two cancer therapeutic monoclonal antibodies. *Protein Cell* 7, 866–877. doi: 10.1007/s13238-016-0337-7

Tan, S., Liu, K., Chai, Y., Zhang, C. W., Gao, S., Gao, G. F., et al. (2017). Distinct PD-L1 binding characteristics of therapeutic monoclonal antibody durvalumab. *Protein Cell* 9, 135–139. doi: 10.1007/s13238-017-0412-8

Ts Mavrova, A., Dimov, S., Yancheva, D., Kolarevic, A., Ilic, B. S., Kocic, G., et al. (2018). Synthesis and DNase I inhibitory properties of some 5,6,7,8-tetrahydrobenzo[4,5]thieno[2,3-d]pyrimidines. *Bioorg. Chem.* 80, 693–705. doi: 10.1016/j.bioorg.2018.07.009

Xue, W., Jin, X., Ning, L., Wang, M., Liu, H., and Yao, X. (2013). Exploring the molecular mechanism of cross-resistance to HIV-1 integrase strand transfer inhibitors by molecular dynamics simulation and residue interaction network analysis. *J. Chem. Inf. Model.* 53, 210–222. doi: 10.1021/ci300541c

Zak, K. M., Grudnik, P., Guzik, K., Zieba, B. J., Musielak, B., Dömling, A., et al. (2016). Structural basis for small molecule targeting of the programmed death ligand 1 (PD-L1). *Oncotarget* 7, 30323–30335. doi: 10.18632/oncotarget.8730

Zak, K. M., Kitel, R., Przetocka, S., Golik, P., Guzik, K., Musielak, B., et al. (2015). Structure of the complex of human programmed death 1, PD-1, and its ligand PD-L1. *Structure* 23, 2341–2348. doi: 10.1016/j.str.2015.09.010

Zarganes-Tzitzikas, T., Konstantinidou, M., Gao, Y., Krzemien, D., Zak, K., Dubin, G., et al. (2016). Inhibitors of programmed cell death 1 (PD-1): a patent review (2010-2015). *Expert Opin. Ther. Pat.* 26, 973–977. doi: 10.1080/13543776.2016.1206527

Zhang, F., Qi, X., Wang, X., Wei, D., Wu, J., Feng, L., et al. (2017a). Structural basis of the therapeutic anti-PD-L1 antibody atezolizumab. *Oncotarget* 8, 90215–90224. doi: 10.18632/oncotarget.21652

Zhang, F., Wei, H., Wang, X., Bai, Y., Wang, P., Wu, J., et al. (2017b). Structural basis of a novel PD-L1 nanobody for immune checkpoint blockade. *Cancer Discov.* 3:17004. doi: 10.1038/celldisc.2017.4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Shi, An, Bai, Bing, Zhou, Liu and Yao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

REVIEW
published: 20 November 2019
doi: 10.3389/fchem.2019.00782

Comparison Study of Computational Prediction Tools for Drug-Target Binding Affinities

Maha Thafar[1,2], *Arwa Bin Raies*[1], *Somayah Albaradei*[1,3], *Magbubah Essack*[1]* and *Vladimir B. Bajic*[1]*

[1] Computer, Electrical and Mathematical Science and Engineering (CEMSE) Division, Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, [2] College of Computers and Information Technology, Taif University, Taif, Saudi Arabia, [3] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

OPEN ACCESS

Edited by:
Jose L. Medina-Franco, National Autonomous University of Mexico, Mexico

Reviewed by:
Julio Caballero, University of Talca, Chile
Humbert Gonzalez-Diaz, University of the Basque Country, Spain
Simone Brogi, University of Pisa, Italy
Giulia Chemi, University of Siena, Siena, Italy, in collaboration with reviewer SB

***Correspondence:**
Vladimir B. Bajic
vladimir.bajic@kaust.edu.sa
Magbubah Essack
magbubah.essack@kaust.edu.sa

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: 09 August 2019
Accepted: 30 October 2019
Published: 20 November 2019

Citation:
Thafar M, Raies AB, Albaradei S, Essack M and Bajic VB (2019) Comparison Study of Computational Prediction Tools for Drug-Target Binding Affinities. Front. Chem. 7:782. doi: 10.3389/fchem.2019.00782

The drug development is generally arduous, costly, and success rates are low. Thus, the identification of drug-target interactions (DTIs) has become a crucial step in early stages of drug discovery. Consequently, developing computational approaches capable of identifying potential DTIs with minimum error rate are increasingly being pursued. These computational approaches aim to narrow down the search space for novel DTIs and shed light on drug functioning context. Most methods developed to date use binary classification to predict if the interaction between a drug and its target exists or not. However, it is more informative but also more challenging to predict the strength of the binding between a drug and its target. If that strength is not sufficiently strong, such DTI may not be useful. Therefore, the methods developed to predict drug-target binding affinities (DTBA) are of great value. In this study, we provide a comprehensive overview of the existing methods that predict DTBA. We focus on the methods developed using artificial intelligence (AI), machine learning (ML), and deep learning (DL) approaches, as well as related benchmark datasets and databases. Furthermore, guidance and recommendations are provided that cover the gaps and directions of the upcoming work in this research area. To the best of our knowledge, this is the first comprehensive comparison analysis of tools focused on DTBA with reference to AI/ML/DL.

Keywords: drug repurposing, drug-target interaction, drug-target binding affinity, artificial intelligence, machine learning, deep learning, information integration, bioinformatics

INTRODUCTION

Experimental confirmation of new drug-target interactions (DTIs) is not an easy task, as *in vitro* experiments are laborious and time-consuming. Even if a confirmed DTI has been used for developing a new drug (in this review compounds that are not approved drugs are also referred to as drugs), the approval for human use of such new drugs can take many years and estimated cost may run over a billion US dollars (Dimasi et al., 2003). Moreover, although huge investments are required for the development of novel drugs, they are often met with failure. In fact, of the 108 new and repurposed drugs reported as Phase II failures between 2008 and 2010, 51% was due to insufficient efficacy as per a Thomson Reuters Life Science Consulting report (Arrowsmith, 2011). This observation highlighted the need for: (1) new, more appropriate drug targets, and (2)

in silico methods that can improve the efficiency of the drug discovery and screen a large number of drugs in the very initial phase of drug discovery process, thus guiding toward those drugs that may exhibit better efficacy. In this regard, methods that predict DTIs and specifically, drug-target binding affinities (DTBA) are of great interest.

Over the last three decades, several methods that predict DTIs have been developed ranging from ligand/receptor-based methods (Cheng et al., 2007; Wang et al., 2013) to gene ontology-based (Mutowo et al., 2016), text-mining-based methods (Zhu et al., 2005), and reverse virtual screening techniques (reverse-docking) (Lee et al., 2016; Vallone et al., 2018; Wang et al., 2019). Development of such methods is still ongoing as each method suffers from different types of limitations. For example, docking simulation is often used in the receptor-based methods; also, docking simulation requires the 3D structures of the target proteins that are not always readily available. Furthermore, this is an expensive process. On the other hand, the ligand-based approaches suffer from low performance when the number of known ligands of target proteins is small, as this approach predicts DTIs based on the similarity between candidate ligands and the known ligands of the target proteins. The limitations associated with gene ontology-based and text-mining-based approaches are the same, their major limitation appears to be what is reported in the text. This also becomes more complicated due to the frequent use of redundant names for drugs and target proteins. Moreover, with the text-mining approach being limited to the current knowledge (i.e., published material), making discovery of new knowledge is not easy.

Other methods such as deep learning (DL), machine learning (ML), and artificial intelligence (AI) in general, avoid these limitations by using models that learn the features of known drugs and their targets to predict new DTIs. Understanding that ML methods are just a subset of AI methods, does not always makes it clear what would be strictly an ML method and what an AI method. This particularly becomes apparent when graph, network, and search analyses methods are combined with conventional (shallow) ML approaches. The situation for DL is clearer, as these methods are a subset of ML approaches based on transformation of the original input data representation across multiple information processing layers, thus distinguishing them from the shallow ML approaches. More recent approaches introduced AI, network analysis, and graph mining (Emig et al., 2013; Ba-Alawi et al., 2016; Luo et al., 2017; Olayan et al., 2018), and ML and DL techniques (Liu Y. et al., 2016; Rayhan et al., 2017; Zong et al., 2017; Tsubaki et al., 2019) to develop prediction models for DTI problem. AI/ML-based methods (we will frequently refer to them in this study as ML methods) are generally feature-based or similarity-based (see DTBA ML-based methods section). Feature-based AI/ML methods can be integrated with other approaches constructing "Ensemble system" as presented in Ezzat et al. (2016), Jiang et al. (2017), and Rayhan et al. (2019). Thus, several comprehensive recent reviews summarized the different studies that predict DTIs using various techniques covering structure-based, similarity-based, network-based, and AI/ML-based methods as presented in Liu Y. et al. (2016), Ezzat et al. (2017, 2018, 2019), Rayhan et al. (2017), Trosset and Cavé (2019), and Wan et al. (2019). Other reviews focused on one aspect which are similarity-based methods (Ding et al., 2014; Kurgan and Wang, 2018) or feature-based methods (Gupta, 2017). Most of the approaches mentioned above address DTI prediction as a simple binary on-off relationship. That is, they simply predict whether the drug and target could interact or not. This approach suffers from two major limitations including: (1) the inability to differentiate between true negative interactions and instances where the lack of information or missing values impede predicting an interaction, and (2) it does not reflect how tightly the drug binds to the target which reflects the potential efficacy of the drug. To overcome these limitations, approaches that focus on DTBA predictions have been developed. We compile this study with the focus on DTBA, which has not been addressed well in the past, but is more critical for estimating usefulness of DTI in early stages of drug development.

DRUG-TARGET BINDING AFFINITY (DTBA)

DTBA indicates the strength of the interaction or binding between a drug and its target (Ma W. et al., 2018). The advantage of formulating drug-target prediction as a binding affinity regression task, is that it can be transformed from regression to either binary classification by setting specific thresholds or to ranking problem (He et al., 2017). This enables different generalization options.

Most *in silico* DTBA prediction methods developed to date use 3D structural information (see **Figure 1**), which was demonstrated to successfully contribute to the drug design (Leach et al., 2006). Some of these methods provide free analysis software as reported by Agrawal et al. (2018). The 3D structure information of proteins is used in the molecular docking analysis and followed by applying search algorithms or scoring functions to assist with the binding affinity predictions (Scarpino et al., 2018; Sledz and Caflisch, 2018). This whole process is used in the structured-based virtual screening (Li and Shah, 2017).

In DTBA predictions, the concept of scoring function (SF) is frequently used. SF reflects the strength of binding affinity between ligand and protein interaction (Abel et al., 2018). When SFs have a prearranged functional form that mimics the relationship between structural features and binding affinity, it is called classical SF. Classical SFs are categorized as Empirical SFs (Guedes et al., 2018), Force field SFs (Huang and Zou, 2006),

Abbreviations: AI, artificial intelligence; ML, machine learning; DL, deep learning; Sim, similarity; aaseq, amino-acid sequence; SPS, structural property sequence; PSC, protein sequence composition; PDM, protein domain and motif; ECFP, extended-connectivity fingerprint; LMCS, ligand maximum common substructure; KronRLS, Kronecker regularized least square; CNN, convolutional neural network; GCNN, graph convolution neural network; FNN, feedforward neural network; ANN, artificial neural network; RNN, recurrent neural network; RBNN, radial basis function neural network; MNN, modular neural network; MLP, multilayer perceptron; RNN, recurrent neural network; FC, fully connected; ReLU, rectified linear unit; CV, cross validation; LDO, leave one drug out; LTO, leave one target out; MSE, mean square error; RMSE, root square of mean square error; CI, concordance index; PCC, Pearson correlation coefficient; NR, nuclear receptors; GPCR, G protein-coupled receptors; IC, ion channels; E, enzymes; KIBA, kinase inhibitor bioactivity.

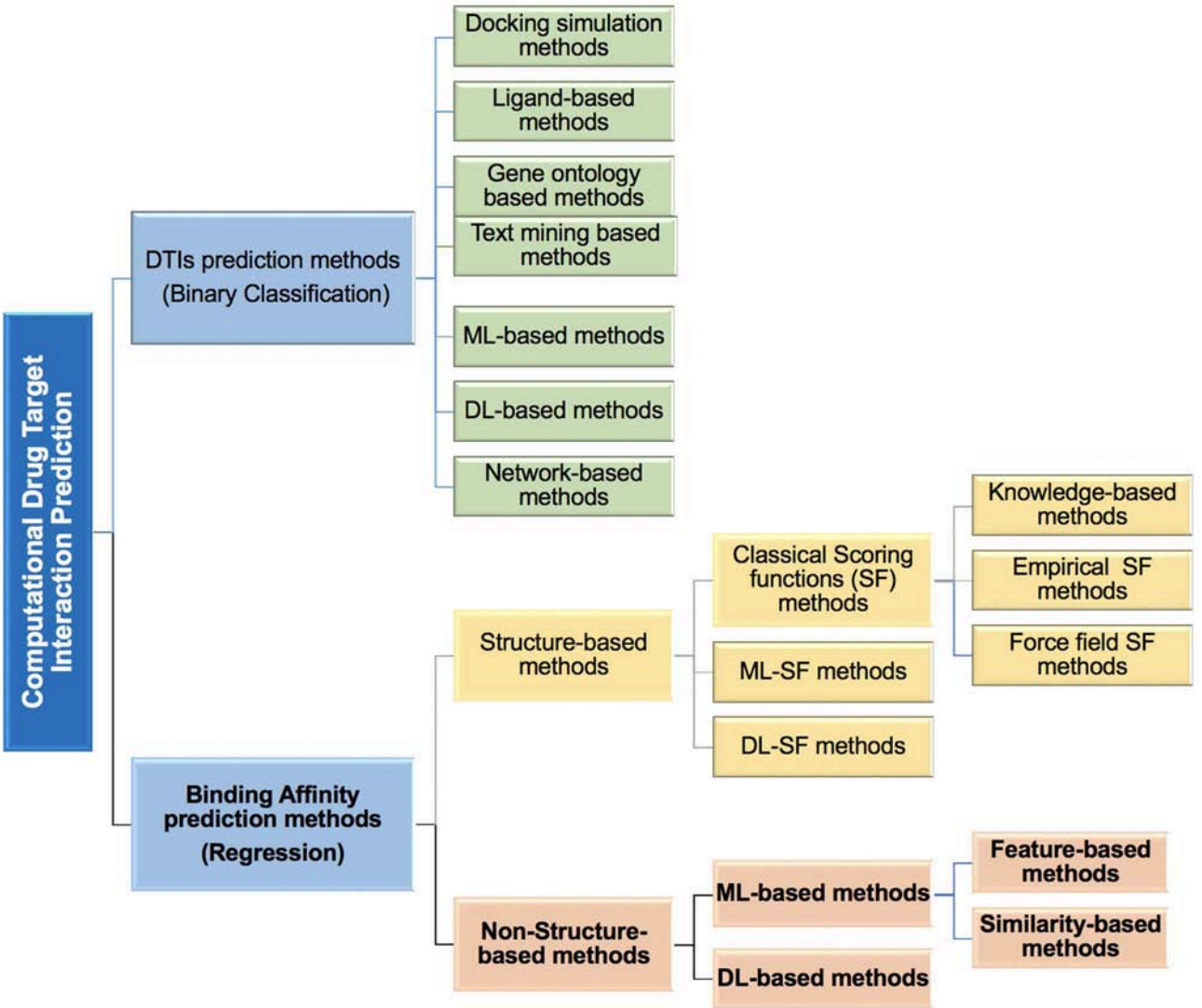


FIGURE 1 | An overview of the different types of computational methods developed to predict drug-target interactions (DTIs) and drug-target binding affinity (DTBA) categories.

and Knowledge-based SFs (Huang and Zou, 2006; Liu et al., 2013). SFs have been used to predict protein-ligand interaction in molecular docking such as with the Binding Estimation After Refinement (BEAR) SF (Degliesposti et al., 2011) which is a post docking tool that uses molecular dynamics to accurately predict protein binding free energies using SF. Several of these classical SFs are summarized in a recent review (Li J. et al., 2019). A specific form of the SF called target-specific SF, is based on energy calculations of interacting compound (i.e., free energy calculations; Ganotra and Wade, 2018; Sun et al., 2018). Other SFs were also developed that do not follow a predetermined functional form. These SFs use ML techniques to infer functional form from training data (Deng et al., 2004; Vert and Jacob, 2008; Kundu et al., 2018). Thus, the ML-based SFs methods are data-driven models that capture the non-linearity relationship in data making the SF more general and more accurate. DL is an emerging research area in different cheminformatic fields including drug design (Jain, 2017; Andricopulo and Ferreira, 2019). SFs that use DL in structure-based methods focused on binding affinity prediction have been developed (Ashtawy and Mahapatra, 2018; Jiménez et al., 2018; Antunes et al., 2019). As all DL models, these DL-based SFs methods learn the features to predict binding affinity without requirement for feature engineering as may be the case in the ML methods. Several reviews have been made covering virtual screening structure-based binding affinity prediction methods including docking techniques, before applying SFs (Kontoyianni, 2017; Li and Shah, 2017), classical SFs (Guedes et al., 2018), or ML-derived SFs (Ain et al., 2015; Heck et al., 2017; Colwell, 2018; Kundu et al., 2018). The main limitations of the structure-based methods are the

requirement for the 3D structure data (including compound and protein) that are scarce. This is compounded by the problem of low-quality structure predicted from docking, which cannot be tested and scaled to large-scale data applications (Karimi et al., 2019). Several publications have discussed the major limitations of structured-based virtual screening (Sotriffer and Matter, 2011; Hutter, 2018).

Non-structure-based methods, overcome most of these limitations since there is no need for the docking process or 3D structural data. Despite the enormous amount of effort and research devoted to binding affinity prediction, there are only a few publications that address the DTBA problem as a non-structure-based approach. This remains a critical and challenging task that requires the development of significantly improved algorithms.

Here, we review methods developed for prediction of DTIs based on binding affinities. Specifically, we focus on the novel methods that utilize non-structure-based binding affinity prediction (shown in bold font in **Figure 1**), which does not require or use 3D structural data. The study provides a comparative analysis of the current DTBA prediction methods. It covers: (a) definitions and calculations associated with binding affinity, (b) the benchmark datasets that are used in DTBA regression problem, (c) computational methods used, (d) evaluation and performance comparison of DTBA prediction methods, and (e) recommendations of areas for improvement and directions in binding affinity prediction research.

MEASURING BINDING AFFINITY

Each ligand/protein has a unique binding affinity constant for specific receptor system which can be used to identify distinct receptors (Weiland and Molinoff, 1981; Bulusu et al., 2016). The equilibrium reaction below describes how a protein (P) binds to its ligand (L) to create the protein-ligand complex (PL) (Du et al., 2016):

$$P+L \overset{K_a}{\Leftrightarrow} PL \tag{1}$$

K_a is the equilibrium association constant (also called binding affinity constant). A high value of K_a indicates a strong binding capacity between the drug/ligand and the receptor/protein (Weiland and Molinoff, 1981; Bulusu et al., 2016). The inverse of the above reaction is when the protein-ligand complex dissociates into its components of a protein and a ligand as explained in the equilibrium reaction below (Du et al., 2016):

$$PL \overset{K_d}{\Leftrightarrow} P+L \tag{2}$$

K_d is the equilibrium dissociation constant, and it is used more often than K_a. Small values of K_d indicate higher affinity (Ma W. et al., 2018). K_d is the inverse of the K_a as illustrated in the equation below (Du et al., 2016):

$$K_d = \frac{1}{K_a} \tag{3}$$

Binding Curve

Figure 2 shows a hypothetical example of a binding curve for two ligands: Ligand 1 and Ligand 2. The x-axis represents the concentration of the ligand, and the y-axis represents the percentage of available binding sites (Θ) in a protein that is occupied by the ligand. The values of Θ range from 0 to 1 (corresponding to the range from 0 to 100% in **Figure 2**). For example, if Θ is 0.5, this means that 50% of the available binding sites are occupied by the ligand. The binding curves help in determining graphically which ligand binds more strongly to the protein at a specific concentration of the ligand (Stefan and Le Novère, 2013). For example, in **Figure 2**, if the concentration of the ligands is 3 μM, Ligand 1 binds to 75% of the binding sites of the protein, while Ligand 2 binds to only 50% of the binding sites. Therefore, Ligand 1 binds more strongly to the protein than Ligand 2. **Figure 2** depicts an example of cooperative binding (if the concentration of the ligand increases, the number of binding sites the ligand occupies increases non-linearly). Cooperative binding is positive if binding of the ligand increases the affinity of the protein and increases the chance of another ligand binding to the protein; otherwise, the cooperative binding is negative (i.e., binding of the ligand to the protein decreases the affinity of the protein and reduces the chance of another ligand binding to the protein; Stefan and Le Novère, 2013).

The equation below shows the relationship between Θ for a protein to which the ligand binds, and K_d of the equilibrium reaction at a given concentration of the ligand $[L]$ (Salahudeen and Nishtala, 2017):

$$\theta = \frac{[L]}{K_d+[L]} \tag{4}$$

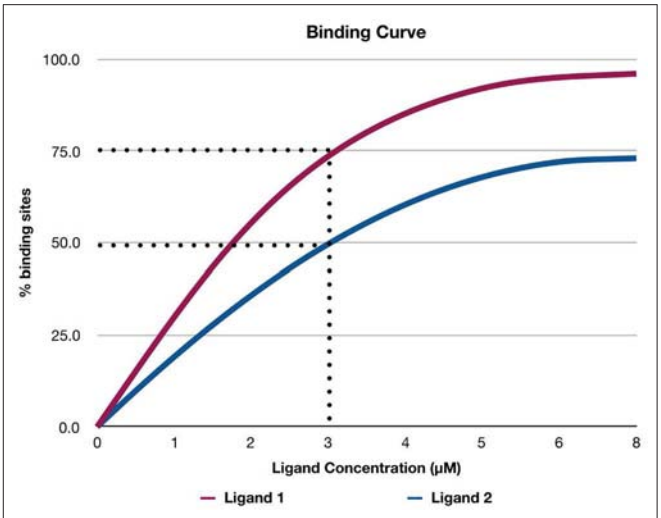


FIGURE 2 | A hypothetical example of a binding curve for ligand 1 and ligand 2. The x-axis shows the concentration of the ligand, and the y-axis shows the percentage of available binding sites (Θ) in a protein that is occupied by the ligand.

K_i and IC_{50} Constants

The inhibitor constant (K_i) is an indicator of the potency of an inhibitor (Bachmann and Lewis, 2005). Inhibitors are compounds (e.g., drugs) that can reduce the activity of enzymes. Enzymes that exhibit overactivity are potential targets for drugs to treat specific diseases, as well as inhibitors of a cascade of events in a pathway. Several drugs act by inhibiting these specific enzymes (Chou and Talalay, 1984; Tang et al., 2017). IC_{50} is the concentration required to produce half-maximum inhibition (Bachmann and Lewis, 2005). K_i is calculated using IC_{50} values, which are the concentration required to produce 50% inhibition (Burlingham and Widlanski, 2003). **Figure 3** provides a hypothetical example of IC_{50} values, with the concentration of the inhibitor represented on the x-axis, and the percentage of enzyme activity represented on the y-axis. The hypothetical example (in **Figure 3**) shows 50% of enzyme activity can be inhibited when the concentration of the inhibitor is 2 μM.

IC_{50} is not an indicator of affinity, but rather indicates the functional strength of the inhibitor. On the other hand, K_i constant reflects the binding affinity of the inhibitor. Lower values of K_i indicate higher affinity. The relationship between IC_{50} and K_i is explained by the equation below (Hulme and Trevethick, 2010):

$$K_i = \frac{IC_{50}}{1 + \frac{[S]}{K_m}} \tag{5}$$

where K_m is the substrate concentration (in the absence of inhibitor) at which the velocity of the reaction is half-maximal, and $[S]$ is the concentration of substrate. More details about K_m can be found in Hulme and Trevethick (2010).

BENCHMARK DATASETS AND SOURCES

Benchmark datasets are used to train models and evaluate their performance on the standardized data. Using these datasets also allow the performance of the newly developed method to be compared to the state-of-the-art methods to establish the best performance. Only a few benchmark datasets have been used to develop *in silico* DTBA prediction methods. When predicting DTIs, the Yamanishi datasets (Yamanishi et al., 2008) are the most popular benchmark datasets. There are four Yamanishi datasets based on family of target proteins, including: (1) nuclear receptors (NR), (2) G protein-coupled receptors (GPCR), (3) ion channels (IC), and (4) enzymes (E). Each dataset contains binary labels to indicate the interacting or non-interacting drug-target pairs (Yamanishi et al., 2008). However, these datasets cannot be used for DTI regression-based models, because the datasets do not indicate the actual binding affinities between known interacting drug-target pairs. That is, actual binding affinity scores are needed to train the models to predict the continuous values that indicate the binding strength between drugs and their targets. Three large-scale benchmark datasets that we name Davis dataset, Metz dataset, and Kinase Inhibitor BioActivity (KIBA) dataset, which provide these binding affinities for interaction strength were used to evaluate DTBA prediction in Davis et al. (2011), Metz et al. (2011), and Tang et al. (2014), respectively. All three datasets are large scale biochemical selectivity assays of the kinase inhibitors. The kinase protein family is used for the reason that this protein family has increased biological activity and is involved in mediating critical pathway signals in cancer cells (Tatar and Taskin Tok, 2019).

In Davis dataset, the K_d value is provided as a measure of binding affinity. The Metz dataset provides the K_i as a measure of binding affinity. When the value of K_d or K_i is small, this indicates strong binding affinity between a drug and its target. KIBA dataset integrates different bioactivities and combines K_d, K_i, and IC_{50} measurements. KIBA score represents a continuous value of the binding affinity that was calculated utilizing K_d, K_i, and IC_{50} scores. The higher KIBA score indicates a lower binding affinity between a drug and its target.

Recently, Feng (2019) also used ToxCast (Judson, 2012) as a benchmark dataset for binding affinity. This dataset is much larger than the other three benchmark datasets. It contains data about different proteins that can help in evaluating the model robustness and scalability. ToxCast contains toxicology data obtained from *in vitro* high-throughput screening of drugs (i.e., chemicals). Several companies have done ToxCast curation with 61 different measurements of binding affinity scores. Other details of this dataset and the method are explained later in section Computational Prediction of Drug-Target Binding Affinities. **Table 1** summarizes the statistics for these four benchmark datasets.

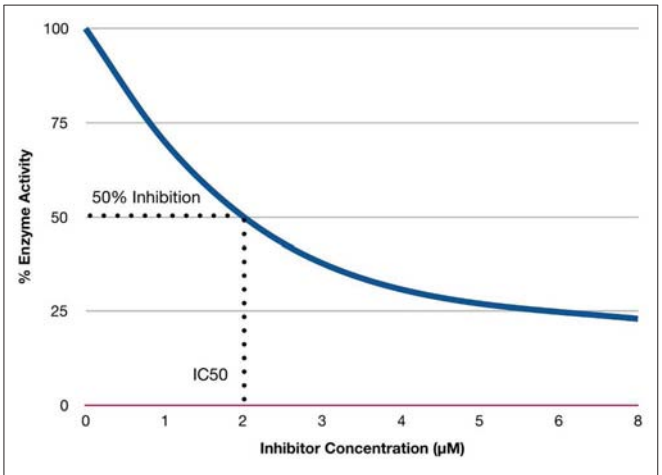


FIGURE 3 | Relationship between concentration of inhibitors and enzymes activity.

TABLE 1 | Binding affinity benchmark datasets statistics.

Datasets	No. of drugs	No. of proteins	Known DTIs
Davis	68	442	30,056
Metz	1,421	156	35,259
Kiba	2,116	229	118,254
ToxCast	7,675	335	530,605

Other benchmark binding affinity datasets provided 3D structure information used to evaluate and validate structure-based methods via scoring functions and docking techniques. These benchmark datasets provide all the binding affinity information for the interactions. We listed these datasets without mentioning any further details since their use is beyond the scope of this study. Most of these datasets have more than one version since they are updated each year by adding more experimental, validated data. These datasets/data sources are: PDBbind (Wang et al., 2004, 2005), BindingDB (Chen et al., 2001; Liu et al., 2007; Gilson et al., 2016), BindingMOAD—(the Mother Of All Databases; Hu et al., 2005; Benson et al., 2007; Ahmed et al., 2015; Smith et al., 2019), CSAR (Smith et al., 2011; Dunbar et al., 2013), AffinDB (Block et al., 2006), Ligand Protein DataBase (LPDB) (Roche et al., 2001), and Protein-Ligand Database (PLD) (Puvanendrampillai and Mitchell, 2003). These datasets are integrated with protein 3D structure information provided in Protein Data Bank (PDB) (Berman et al., 2000; Westbrook et al., 2003) adding more information. All these resources are publicly available, and some of them have associated web-tools aiming to facilitate accessing and searching information.

COMPUTATIONAL PREDICTION OF DRUG-TARGET BINDING AFFINITIES

There are few cheminformatics methods developed to predict continuous DTBA that do not use the 3D structure data. These methods are data-driven and use AI/ML/DL techniques for regression task rather than classification. To our knowledge, there are only six state-of-art methods developed for DTBA prediction. These we describe in what follows.

Artificial Intelligence and Machine Learning-Based Methods

AI/ML and statistical analysis approaches have been applied across different stages of the drug development and design pipelines (Lima et al., 2016) including target discovery (Ferrero et al., 2017), drug discovery (Hutter, 2009; Raschka et al., 2018; Vamathevan et al., 2019), multi-target drug combination prediction (Tang et al., 2014; Vakil and Trappe, 2019), and drug safety assessment (Raies and Bajic, 2016, 2018; Lu et al., 2018). AI/ML approaches are generally either feature-based or similarity-based. The feature-based approaches use known DTIs chemical descriptors for drugs and the descriptors for the targets to generate feature vectors. On the other hand, similarity-based AI/ML approaches use the "guilt by association" rule. Using this rule is based on the assumptions that similar drugs tend to interact with similar targets and similar targets are targeted by similar drugs. Such AI/ML approaches that predict binding affinity of DTIs were used to develop state-of-the-art DTBA prediction methods, KronRLS (Pahikkala et al., 2015) and SimBoost (He et al., 2017).

KronRLS

Regularized least-square (RLS) is an efficient model used in different types of applications (Pahikkala et al., 2012a,b). Van Laarhoven et al. (2011) used RLS for the binary prediction of DTIs and achieved outstanding performance. Later, the RLS model was amended to develop a method that is suitable for DTBA prediction named, Kronecker-Regularized Least Squares (KronRLS) (Pahikkala et al., 2015). This method is a similarity-based method that used different types of drug-drug similarity and protein-protein similarity score matrices as features. The problem is formulated as regression or rank prediction problem as follows: a set D of drugs $\{d_1, d_2,..., d_i\}$ and a set T of protein targets $\{t_1, t_2,..., t_i\}$ are given with the training data $X = \{x_1, x_2,..., x_n\}$ that is a subset from all possible generated drug-target pairs $X \subset \{d_i \times t_j\}$. Each row of X (i.e., feature vector) is associated with the label y_i, $y_i \in Y_n$, where Y_n is the label vector that represents a binding affinity. To learn the prediction function f, a minimizer of the following objective function J is defined as:

$$J(f) = \sum_{i=1}^{m} (y_i - f(x_i))^2 + \lambda \| f \|_k^2 \quad (6)$$

Here $||f||_k$ is the norm of f, $\lambda > 0$ is regularization parameter defined by the user, and K is the kernel function (i.e., similarity) that is associated with the norm. The objective function to be minimized during optimization process is defined as:

$$f(x) = \sum_{i=1}^{m} a_i K(x, x_i) \quad (7)$$

The kernel function K in the equation above is the symmetric similarity matrix $n \times n$ for all possible drug-target pairs. This kernel function is the Kronecker product of two other similarity matrices: $K = K_d \otimes K_t$, where K_d is the drug chemical structure similarity matrix computed using the PubChem structure clustering tool, and K_t is the protein sequence similarity matrix computed using both original and normalized versions of the Smith-Waterman (SW) algorithm (Yamanishi et al., 2008; Ding et al., 2014). There are two scenarios of the training data. If the training set $X = \{d_i \times t_j\}$ contains all possible pairs, the parameter vector a in Equation (7) can be obtained by solving the system of linear equations:

$$(K + I)\, a = y \quad (8)$$

where I is the identity matrix. For the second scenario, if only a subset of $\{d_i \times t_j\}$ is used as the training data, such as $X \subset \{d_i \times t_j\}$, the vector y has missing values for binding affinity and for determining the parameter a, conjugate gradient with Kronecker algebraic optimization is needed to solve the system of linear Equation (8).

SimBoost

SimBoost (He et al., 2017) is a novel non-linear method that has been developed to predict DTBA as a regression task using gradient boosting regression trees. This method uses both similarity matrices and constructed features. The definition of the training data is similar to the KronRLS method. Thus, SimBoost requires a set of, (1) drugs (D), (2) targets (T), (3) drug-target

pairs (that are associated with user-defined features), and (4) binding affinity such that $y_i \in Y_n$ (where Y_n is the binding affinity vector). SimBoost is used to generate features for each drug, target, and drug-target pair. There are three types of features:

Type-1 features are object-based features for every single drug and target. This type of features reflects the statistics and similarity information such as score, histogram, a frequency for every single object (drug or target).

Type-2 features are similar to network-based features. Here, two networks are built, one network for drug-drug similarity, and the other network for target-target similarity. For the drug-drug similarity network, each drug is a graph node, and the nodes connected through edges. Edges are determined using the similarity score that is higher than the user-defined threshold. The construction of the second target-target similarity network is similar to the drug-drug network. For each network, we extract features. These features include statistics of node neighbors, page rank, betweenness, and eigencentrality (introduced in Newman, 2018).

Type-3 features are heterogeneous network-based features from the drug-target network, where drugs and targets are connected based on binding affinity continuous value. We extract other features from this network such as the latent vectors using matrix factorization (Liu Y. et al., 2016), and the normal ones, including betweenness, closeness, and eigencentrality.

A feature vector is constructed for each (drug, target) pair by concatenating type-1 and type-2 feature vector for each d_i and t_j and type-3 feature vector for each pair (d_i, t_j). After finishing feature engineering, the feature vector is feed to the gradient boosting regression trees. In this model, the predicted score $\hat{y}_i$ for each input data x_i that is represented by its feature vector, is computed using the following:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{9}$$

Here, B is the number of regression trees, $\{f_k\}$ is the set of trees, and F represents the space of all possible trees. The following is the regularized objective function L used to learn the f_k:

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{10}$$

Here, l is a differentiable loss function that evaluates the prediction error. The Ω function measures the model complexity to avoid overfitting. The model is trained additively, at each iteration t, F is searched to find a new tree f_t. This new tree f_t optimizes the following objective function:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) = \sum_{i=1}^{t} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \sum_{i=1}^{t} \Omega(f_i) \tag{11}$$

A gradient boosting algorithm iteratively adds trees that optimize the approximate objective at specific step for several user-defined iterations. SimBoost used similarity matrices are the same as KronRLS and are obtained using drug-drug similarity (generated by PubChem clustering based on the chemical structure) and target-target similarity (generated using the SW algorithm based on protein sequences).

Deep Learning-Based Methods

Recently and in this big data era, DL approaches have been successfully used to address diverse problems in bioinformatics/cheminformatics applications (Ekins, 2016; Kalkatawi et al., 2019; Li Y. et al., 2019) and more specifically in drug discovery as discussed in detail in Chen et al. (2018), Jing et al. (2018), and Ekins et al. (2019). DL algorithms developed to predict DTBA sometimes show superior performance when compared to conventional ML algorithms (Öztürk et al., 2018, 2019; Karimi et al., 2019). These DL-based algorithms developed to predict DTBA differ from each other in two main aspects. The first is concerning the representation of input data. For example, Simplified Molecular Input Line Entry System (SMILES), Ligand Maximum Common Substructure (LMCS) Extended Connectivity Fingerprint (ECFP), or a combination of these features can be used as drug features (see **Table 4**). The second is concerning the DL system architecture that is developed based on different neural network (NN) types (Krig, 2016) elaborated on below. The NN types differ in their structure that in some cases include the number of layers, hidden units, filter sizes, or the incorporated activation function. Each type of NN has its inherent unique strengths that make them more suitable for specific kinds of applications. The most popular NN types include the Feedforward Neural Network (FNN), Radial Basis Function Neural Network (RBNN), Multilayer Perceptron (MLP), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Modular Neural Network (MNN) (Schmidhuber, 2015; Liu et al., 2017). FNN and CNN have been used in algorithms discussed below to predict DTBA.

FNN, also known as a front propagated wave, is the simplest type of artificial NN (ANN) (Michelucci, 2018). In this type, the information only moves in one direction, from the input nodes to the output nodes, unlike more complex kinds of NN that have backpropagation. Nonetheless, it is not restricted to having a single layer, as it may have multiple hidden layers. Like all NN, FNN also incorporates an activation function. Activation function (Wu, 2009) is represented by a node which is added to the output layer or between two layers of any NN. Activation function node decides what output a neuron should produce, e.g., should it be activated or not. The form of the activation function is the non-linear transformation of the input signal to an output signal that serves as the input of a subsequent layer or the final output. Example of activation functions includes sigmoid, tanh, Rectified Linear Unit (ReLU), and variants of them.

On the other hand, CNN uses a variation of multilayer perceptron. Its architecture incorporates convolution layers which apply k filters on the input to systematically capture the presence of some discriminative features and create feature maps (Liu et al., 2017). Those filters are automatically learned based on the desired output, which maximizes the algorithms ability to identify true positive cases. This is achieved through a

loss layer (loss function) which penalizes predictions based on their deviations from the training set. When many convolutional layers are stacked, more abstracted features are automatically detected. Usually, a pooling layer follows a convolution layer to limit the dimension and keep only the essential elements. The common types of pooling are average pooling and max pooling. Average pooling finds the average value for each patch on the feature map. Max pooling finds the maximum value for each patch of the feature map. The pooling layer produces a down-sampled feature map which reduces the computational cost. After features extraction and features selection automatically performed by the convolutional layers and pooling layers, fully connected layers are usually used to perform the final prediction.

The general design used for the prediction of DTBA start with the representation of the input data for the drug and target, then different NN types with various structures are applied to learn features (i.e., embedding). Subsequently, the features of each drug-target pair are concatenated to create feature vectors for all drug-target pairs. The fully connected (FC) layers are fed with these feature vectors for the prediction task. **Figure 4** provides a step-by-step depiction of this general framework.

DeepDTA

DeepDTA, introduced in Öztürk et al. (2018), is the first DL approach developed to predict DTBA, and it does not incorporate 3D structural data for prediction, i.e., it is non-structure-based method. DeepDTA uses SMILES, the one-dimensional representation of the drug chemical structure (Weininger, 1988, 1990), as representation of the drug input data for drugs, while the amino acid sequences are used to represent the input data for proteins. Integer/label encoding was used to encode drug SMILES. For example, the [1 3 63 1 63 5] label encodes the "CN=C=O" SMILES. The protein sequences are similarly encoded. More details about data preprocessing and representation are explained in Öztürk et al. (2018). A CNN (Liu et al., 2017) that contains three 1D convolutional layers following by max-pooling function (called the first CNN block) was applied on the drug embedding to learn latent features for each drug. All three 1D convolution layers in each CNN block consists of 32, 64, and 96 filters, respectively. An identical CNN block was constructed and applied on protein embedding as well. Subsequently, the feature vectors for each drug-target pair are concatenated and fed into the three FC layers coined DeepDTA. First two FC layers contain a similar number of hidden nodes equal to 1,024, and a dropout layer follows each one of them to avoid overfitting as a regularization technique, as introduced in Srivastava et al. (2014). The last FC layer has a smaller number of nodes equal to 512 that is followed by the output layer. ReLU (Nair and Hinton, 2010) layer is implements $J(x) = max(0, x)$ that was used as the activation function (explained above). This model is following the general architecture that is illustrated in **Figure 2**, but with a different structure. Also, DeepDTA tunes several hyper-parameters such as the number of filters, filter length of the drug, filter length of the protein, hidden neurons number, batch size, dropout, optimizer, and learning rate in the validation step. The goal of this model is to minimize the difference between the predicted binding affinity value and the real binding affinity value in the training session. The goal of this model is to minimize the difference between the predicted binding affinity value and the real binding affinity value in the training session. DeepDTA performance significantly increased when using two CNN-blocks to learn feature representations of drugs and proteins. This study showed that performance is lower when using CNN to learn protein representation from the amino-acid sequence compared to other studies that are using CNN in their algorithms. This poor performance suggests CNN could not handle the order relationship in the amino-acid sequence, captured in the structural data. Öztürk et al. (2018), suggests avoiding this limitation by using an architecture more suitable for learning from long sequences of proteins, such as Long-Short Term Memory (LSTM).

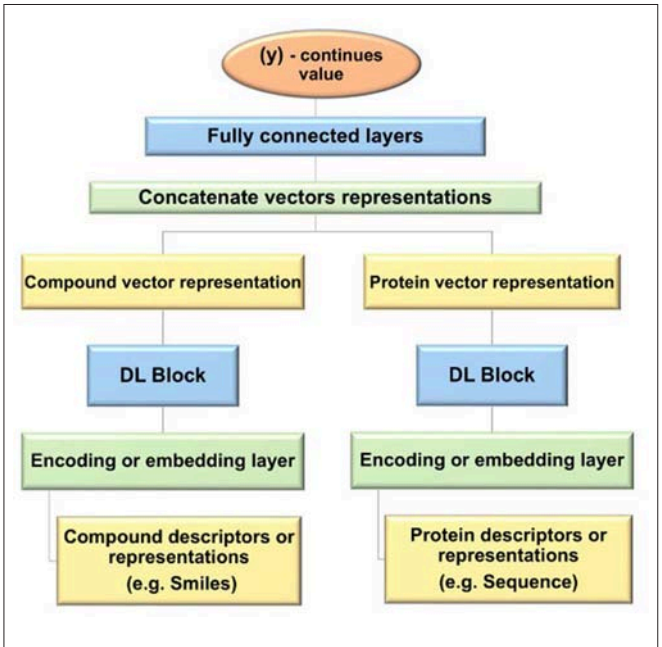


FIGURE 4 | Flowchart of the general framework of deep learning (DL) models used for drug-target binding affinity (DTBA) prediction.

WideDTA

To overcome the difficulty of modeling proteins using their sequences, the authors of DeepDTA attempted to improve the performance of DTBA prediction by developing a new method names WideDTA (made available through the e-print archives, arXiv) (Öztürk et al., 2019). WideDTA uses input data such as Ligand SMILES (LS) and amino acid sequences for protein sequences (PS), along with two other text-based information sources Ligand Maximum Common Substructure (LMCS) for drugs and Protein Domains and Motifs (PDM) based on PROSITE. Unlike DeepDTA, WideDTA represents PS and LS as a set of words instead of their full-length sequences. A word in PS is three-residues in the sequence, and a word in LS is 8-residues in the sequence. They claim, shorter lengths of residues that represent the features of the protein, are not detected using the full-length sequences due to the low signal to noise ratio. Thus, they proposed the word-based model instead of a character-based

model. WideDTA is a CNN DL model that uses as input all four text-based information sources (PS, LS, LMCS, PDM) and predict binding affinity. It first uses the Keras Embedding layer (Erickson et al., 2017) to represent words with 128-dimensional dense vectors to fed integer encode inputs. Then, it sequentially applies two 1D-CNN layers with 32 and 64 filters, followed by a max-pooling layer by the activation function layer, ReLU:

$$Features <- ReLU(pool(conv2(conv1(input))))$$

Four models with the same architecture are used to extract features from each of the text-based information sources (PS, LS, LMCS, PDM). The output features from each model are then concatenated and fed to three fully connected (FC) layers (with two drop out layers to avoid overfitting problems) that predict the binding affinity.

PADME

PADME (Protein And Drug Molecule interaction prEdiction; Feng, 2019), is another DL-based method that applies drug-target features and fingerprints to different deep neural network architectures, to predict the binding affinity values. There are two versions of PADME. The first one called PADME-ECFP, uses the Extended-Connectivity Fingerprint (Rogers and Hahn, 2010) as input features that represent drugs. The second version called PADME-GraphConv integrates Molecular Graph Convolution (MGC) (Liu et al., 2019) into the model. This is done by adding one more Graph Convolution Neural Network (GCNN) layer (which is a generalization of CNN), which is used to learn the latent features of drugs from SMILES (i.e., from graphical representation). Both PADME versions use Protein Sequence Composition (PSC) (Michael Gromiha, 2011) descriptors, which contain rich information to represent the target proteins. After generating the feature vectors for each drug and target protein, a feature vector for each drug-target pair is fed into a simple FNN to predict the DTBA. Techniques used for regularization in the FNN includes dropout, early stopping, and batch normalization. The ReLU activation functions are used for the FC layers. The cross-validation (CV) process revealed the best hyperparameter (such as batch size, dropout rate, etc.) or combination thereof that is fixed and used to evaluate the test data.

DeepAffinity

DeepAffinity (Karimi et al., 2019) is a novel interpretable DL model for DTBA prediction, which relies only on using the SMILES representation of drugs and the structural property sequence (SPS) representation that annotates the sequence with structural information to represent the proteins. The SPS is better than other protein representations because it gives structural details and higher resolution of sequences (specifically among proteins in the same family), that benefits regression task. The SPS being better than other protein representations may also be as a consequence of the SPS sequence being shorter than other sequences. Both drug SMILES and protein SPS are encoded into embedding representations using a recurrent neural network (RNN) (Ghatak, 2019). RNN model named seq2seq (Shen and Huang, 2018) is used widely and successfully in natural language processing. The seq2seq model is an auto-encoder model that consists of a recurrent unit called "encoder" that maps sequence (i.e., SMILES/SPS) to a fixed dimensional vector, and other recurrent unit called "decoder" that map back the fixed-length vector into the original sequence (i.e., SMILES/SPS). These representation vectors that have been learned in an unsupervised fashion capture the non-linear mutual dependencies among compound atoms or protein residues. Subsequently, the RNN encoders and its attention mechanisms which are introduced to interpret the predictions, are coupled with a CNN model to develop feature vectors for the drugs and targets separately. The CNN model consists of a 1D convolution layer followed by a max-pooling layer. The output representation of the CNNs for both the drugs and targets are then concatenated and fed into FC layers to output the final results, DTBA values. The entire unified RNN-CNN pipeline, including data representation, embedding learning (unsupervised learning), and joint supervised learning trained from end to end, achieved very high accuracy results compared to ML models that use the same dataset (Karimi et al., 2019).

EVALUATION OF THE STATE-OF-THE-ART METHODS

Since KronRLS, SimBoost, DeepDTA, DeepAffinity, WideDTA, and PADME are the only computational non-structure-based methods developed for prediction of DTBA to-date, we consider them the baseline methods. Here, we compare the performance of KronRLS, SimBoost, DeepDTA, WideDTA, and PADME, using the same benchmark datasets for evaluation. We excluded DeepAffinity from this comparison since it used different datasets which are based on BindingDB database (Liu et al., 2007). Also, when methods have more than one version, the comparison only includes the version that performs the best, based on identical evaluation metrics published for each method.

Evaluation Metrics

The evaluation of the performance in these regression-based models uses five metrics:

- *Concordance Index (CI)*, first introduced by Gönen and Heller (2005), and was used first for evaluation in the development of KronRLS. *CI* is a ranking metric for continuous values that measure whether the predicted binding affinity values of two random drug-target pairs were predicted in the same order as their actual values were:

$$CI = \frac{1}{Z} \sum_{s_i > s_j} h(b_i - b_j) \tag{12}$$

where b_i is the prediction value for the larger affinity s_i, b_j is the prediction value for the smaller affinity s_j, Z is a normalization constant, and $h(x)$ is the Heaviside step function (Davies, 2012), which is a discontinuous function defined as:

$$h(x) = \begin{cases} 1, & x>0 \\ 0.5, & x=0 \ \& \\ 0, & x<0 \end{cases}$$

where its value is either equal to zero when the input is negative or equal to one when the input is positive.

- *Mean Square Error (MSE)* (Wackerly et al., 2014) is commonly used as a loss function (i.e., error function) in regression task to measure how close the fitted line, that is represented by connecting the estimated values, is to the actual data points. The following formula defines the *MSE*, in which *P* denotes the prediction vector, *Y* denotes the vector of the actual outputs, and *n* is the number of samples. The square is used to ensure the negative values do not cancel the positive values. The value of *MSE* is close to zero, thus the smaller the *MSE*, the better the performance of the regressor (i.e., estimator):

$$MSE = \sum_{i=1}^{n} (P_i - Y_i) \tag{13}$$

- *Root Mean Squared Error (RMSE)* (Wackerly et al., 2014) is another metric to evaluate the regressor where it is the square root of *MSE*.

$$RMSE = \sqrt[2]{MSE} \tag{14}$$

RMSE is the distance, on average, of data points from the fitted line.

- *Pearson correlation coefficient (PCC)* (also known as Person's *R*; Kullback and Leibler, 1951) measures the difference between the actual values and the predicted values by measuring the linear correlation (association) between these two variables. The range of *PCC* is between +1 and −1, where +1 is a total positive linear correlation, −1 is a total negative linear correlation, and 0 is a non-linear correlation which indicates that there is no relationship between the actual values and the predicted values. The formula of *PCC* is defined as follows:

$$PCC = \frac{cov\,(p,y)}{std\,(p)\,std\,(y)} \tag{15}$$

where *cov* denotes the covariance between original values *y* and predicted values, and *std* denotes the standard deviation. The disadvantage is, *PCC* is only informative when used with variables that have linear correlation, as *PCC* results are misleading when used with non-linearly associated variables (Liu J. et al., 2016).

- *R-squared* (R^2) (Kassambara, 2018) is the proportion of variation in the outcome that is explained by the predictor variables. The R^2 corresponds to the squared correlation between the actual values and the predicted values in multiple regression models. The higher the *R*-squared, the better the model.

CI and *RMSE* are the only evaluation metrics reported by all the baseline methods; other metrics are reported but not by all the methods compared in this section. Also, *RMSE* and *MSE* represent the error function of the same type of error (i.e., mean square error) so reporting one of them is enough.

Validation Settings

The performance of the methods in different prediction tasks is evaluated using various CV settings. The chosen setting can affect accuracy and make the evaluation results less realistic. KronRLS (Pahikkala et al., 2015) reported using three different CV settings that make the performance evaluation more accurate and realistic. One can split the input data (that is, how the set of drug-target pairs and their affinity labels, are split into training and testing dataset) in various ways, and this splitting of data defines the validation settings used. There are three main ways used to split input data:

- *Setting 1 (S1):* Random drug-target pair which correspond to regular *k*-fold CV that split the data into *k*-folds randomly, and keeps one of these folds for testing. That is, the training phase includes a significant portion of all the drug-target pairs, while the testing phase includes the remaining random pairs.
- *Setting 2 (S2):* New drug, which means the drug is missing from the training data corresponding to leave one drug out validation (LDO).
- *Setting 3 (S3):* New target, which means the target is missing from the training data corresponding to leave one target out validation (LTO).

KronRLS and PADME methods used these settings to evaluate subsequently developed DTI and DTBA prediction methods.

Method Comparison

Tables 2, **3** summarize the performance of the baseline methods using all CV settings based on *RMSE* and *CI*, respectively. SimBoost and PADME reported *RMSE* in their respective publications. However, DeepDTA and WideDTA reported only *MSE*, so we calculated *RMSE* by taking the square root of their reported *MSE* values as defined by Equation (13). The KronRLS method did not report *RMSE* or *MSE*. However, the SimBoost paper calculated and reported *RMSE* for the KronRLS method (included in **Table 2**). Some of these baseline methods were only evaluated based on select datasets, while others only applied specific settings. All three dataset (Davis, Metz, and KIBA) were used to evaluate the performances of the SimBoost and PADME methods (based on self-reported results). The performance of PADME was also assessed using the ToxCast dataset. PADME is the first to use the ToxCast dataset. Moreover, PADME performances are reported using each dataset with the three settings (S1, S2, and S3) described above. However, SimBoost only provides its performance using one setting (S1) for each dataset.

Thus, we added performance results at specific settings not found in the original manuscripts, as calculated and reported in studies published later, to compare differences in performance (these are denoted by stars *, see **Tables 2**, **3** legend). In some instances, the results reported by other methods differ from the self-reported results. There are two reasons the results difference. The first is using different statistics of the datasets. For example, some methods, such as PADME, filter the KIBA dataset as well as adjusts the thresholds of other settings. The authors of PADME explained in their study, "*Because of the limitations of SimBoost and KronRLS, we filtered the datasets… Considering the huge*

TABLE 2 | RMSE calculated using multiple settings for all baseline methods.

Dataset	Setting	Method/reference of results					
		KronRLS	**SimBoost**	**DeepDTA**	**PADME-EFP**	**PADME-GC**	**WideDTA**
Davis (K_d)	*s1*	0.608*** 0.61562** 0.57294*	**0.247** 0.53103** 0.48197*	0.5109	**0.43219**	0.43225	0.5119
	s2	0.84048*	N/A	N/A	**0.78535**	0.80644	N/A
	s3	0.65964*	N/A	N/A	**0.56005**	0.57840	N/A
Metz (K_i)	*s1*	0.562*** 0.78128*	0.1660 0.58154*	N/A	**0.55293**	0.59926	N/A
	s2	0.78429*	N/A	N/A	**0.71170**	0.74292	N/A
	s3	0.89889*	N/A	N/A	**0.79154**	0.81893	N/A
KIBA	*s1*	0.620*** 0.64109** 0.65664*	**0.204** 0.47117** 0.46888*	0.4405	0.43214	**0.418691**	0.42308
	s2	0.70243*	N/A	N/A	**0.60201**	0.62029	N/A
	s3	0.68111*	N/A	N/A	**0.61677**	0.62345	N/A
ToxCast	*s1*	N/A	N/A	N/A	**0.40563**	0.40779	N/A
	s2	N/A	N/A	N/A	0.4485	**0.44502**	N/A
	s3	N/A	N/A	N/A	**0.48698**	0.49439	N/A

*The star symbols denote results that are not self-reported, i.e., the single star * indicates that PADME reported the other methods results, double stars ** indicates that DeepDTA reported the other methods results, and the triple stars *** indicates that SimBoost reported the other methods results. Missing data are indicated with N/A. The best values for each setting are indicated in bold font.*

compound similarity matrix required and the time-consuming matrix factorization used in SimBoost, it would be infeasible to work directly on the original KIBA dataset. Thus, we had to filter it rather aggressively so that the size becomes more manageable." Therefore, the authors of PADME reported different values for the *RMSE* scores of KronRLS and SimBoost, as shown in **Table 2**. The second reason is related to the CV settings such as the number of folds, the random seeds to split the data into training and testing, and the number of repeated experiments. The best values for each setting are indicated in bold font in **Table 2**.

Tables 2, **3** show that the SimBoost, DeepDTA, and WideDTA methods cannot handle the new drug and target settings (indicated by the missing data). From the methods that provide performances for all settings, we observe better performances using S1 setting (random pairs) compared to both S2 and S3 settings. The better performances acquired using S1 setting is expected for all methods and all datasets since it is the most informative. Better performances were also observed for S3 setting as compared to S2 setting, suggesting that the prediction of DTBA for new targets is more straightforward than the prediction of DTBA for new drugs (Pahikkala et al., 2015). However, we observe better performances for S2 setting than S3 setting when the number of targets is much lower than the number of drugs, as is the case for the Metz and ToxCast datasets.

From **Tables 2**, **3**, we further conclude that overall, the DL-based methods outperform AI/ML-based methods in predicting DTBA. However, SimBoost error rate is smaller than other methods for specific datasets indicating that there are some characteristics of SimBoost and KronRLS that can improve prediction performance. In **Table 4**, we provide a comparison of all methods to summarize the characteristics of the methods shedding light on the differences that may be contributing to improved performance. The two AI/ML methods are similarity-based (SimBoost combines similarity and features), while the DL methods are features-based. These features were obtained automatically from the raw data using DL without doing any handcrafted feature engineering as in ML. Thus, developing DL-based methods for DTBA prediction eliminates the limitation of the ML methods associated with manual alteration of data. Different representations for both drugs and targets also present advantages discussed separately with each method above, and we provide recommendations concerning the use of different representation in the last section below.

The comparison table also shows all DL-based methods reported up to now, used CNN to learn the features for both drugs and targets. The robust feature of CNN is its ability to capture local dependencies for both sequence and structure data. CNN is additionally computationally efficient since it uses unique convolution and pooling operations and performs parameter sharing (Defferrard et al., 2016). All DL methods use the same activation function, ReLU, which is the most widely used activation function for many reasons (Gupta, 2017). First, ReLU is non-linear function so it can easily backpropagate an error. Second, ReLU can have multiple layers of neurons, but it does not activate all these neurons at the same time. The last advantage of ReLU function is that it converts negative values of the input to zero values, and the neurons are not activated, so the network will be sparse which means easy and efficient of computation.

We can also observe from **Table 4**, that KronRLS, SimBoost, and PADME methods are suitable for both classification and regression problems. It is better to generalize the model to work on more than one application by making it suitable for both

TABLE 3 | CI across multiple datasets of all baseline methods.

Dataset	Setting	Method/reference of results					
		KronRLS	**SimBoost**	**DeepDTA**	**PADME-ECFP**	**PADME-GC**	**WideDTA**
Davis (K_d)	*s1*	0.8830 0.8710** 0.87578*	0.8840 0.872** 0.8871*	0.8780	0.90388	**0.90389**	0.8860
	s2	**0.7480** 0.69245*	N/A	N/A	0.71630	0.72001	N/A
	s3	**0.8610** 0.80751*	N/A	N/A	0.85503	0.84483	N/A
Metz (K_i)	*s1*	0.7930 0.748522*	**0.8510** 0.79439*	N/A	0.80756	0.79400	N/A
	s2	0.7360 0.70916*	N/A	N/A	**0.74240**	0.74104	N/A
	s3	0.6660 0.647*	N/A	N/A	0.69830	**0.70796**	N/A
KIBA	*s1*	0.782** 0.7831*	0.8470 0.836** 0.84046*	0.8630	0.85745	0.86370	**0.8750**
	s2	0.6890*	N/A	N/A	**0.77310**	0.75450	N/A
	s3	0.7122*	N/A	N/A	**0.77167**	0.76790	N/A
ToxCast	*s1*	N/A	N/A	N/A	0.79655	**0.79871**	N/A
	s2	N/A	N/A	N/A	0.72057	**0.7286**	N/A
	s3	N/A	N/A	N/A	0.68481	**0.69050**	N/A

*The star symbols denote results that are not self-reported, i.e., the single star * indicates that PADME reported the other methods results, and the double stars ** indicates that DeepDTA reported the other methods results. Missing data are indicated with N/A. The best values for each setting are indicated in bold font.*

DTBA and DTIs predictions using the appropriate benchmark datasets and correct evaluation metrics.

LIMITATIONS OF AI/ML/DL-BASED METHODS

AI/ML/DL-based computational models developed for DTBA prediction show promising results. However, all such models suffer from limitations that if avoided, may improve performance.

AI/ML-Based Methods

Similarity-based approaches used by these methods usually do not take into considerations the heterogeneous information defined in the relationship network. Avoiding this limitation requires integrating a feature-based approaches with the similarity-based approaches. Another limitation is that AI/ML-based models require extensive training, and each application requires specific training for the application-specific purpose. Moreover, shallow network-based methods with sequence data usually do not learn well some of the crucial features (such as distance correlation) that may be needed for accurate prediction.

DL-Based Methods

The use of these methods is currently trending despite DL models creating "black boxes" that are difficult to interpret due to the learning features integrated into the data for modeling. Limitations faced with the use of DL models involve the requirement of the large amount of high-quality data, which are frequently kept private and is very expensive to generate. Not using a sufficiently large volume of high-quality data affects the reliability and performance of DL models. The other limitation is that the engineered features (generated automatically), are not intuitive, and the DL-based models developed lack rational interpretation of the biological/chemical aspects of the problem in question.

DISCUSSION

Here we attempt to extract useful insights from the characteristics of the methods developed for DTBA prediction, suggest possible future avenues to improve predictions, and highlight the existing problems that need a solution. Our recommendations are grouped under several sub-sections to focus on different aspects of improvements of prediction performance of DTBA.

Using More Comprehensive Information

Integrating information from different sources of drug and target data can improve the prediction performance. These sources can include but are not limited to drug side-effects, drug-disease association, and drug interactions. For targets, examples of other sources of information are protein-protein interaction, protein-diseases association, and genotype-phenotype association. To the best of our knowledge, no

TABLE 4 | Baseline methods features.

Characteristics	Methods				
	1. KronRLS	**2. SimBoost**	**3. DeepDTA**	**4. WideDTA**	**5. PADME**
Datasets	Davis, Metz	Davis, Metz, Kiba	Davis, Kiba	Davis, Kiba	Davis, Metz, Kiba, ToxCast
ML/DL	AI/ML	AI/ML	DL	DL	DL
Similarity (OR) Feature based method	Similarity-based	Similarity and feature based	Feature-based	Feature-based	Feature-based
Drug representation (or features)	PubChem Sim Chemical kernels	PubChem Sim + statistical and network features	SMILES	SMILES + LMCS	SMILES / ECFP
Protein representation (or features)	SW sim score, Normalized SW sim score	SW sim score	aaseq	aaseq + PDM	PSC
NN type for features learning			CNN	two 1D-CNN	GCNN
NN type for prediction			3 FC layers	FC layer	Feedforward NN
Regressor/OR/activation function	KronRLS model	Gradient boosting model	ReLU	ReLU	ReLU
Validation setting	S1, S2, S3	S1	S1	S1	S1, S2, S3
Cross Validation	Repeated 10-folds CV, Nested CV, LDO-CV, LTO-CV	10 times 5 folds CV, LDO-CV, LTO-CV	5 folds CV	6 folds CV	5 folds CV, LDO-CV, LTO-CV
Performance metrics	CI, MSE	CI, RMSE	CI, MSE, PCC	CI, MSE, PCC	CI, RMSE, R^2
Classification/Regression	Both	Both	Regression	Regression	Both
Year	2014	2017	2018	2019	2018

ML, Machine Learning; DL, Deep Learning; Sim, Similarity; aaseq, amino-acid sequence; SPS, structural property sequence; PSC, protein sequence composition; PDM, protein domain and motif; ECFP, extended-connectivity fingerprint; LMCS, ligand maximum common substructure; KronRLS, Kronecker Regularized Least Square; CNN, convolutional neural network; GCNN, graph convolution neural network; RNN, recurrent neural network; FC, fully connected; ReLU, rectified linear unit; CV, cross validation; LDO, leave one drug out; LTO, leave one target out; MSE, Mean Square Error; RMSE, root square of mean square error; CI, concordance index; PCC, Pearson correlation coefficient.

method uses such information for DTBA prediction except KronRLS, which integrates some other sources of information in the form of similarity matrices. However, there are different DTIs prediction works that integrate different sources of information, which help in boosting the prediction performance. For example, some studies predicted DTIs by integrating drug side-effects information (Campillos et al., 2008; Mizutani et al., 2012), or drug-diseases interaction (Wang W. et al., 2014; Luo et al., 2017). Other studies used public gene expression data (Sirota et al., 2011), gene ontology (Tao et al., 2015), transcriptional response data (Iorio et al., 2010), or have integrated several of these resources (Alshahrani and Hoehndorf, 2018). DTBA prediction methods can benefit from these previous studies through integration of these different sources of information.

Input Data Representation

Different representations can be used for both drugs and targets (see **Table 4**). For example, SMILES, max common substructure, and different kinds of fingerprints can be used to represent drugs. These representations significantly affect the prediction performance. Thus, it is essential to start with appropriate representations by deciding which features from these representations are intended to obtain. Each representation has its own advantages as discussed above when comparing methods.

Similarity Calculation, Selection, and Information Fusion

There are several types of similarities that can be calculated using different sources of information, such as the multiple drug-drug similarities based on the chemical structures or based on side-effects. There are also other drug-drug similarities based on specific SMILES embeddings. The same goes for the target-target similarities, which can use other sources of information such as amino-acid sequence, nucleotide sequences, or protein-protein interaction network. Choosing suitable drug-drug and target-target similarities also contribute significantly to the prediction performance under different settings (either for DTBA or DTI prediction). If all similarities are combined, it will lead to introducing some noise as well as the most informative similarities will be affected by the less informative similarities. Thus, it is essential to apply a similarity selection method in order to select the most informative and robust subset of similarities among all similarities as introduced in Olayan et al. (2018). Integrating multiple similarities (i.e., a subset of similarities) has the advantage of complementary information for different similarities as well as avoiding dealing with a different scale. One could use the Similarity Network Fusion (SNF) (Wang B. et al., 2014) algorithm for data integration in a non-linear fashion to predict DTBA with multiple similarities. There are other integration algorithms or functions such as SUM, AVG, and MAX functions. Also, multi-view graph autoencoder algorithm

(GAE) (Baskaran and Panchavarnam, 2019) proved its efficiency in integrating drug similarities (Ma T. et al., 2018).

Integration of Computational Methods

Future *in silico* methods for DTBA prediction will benefit from the integration of diverse methods and approaches. Methods can be developed using different techniques, such as network analysis (Zong et al., 2019), matrix factorization (Ezzat et al., 2017), graph embeddings (Crichton et al., 2018), and more. Feature-based models and similarity-based models can be combined as well, as has been done in the SimBoost method. Furthermore, AI/ML/DL methods can be combined in different ways, (1) by combining some essential hand-crafted features from AI/ML and auto-generated features from DL, (2) using AI/ML for feature engineering and DL for prediction.

Network Analysis and Graph Mining Techniques

Since graph mining and graph embedding approaches are very successful in the prediction of DTIs (Luo et al., 2017; Olayan et al., 2018), we can apply some of these techniques to DTBA. To apply this technique to DTBA we can formulate a weighted undirected heterogeneous graph G(V, E), where V is the set of vertices (i.e., drugs and targets), and E is the set of edges that represent the binding strength values. Multiple target-target similarities and drug-drug similarities can be integrated into the DTBA graph to construct a complete interaction network. After that, graph mining techniques such as Daspfind (Ba-Alawi et al., 2016) that calculate simple path scores between drug and target can be applied. Also, graph embedding techniques such as DeepWalk (Perozzi et al., 2014), node2vec (Grover and Leskovec, 2016), metapath2vec (Dong et al., 2017; Zhu et al., 2018), or Line (Tang et al., 2015) can be applied to the DTBA graph to obtain useful features for prediction. There are different graph embedding techniques that can be used for features learning and representation as summarized by Cai et al. (2018) and Goyal and Ferrara (2018a,b). To the best of our knowledge no published DTBA prediction method formulate the problem as a weighted graph and apply such techniques.

Deep Learning

For the computational prediction of DTIs and DTBA, DL and features learning (i.e., embedding) are currently the most popular techniques since they are efficient in generating features and addressing scalability for large-scale data. DL techniques are capable of learning features of the drugs, targets, and the interaction network. Furthermore, when using heterogeneous information sources for drugs and targets, DL techniques can be applied to obtain additional useful features. DL techniques including different types of NN can extract useful features not just from the sequence-based representation of drug (i.e., SMILES) and protein (i.e., amino acid) as done by Öztürk et al. (2018, 2019), but also from the graph-based representation. For example, CNN, or GCNN can be applied on SMILES (that are considered graphs) to capture the structural information of the molecules (i.e., drugs). It is highly recommended to attempt to apply DL and feature learning techniques on graph-based techniques as well as a heterogeneous graph that combine different information about drugs and targets to enhance the DTBA predictive model. Several steps should be applied to develop a robust DL model: starting with selecting the suitable data representation, deciding about NN type and DL structures, then choosing the optimal hyperparameter set. The decisive advantage of the DL techniques worth mentioning is to implement the running of code on the Graphics Processing Unit (GPU). In terms of time complexity, DL-based methods that run on GPUs, drastically decrease computational time compared to running the method on a CPU. Guidelines to accelerate drug discovery applications using GPU as well as a comparison of recent GPU and CPU implementations are provided in Gawehn et al. (2018).

Multi-Output Regression Methods

Given that DTBA can be measured using several output properties (e.g., IC_{50} and K_i,), it is a laborious task to develop one model to predict each property individually. Therefore, it is much more efficient to generate a model that can predict several output properties, such as multi-output regression models (also known as multi-target regression), which aims at predicting several continuous values (Borchani et al., 2015). Multi-output regression differs from multi-label classification, which aims at predicting several binary labels (e.g., positive or negative; Gibaja and Ventura, 2014). Multi-output regression methods take into consideration correlations between output properties in addition to input conditions (e.g. organism and cell line). Borchani et al. (2015) recently wrote a review that covers more in-depth details regarding the multi-output regression methods. Moreover, Mei and Zhang (2019) demonstrated how multi-label classification methods could be applied for DTI prediction. In this study, each drug is considered a class label, and target genes are considered input data for training. To the best of our knowledge, multi-output regression methods have not been applied for DTBA prediction. The main challenge in applying multi-output regression to DTBA is missing data. Output properties (and sometimes input conditions) may not be available for all drug-target pairs in the dataset. However, several multi-label classification methods have been applied for handling missing data in multi-output datasets (Wu et al., 2014; Xu et al., 2014; Yu et al., 2014; Jain et al., 2016).

Validation Settings

Overall, the methods further show that three settings for the CV are used to evaluate the prediction model. However, there are still many studies that only use the typical CV setting of random pair for evaluation (S1 setting), which leads to overoptimistic prediction results. Thus, models should be evaluated using all three settings. Models can also be evaluated using (a rarely used) fourth setting wherein both the drug and target are new (Pahikkala et al., 2015; Cichonska et al., 2017), and it is even better to evaluate the model under this setting as well, to see how good it is in predicting DTI when both the drug and the target are new. Evaluating the model under the four settings will avoid over-optimistic results. The CV is essential for adjusting

the hyperparameters for both AI/ML and DL models. It is also essential to handle the overfitting problem. Overfitting happens when a model learns many details, including noise from the training data and fits the training data very well but cannot fit the test data well (Domingos, 2012). Overfitting can be evaluated by assessing how good the model is fitted to training data using some strategies that were recommended in Scior et al. (2009) and Raies and Bajic (2016) using two statistical parameters: *S*, standard error of estimation (Cronin and Schultz, 2003), and R^2, coefficient of multiple determination (Gramatica, 2013), which will not be discussed in detail in this review.

Evaluation Metrics

The choice of the suitable measure to evaluate DTBA prediction model is very important. Since DTBA prediction is a regression model, the evaluation metrics commonly used is *CI* and *RMSE,* as explained above. Nonetheless, other metrics (such as *R* and *PCC*) are partially used in assessment of DTBA prediction models. Using several metrics is essential as every metric carries disadvantages, which forces researchers to consider multiple evaluation metrics (Bajić, 2000) in performance evaluation to assess the model effectiveness in an accurate manner and from different perspectives. For example, *MSE* and *RMSE* are more sensitive to outliers (Chai and Draxler, 2014). *RMSE* is not a good indicator of average model performance and is a misleading indicator of average error. Thus, Mean Absolute Error (MAE) would be a better metric, as suggested by Willmott et al. (2009). So, it is better to have multiple evaluation metrics to get benefit from each one's strengths and evaluate the model from a different perspective.

CONCLUSION

Both DTIs and DTBA predictions play a crucial role in the early stages of drug development and drug repurposing. However, it is more meaningful and informative to predict DTBA rather than predicting just on/off interaction between drug and target. An overview of the computational methods developed for DTBA prediction are summarized, but we specifically focused with more details on the recent AI/ML/DL-based methods developed to predict DTBA without the limitations imposed by 3D structural data. The available datasets for DTBA are summarized, and the benchmark datasets are discussed with details including definitions, sources, and statistics. For future research, computational prediction of DTBA remains an open problem. There is a lot of space to improve the existing computational methods from different angles as discussed in the recommendations. As the data is growing so fast, it is important to keep updating the prediction and updating evaluation datasets as well. After updating the data, it is necessary to customize, refine, and scale the current DTBA models, and to develop more efficient models as well.

AUTHOR CONTRIBUTIONS

MT designed the study and wrote the first draft of the manuscript. MT and AR designed the figures. AR and SA contributed to discussions and writing of specific sections of the manuscript. ME and VB supervised and critically revised the manuscript. All authors read and approved the final manuscript.

FUNDING

VB has been supported by the King Abdullah University of Science and Technology (KAUST) Baseline Research Fund (BAS/1/1606-01-01) and ME has been supported by KAUST Office of Sponsored Research (OSR) Award No. FCC/1/1976-24-01.

ACKNOWLEDGMENTS

The research reported in this publication was supported by the King Abdullah University of Science and Technology (KAUST).

REFERENCES

Abel, R., Manas, E. S., Friesner, R. A., Farid, R. S., and Wang, L. (2018). Modeling the value of predictive affinity scoring in preclinical drug discovery. *Curr. Opin. Struct. Biol.* 52, 103–110. doi: 10.1016/j.sbi.2018.09.002

Agrawal, P., Raghav, P. K., Bhalla, S., Sharma, N., and Raghava, G. P. S. (2018). Overview of free software developed for designing drugs based on protein-small molecules interaction. *Curr. Top. Med. Chem.* 18, 1146–1167. doi: 10.2174/1568026618666180816155131

Ahmed, A., Smith, R. D., Clark, J. J., Dunbar, J. B., and Carlson, H. A. (2015). Recent improvements to Binding MOAD: a resource for protein–ligand binding affinities and structures. *Nucleic Acids Res.* 43, D465–D469. doi: 10.1093/nar/gku1088

Ain, Q. U., Aleksandrova, A., Roessler, F. D., and Ballester, P. J. (2015). Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 5, 405–424. doi: 10.1002/wcms.1225

Alshahrani, M., and Hoehndorf, R. (2018). Drug repurposing through joint learning on knowledge graphs and literature. *bioRXiv [Preprint]*. doi: 10.1101/385617

Andricopulo, A. D., and Ferreira, L. L. G. (2019). Chemoinformatics approaches to structure- and ligand-based drug design. *Front. Media SA*. 9:1416. doi: 10.3389/978-2-88945-744-1

Antunes, D. A., Abella, J. R., Devaurs, D., Rigo, M. M., and Kavraki, L. E. (2019). Structure-based methods for binding mode and binding affinity prediction for peptide-MHC complexes. *Curr. Top. Med. Chem.* 18, 2239–2255. doi: 10.2174/1568026619666181224101744

Arrowsmith, J. (2011). Trial watch: phase II failures: 2008–2010. *Nat. Rev. Drug Discov.* 10, 328–329. doi: 10.1038/nrd3439

Ashtawy, H. M., and Mahapatra, N. R. (2018). Task-specific scoring functions for predicting ligand binding poses and affinity and for screening enrichment. *J. Chem. Inf. Model* 58, 119–133. doi: 10.1021/acs.jcim.7b00309

Ba-Alawi, W., Soufan, O., Essack, M., Kalnis, P., and Bajic, V. B. (2016). DASPfind: new efficient method to predict drug-target interactions. *J. Cheminform.* 8:15. doi: 10.1186/s13321-016-0128-4

Bachmann, K. A., and Lewis, J. D. (2005). Predicting inhibitory drug—drug interactions and evaluating drug interaction reports using inhibition constants. *Ann. Pharmacother.* 39, 1064–1072. doi: 10.1345/aph.1E508

Bajić, V. B. (2000). Comparing the success of different prediction software in sequence analysis: a review. *Brief. Bioinformatics* 1, 214–228. doi: 10.1093/bib/1.3.214

Baskaran, S., and Panchavarnam, P. (2019). Data integration using through attentive multi-view graph auto-encoders. *Int. J. Sci. Res. Comp. Sci. Eng. Inf. Technol.* 5, 344–349. doi: 10.32628/CSEIT195394

Benson, M. L., Smith, R. D., Khazanov, N. A., Dimcheff, B., Beaver, J., Dresslar, P., et al. (2007). Binding MOAD, a high-quality protein ligand database. *Nucl. Acids Res.* 36, D674–D678. doi: 10.1093/nar/gkm911

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Block, P., Sotriffer, C. A., Dramburg, I., and Klebe, G. (2006). AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res.* 34, D522–526. doi: 10.1093/nar/gkj039

Borchani, H., Varando, G., Bielza, C., and Larrañaga, P. (2015). A survey on multi-output regression. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 5, 216–233. doi: 10.1002/widm.1157

Bulusu, K. C., Guha, R., Mason, D. J., Lewis, R. P., Muratov, E., Motamedi, Y. K., et al. (2016). Modelling of compound combination effects and applications to efficacy and toxicity: state-of-the-art, challenges and perspectives. *Drug Discov. Today* 21, 225–238. doi: 10.1016/j.drudis.2015.09.003

Burlingham, B. T., and Widlanski, T. S. (2003). An intuitive look at the relationship of K_i and IC50: a more general use for the dixon plot. *J. Chem. Educ.* 80:214. doi: 10.1021/ed080p214

Cai, H., Zheng, V. W., and Chang, K. C. (2018). A comprehensive survey of graph embedding: problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.* 30, 1616–1637. doi: 10.1109/TKDE.2018.2807452

Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J., and Bork, P. (2008). Drug target identification using side-effect similarity. *Science* 321, 263–266. doi: 10.1126/science.1158140

Chai, T., and Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* 7, 1247–1250. doi: 10.5194/gmd-7-1247-2014

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discov. Today* 23, 1241–1250. doi: 10.1016/j.drudis.2018.01.039

Chen, X., Liu, M., and Gilson, M. (2001). BindingDB: a web-accessible molecular recognition database. *Comb. Chem. High Throughput Screen.* 4, 719–725. doi: 10.2174/1386207013330670

Cheng, A. C., Coleman, R. G., Smyth, K. T., Cao, Q., Soulard, P., Caffrey, D. R., et al. (2007). Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* 25, 71–75. doi: 10.1038/nbt1273

Chou, T. C., and Talalay, P. (1984). Quantitative analysis of dose-effect relationships: the combined effects of multiple drugs or enzyme inhibitors. *Adv. Enzyme Regul.* 22, 27–55. doi: 10.1016/0065-2571(84)90007-4

Cichonska, A., Ravikumar, B., Parri, E., Timonen, S., Pahikkala, T., Airola, A., et al. (2017). Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. *PLoS Comput. Biol.* 13:e1005678. doi: 10.1371/journal.pcbi.1005678

Colwell, L. J. (2018). Statistical and machine learning approaches to predicting protein–ligand interactions. *Curr. Opin. Struct. Biol.* 49, 123–128. doi: 10.1016/j.sbi.2018.01.006

Crichton, G., Guo, Y., Pyysalo, S., and Korhonen, A. (2018). Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches. *BMC Bioinformatics* 19:176. doi: 10.1186/s12859-018-2163-9

Cronin, M. T. D., and Schultz, T. W. (2003). Pitfalls in QSAR. *J. Mol. Struct.* 622, 39–51. doi: 10.1016/S0166-1280(02)00616-4

Davies, B. (2012). *Integral Transforms and Their Applications.* New York, NY: Springer Science and Business Media.

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., et al. (2011). Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* 29, 1046–1051. doi: 10.1038/nbt.1990

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering," in *Advances in Neural Information Processing Systems* Vol. 29, eds D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Barcelona: Curran Associates, Inc.), 3844–3852.

Degliesposti, G., Portioli, C., Parenti, M. D., and Rastelli, G. (2011). BEAR, a novel virtual screening methodology for drug discovery. *J. Biomol. Screen.* 16, 129–133. doi: 10.1177/1087057110388276

Deng, W., Breneman, C., and Embrechts, M. J. (2004). Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods. *J Chem. Inf. Comput. Sci.* 44, 699–703. doi: 10.1002/chin.200426198

Dimasi, J. A., Hansen, R. W., and Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *J. Health Econ.* 22, 151–185. doi: 10.1016/S0167-6296(02)00126-1

Ding, H., Takigawa, I., Mamitsuka, H., and Zhu, S. (2014). Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Brief. Bioinformatics* 15, 734–747. doi: 10.1093/bib/bbt056

Domingos, P. M. (2012). A few useful things to know about machine learning. *Commun. ACM* 55, 78–87. doi: 10.1145/2347736.2347755

Dong, Y., Chawla, N. V., and Swami, A. (2017). "Metapath2Vec: scalable representation learning for heterogeneous networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS: ACM). doi: 10.1145/3097983.3098036

Du, X., Li, Y., Xia, Y.-L., Ai, S.-M., Liang, J., Sang, P., et al. (2016). Insights into protein–ligand interactions: mechanisms, models, and methods. *Int. J. Mol. Sci.* 17:144. doi: 10.3390/ijms17020144

Dunbar, J. B. Jr., Smith, R. D., Damm-Ganamet, K. L., Ahmed, A., Esposito, E. X., Delproposto, J., et al. (2013). CSAR data set release 2012: ligands, affinities, complexes, and docking decoys. *J. Chem. Inf. Model* 53, 1842–1852. doi: 10.1021/ci4000486

Ekins, S. (2016). The next era: deep learning in pharmaceutical research. *Pharm. Res.* 33, 2594–2603. doi: 10.1007/s11095-016-2029-7

Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., et al. (2019). Exploiting machine learning for end-to-end drug discovery and development. *Nat. Mater.* 18, 435–441. doi: 10.1038/s41563-019-0338-z

Emig, D., Ivliev, A., Pustovalova, O., Lancashire, L., Bureeva, S., Nikolsky, Y., et al. (2013). Drug target prediction and repositioning using an integrated network-based approach. *PLoS ONE* 8:e60618. doi: 10.1371/journal.pone.0060618

Erickson, B. J., Korfiatis, P., Akkus, Z., Kline, T., and Philbrick, K. (2017). Toolkits and libraries for deep learning. *J. Digit. Imaging* 30, 400–405. doi: 10.1007/s10278-017-9965-6

Ezzat, A., Wu, M., Li, X., and Kwoh, C.-K. (2019). Computational prediction of drug-target interactions via ensemble learning. *Methods Mol Biol.* 1903, 239–254. doi: 10.1007/978-1-4939-8955-3_14

Ezzat, A., Wu, M., Li, X.-L., and Kwoh, C.-K. (2016). Drug-target interaction prediction via class imbalance-aware ensemble learning. *BMC Bioinformatics* 17:509. doi: 10.1186/s12859-016-1377-y

Ezzat, A., Wu, M., Li, X.-L., and Kwoh, C.-K. (2018). Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Brief Bioinform.* 20, 1337–1357. doi: 10.1093/bib/bby002

Ezzat, A., Zhao, P., Wu, M., Li, X.-L., and Kwoh, C.-K. (2017). Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 646–656. doi: 10.1109/TCBB.2016.2530062

Feng, Q. (2019). *PADME: A Deep Learning-based Framework for Drug-Target Interaction Prediction* (Master thesis), Simon Fraser University, Burnaby, BC, Canada.

Ferrero, E., Dunham, I., and Sanseau, P. (2017). *In silico* prediction of novel therapeutic targets using gene–disease association data. *J. Transl. Med.* 15:182. doi: 10.1186/s12967-017-1285-6

Ganotra, G. K., and Wade, R. C. (2018). Prediction of drug–target binding kinetics by comparative binding energy analysis. *ACS Med. Chem. Lett.* 9, 1134–1139. doi: 10.1021/acsmedchemlett.8b00397

Gawehn, E., Hiss, J. A., Brown, J. B., and Schneider, G. (2018). Advancing drug discovery via GPU-based deep learning. *Expert Opin. Drug Discov.* 13, 579–582. doi: 10.1080/17460441.2018.1465407

Ghatak, A. (2019). Recurrent neural networks (RNN) or sequence models. *Deep Learn. R* 1, 207–237. doi: 10.1007/978-981-13-5850-0_8

Gibaja, E., and Ventura, S. (2014). Multilabel learning: a review of the state of the art and ongoing research. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 4, 411–444. doi: 10.1002/widm.1139

Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J. (2016). BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* 44, D1045–D1053. doi: 10.1093/nar/gkv1072

Gönen, M., and Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 92, 965–970. doi: 10.1093/biomet/92.4.965

Goyal, P., and Ferrara, E. (2018a). GEM: a Python package for graph embedding methods. *J. Open Source Softw.* 3:876. doi: 10.21105/joss.00876

Goyal, P., and Ferrara, E. (2018b). Graph embedding techniques, applications, and performance: a survey. *Knowl. Based Syst.* 151, 78–94. doi: 10.1016/j.knosys.2018.03.022

Gramatica, P. (2013). On the development and validation of QSAR models. *Methods Mol. Biol.* 930, 499–526. doi: 10.1007/978-1-62703-059-5_21

Grover, A., and Leskovec, J. (2016). node2vec: scalable feature learning for networks. *KDD* 2016, 855–864. doi: 10.1145/2939672.2939754

Guedes, I. A., Pereira, F. S. S., and Dardenne, L. E. (2018). Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. *Front. Pharmacol.* 9:1089. doi: 10.3389/fphar.2018.01089

Gupta, D. (2017). *"Fundamentals of Deep Learning–Activation Functions and When to Use Them?* [Online]. Available online at: https://www.analyticsvidhya.com/blog/2017/10/fundamentals-deep-learning-activationfunctions-when-to-use-them [accessed August, 1 2019].

He, T., Heidemeyer, M., Ban, F., Cherkasov, A., and Ester, M. (2017). SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *J. Cheminform.* 9:24. doi: 10.1186/s13321-017-0209-z

Heck, G. S., Pintro, V. O., Pereira, R. R., De Ávila, M. B., Levin, N. M. B., and De Azevedo, W. F. (2017). Supervised machine learning methods applied to predict ligand- binding affinity. *Curr. Med. Chem.* 24, 2459–2470. doi: 10.2174/0929867324666170623092503

Hu, L., Benson, M. L., Smith, R. D., Lerner, M. G., and Carlson, H. A. (2005). Binding MOAD (mother of all databases). *Proteins* 60, 333–340. doi: 10.1002/prot.20512

Huang, S.-Y., and Zou, X. (2006). An iterative knowledge-based scoring function to predict protein–ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.* 27, 1866–1875. doi: 10.1002/jcc.20504

Hulme, E. C., and Trevethick, M. A. (2010). Ligand binding assays at equilibrium: validation and interpretation. *Br. J. Pharmacol.* 161, 1219–1237. doi: 10.1111/j.1476-5381.2009.00604.x

Hutter, M. C. (2009). *In silico* prediction of drug properties. *Curr. Med. Chem.* 16, 189–202. doi: 10.2174/092986709787002736

Hutter, M. C. (2018). The current limits in virtual screening and property prediction. *Future Med. Chem.* 10, 1623–1635. doi: 10.4155/fmc-2017-0303

Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., et al. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14621–14626. doi: 10.1073/pnas.1000138107

Jain, A. (2017). *Deep Learning in Chemoinformatics Using Tensor Flow*. (Doctoral dissertation). UC Irvine.

Jain, H., Prabhu, Y., and Varma, M. (2016). "Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 935–944. doi: 10.1145/2939672.2939756

Jiang, J., Wang, N., Chen, P., Zhang, J., and Wang, B. (2017). DrugECs: an ensemble system with feature subspaces for accurate drug-target interaction prediction. *Biomed Res. Int.* 2017, 1–10. doi: 10.1155/2017/6340316

Jiménez, J., Škalič, M., Martínez-Rosell, G., and De Fabritiis, G. (2018). KDEEP: protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J. Chem. Inf. Model.* 58, 287–296. doi: 10.1021/acs.jcim.7b00650

Jing, Y., Bian, Y., Hu, Z., Wang, L., and Xie, X.-Q. S. (2018). Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. *AAPS J.* 20:58. doi: 10.1208/s12248-018-0210-0

Judson, R. (2012). US EPA—ToxCast and the Tox21 program: perspectives. *Toxicol. Lett.* 211:S2. doi: 10.1016/j.toxlet.2012.03.016

Kalkatawi, M., Magana-Mora, A., Jankovic, B., and Bajic, V. B. (2019). DeepGSR: an optimized deep-learning structure for the recognition of genomic signals and regions. *Bioinformatics* 35, 1125–1132. doi: 10.1093/bioinformatics/bty752

Karimi, M., Wu, D., Wang, Z., and Shen, Y. (2019). DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* 35, 3329–3338. doi: 10.1093/bioinformatics/btz111

Kassambara, A. (2018). *Machine Learning Essentials: Practical Guide in R*. STHDA.

Kontoyianni, M. (2017). Docking and virtual screening in drug discovery. *Methods Mol. Biol.* 1647, 255–266. doi: 10.1007/978-1-4939-7201-2_18

Krig, S. (2016). "Feature learning and deep learning architecture survey," in *Computer Vision Metrics* (Cham: Springer), 375–514. doi: 10.1007/978-3-319-33762-3_10

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* 22, 79–86. doi: 10.1214/aoms/1177729694

Kundu, I., Paul, G., and Banerjee, R. (2018). A machine learning approach towards the prediction of protein–ligand binding affinity based on fundamental molecular properties. *RSC Adv.* 8, 12127–12137. doi: 10.1039/C8RA00003D

Kurgan, L., and Wang, C. (2018). Survey of similarity-based prediction of drug-protein interactions. *Curr. Med. Chem.* 26:1. doi: 10.2174/0929867326666190808154841

Leach, A. R., Shoichet, B. K., and Peishoff, C. E. (2006). Prediction of Protein-Ligand Interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* 49, 5851–5855. doi: 10.1021/jm060999m

Lee, A., Lee, K., and Kim, D. (2016). Using reverse docking for target identification and its applications for drug discovery. *Expert Opin. Drug Discov.* 11, 707–715. doi: 10.1080/17460441.2016.1190706

Li, J., Fu, A., and Zhang, L. (2019). An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdiscip. Sci.* 11, 320–328. doi: 10.1007/s12539-019-00327-w

Li, Q., and Shah, S. (2017). Structure-Based Virtual Screening. *Methods Mol. Biol.* 1558, 111–124. doi: 10.1007/978-1-4939-6783-4_5

Li, Y., Huang, C., Ding, L., Li, Z., Pan, Y., and Gao, X. (2019). Deep learning in bioinformatics: introduction, application, and perspective in the big data era. *Methods* 166, 4–21. doi: 10.1101/563601

Lima, A. N., Philot, E. A., Trossini, G. H. G., Scott, L. P. B., Maltarollo, V. G., and Honorio, K. M. (2016). Use of machine learning approaches for novel drug discovery. *Expert. Opin. Drug Discov.* 11, 225–239. doi: 10.1517/17460441.2016.1146250

Liu, J., Tang, W., Chen, G., Lu, Y., Feng, C., and Tu, X. M. (2016). Correlation and agreement: overview and clarification of competing concepts and measures. *Shanghai Arch. Psychiatry* 28, 115–120. doi: 10.11919/j.issn.1002-0829.216045

Liu, K., Sun, X., Jia, L., Ma, J., Xing, H., Wu, J., et al. (2019). Chemi-Net: a molecular graph convolutional network for accurate drug property prediction. *Int. J. Mol. Sci.* 20:3389. doi: 10.3390/ijms20143389

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., and Gilson, M. K. (2007). BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* 35, D198–D201. doi: 10.1093/nar/gkl999

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing* 234, 11–26. doi: 10.1016/j.neucom.2016.12.038

Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X.-L. (2016). Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput. Biol.* 12:e1004760. doi: 10.1371/journal.pcbi.1004760

Liu, Y., Xu, Z., Yang, Z., Chen, K., and Zhu, W. (2013). A knowledge-based halogen bonding scoring function for predicting protein-ligand interactions. *J. Mol. Model.* 19, 5015–5030. doi: 10.1007/s00894-013-2005-7

Lu, J., Lu, D., Fu, Z., Zheng, M., and Luo, X. (2018). Machine learning-based modeling of drug toxicity. *Methods Mol. Biol.* 1754, 247–264. doi: 10.1007/978-1-4939-7717-8_15

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* 8:573. doi: 10.1038/s41467-017-00680-8

Ma, T., Xiao, C., Zhou, J., and Wang, F. (2018). "Drug similarity integration through attentive multi-view graph auto-encoders," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 3477–3483.

Ma, W., Yang, L., and He, L. (2018). Overview of the detection methods for equilibrium dissociation constant KD of drug-receptor interaction. *J. Pharm. Anal.* 8, 147–152. doi: 10.1016/j.jpha.2018.05.001

Mei, S., and Zhang, K. (2019). A multi-label learning framework for drug repurposing. *Pharmaceutics* 11:466. doi: 10.3390/pharmaceutics11090466

Metz, J. T., Johnson, E. F., Soni, N. B., Merta, P. J., Kifle, L., and Hajduk, P. J. (2011). Navigating the kinome. *Nat. Chem. Biol.* 7, 200–202. doi: 10.1038/nchembio.530

Michael Gromiha, M. (2011). *Protein Bioinformatics: From Sequence to Function*. New Delhi: Academic Press.

Michelucci, U. (2018). Feedforward neural networks. *Appl. Deep Learn.* 1, 83–136. doi: 10.1007/978-1-4842-3790-8_3

Mizutani, S., Pauwels, E., Stoven, V., Goto, S., and Yamanishi, Y. (2012). Relating drug–protein interaction network with drug side effects. *Bioinformatics* 28, i522–i528. doi: 10.1093/bioinformatics/bts383

Mutowo, P., Bento, A. P., Dedman, N., Gaulton, A., Hersey, A., Lomax, J., et al. (2016). A drug target slim: using gene ontology and gene ontology annotations to navigate protein-ligand target space in ChEMBL. *J. Biomed. Semantics* 7:59. doi: 10.1186/s13326-016-0102-0

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (Haifa). Available online at: https://web.cs.toronto.edu/

Newman, M. (2018). "Mathematics of networks," in *Networks* (Oxford: Oxford University Press). doi: 10.1093/oso/9780198805090.003.0006

Olayan, R. S., Ashoor, H., and Bajic, V. B. (2018). DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. *Bioinformatics* 34, 3779–3779. doi: 10.1093/bioinformatics/bty417

Öztürk, H., Özgür, A., and Ozkirimli, E. (2018). DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics* 34, i821–i829. doi: 10.1093/bioinformatics/bty593

Öztürk, H., Ozkirimli, E., and Özgür, A. (2019). WideDTA: prediction of drug-target binding affinity. *arXiv:1902.04166*. Available online at: https://arxiv.org/abs/1902.04166 (accessed July 15, 2019).

Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szwajda, A., Tang, J., et al. (2015). Toward more realistic drug-target interaction predictions. *Brief. Bioinformatics.* 16, 325–337. doi: 10.1093/bib/bbu010

Pahikkala, T., Okser, S., Airola, A., Salakoski, T., and Aittokallio, T. (2012a). Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations. *Algorithms Mol. Biol.* 7:11. doi: 10.1186/1748-7188-7-11

Pahikkala, T., Suominen, H., and Boberg, J. (2012b). Efficient cross-validation for kernelized least-squares regression with sparse basis expansions. *Mach. Learn.* 87, 381–407. doi: 10.1007/s10994-012-5287-6

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "DeepWalk: online learning of social representations," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY: ACM). doi: 10.1145/2623330.2623732

Puvanendrampillai, D., and Mitchell, J. B. O. (2003). L/D protein ligand database (PLD): additional understanding of the nature and specificity of protein-ligand complexes. *Bioinformatics* 19, 1856–1857. doi: 10.1093/bioinformatics/btg243

Raies, A. B., and Bajic, V. B. (2016). *In silico* toxicology: computational methods for the prediction of chemical toxicity. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 6, 147–172. doi: 10.1002/wcms.1240

Raies, A. B., and Bajic, V. B. (2018). *In silico* toxicology: comprehensive benchmarking of multi-label classification methods applied to chemical toxicity data. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 8:e1352. doi: 10.1002/wcms.1352

Raschka, S., Scott, A. M., Huertas, M., Li, W., and Kuhn, L. A. (2018). Automated inference of chemical discriminants of biological activity. *Methods Mol. Biol.* 1762, 307–338. doi: 10.1007/978-1-4939-7756-7_16

Rayhan, F., Ahmed, S., Farid, D. M., Dehzangi, A., and Shatabda, S. (2019). CFSBoost: cumulative feature subspace boosting for drug-target interaction prediction. *J. Theor. Biol.* 464, 1–8. doi: 10.1016/j.jtbi.2018.12.024

Rayhan, F., Ahmed, S., Shatabda, S., Farid, D. M., Mousavian, Z., Dehzangi, A., et al. (2017). iDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting. *Sci. Rep.* 7:17731. doi: 10.1038/s41598-017-18025-2

Roche, O., Kiyama, R., and Brooks, C. L. 3rd. (2001). Ligand-protein database: linking protein-ligand complex structures to binding data. *J. Med. Chem.* 44, 3592–3598. doi: 10.1021/jm000467k

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi: 10.1021/ci100050t

Salahudeen, M. S., and Nishtala, P. S. (2017). An overview of pharmacodynamic modelling, ligand-binding approach and its application in clinical practice. *Saudi Pharm. J.* 25, 165–175. doi: 10.1016/j.jsps.2016.07.002

Scarpino, A., Ferenczy, G. G., and Keseru, G. M. (2018). Comparative evaluation of covalent docking tools. *J. Chem. Inf. Model.* 58, 1441–1458. doi: 10.1021/acs.jcim.8b00228

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Scior, T., Medina-Franco, J. L., Do, Q. T., Martínez-Mayorga, K., Yunes Rojas, J. A., et al. (2009). How to recognize and workaround pitfalls in QSAR studies: a critical review. *Curr. Med. Chem.* 16, 4297–4313. doi: 10.2174/092986709789578213

Shen, M., and Huang, R. (2018). "A personal conversation assistant based on Seq2seq with Word2vec cognitive map," in *2018 7th International Congress on Advanced Applied Informatics (IIAI-AAI)* (Yonago: IEEE CPS). doi: 10.1109/IIAI-AAI.2018.00136

Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., et al. (2011). Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3:96ra77. doi: 10.1126/scitranslmed.3001318

Sledz, P., and Caflisch, A. (2018). Protein structure-based drug design: from docking to molecular dynamics. *Curr. Opin. Struct. Biol.* 48, 93–102. doi: 10.1016/j.sbi.2017.10.010

Smith, R. D., Clark, J. J., Ahmed, A., Orban, Z. J., Dunbar, J. B. Jr., and Carlson, H. A. (2019). Updates to binding MOAD (mother of all databases): polypharmacology tools and their utility in drug repurposing. *J. Mol. Biol.* 431, 2423–2433. doi: 10.1016/j.jmb.2019.05.024

Smith, R. D., Dunbar, J. B. Jr., Ung, P. M.-U., Esposito, E. X., Yang, C.-Y., Wang, S., et al. (2011). CSAR benchmark exercise of 2010: combined evaluation across all submitted scoring functions. *J. Chem. Inf. Model.* 51, 2115–2131. doi: 10.1021/ci200269q

Sotriffer, C., and Matter, H. (2011). The challenge of affinity prediction: scoring functions for structure-based virtual screening. *Methods Princ. Med. Chem.* 1, 177–221. doi: 10.1002/9783527633326.ch7

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Stefan, M. I., and Le Novère, N. (2013). Cooperative binding. *PLoS Comput. Biol.* 9:e1003106. doi: 10.1371/journal.pcbi.1003106

Sun, H., Duan, L., Chen, F., Liu, H., Wang, Z., Pan, P., et al. (2018). Assessing the performance of MM/PBSA and MM/GBSA methods. 7. Entropy effects on the performance of end-point binding free energy calculation approaches. *Phys. Chem. Chem. Phys.* 20, 14450–14460. doi: 10.1039/C7CP07623A

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). "LINE: large-scale information network embedding," in *Proceedings of the 24th International Conference on World Wide Web* (Florence: International World Wide Web Conferences Steering Committee). doi: 10.1145/2736277.2741093

Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., et al. (2014). Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J. Chem. Inf. Model.* 54, 735–743. doi: 10.1021/ci400709d

Tang, Z., Roberts, C. C., and Chia-En, A. C. (2017). Understanding ligand-receptor non-covalent binding kinetics using molecular modeling. *Front. Biosci.* 22:960. doi: 10.2741/4527

Tao, C., Sun, J., Jim Zheng, W., Chen, J., and Xu, H. (2015). Colorectal cancer drug target prediction using ontology-based inference and network analysis. *Database*. 2015:bav015. doi: 10.1093/database/bav015

Tatar, G., and Taskin Tok, T. (2019). Structure prediction of eukaryotic elongation factor-2 kinase and identification of the binding mechanisms of its inhibitors: homology modeling, molecular docking and molecular dynamics simulation. *J. Biomol. Struct. Dyn.* 18, 1–16. doi: 10.1080/07391102.2019.1592024

Trosset, J.-Y., and Cavé, C. (2019). "*In silico* drug–target profiling," in *Target Identification and Validation in Drug Discovery: Methods and Protocols*. eds J. Moll and S. Carotta (New York, NY: Springer), 89–103.

Tsubaki, M., Tomii, K., and Sese, J. (2019). Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* 35, 309–318. doi: 10.1093/bioinformatics/bty535

Vakil, V., and Trappe, W. (2019). Drug combinations: mathematical modeling and networking methods. *Pharmaceutics* 11:e208. doi: 10.3390/pharmaceutics11050208

Vallone, A., D'alessandro, S., Brogi, S., Brindisi, M., Chemi, G., Alfano, G., et al. (2018). Antimalarial agents against both sexual and asexual parasites stages: structure-activity relationships and biological studies of the Malaria Box compound 1-[5-(4-bromo-2-chlorophenyl) furan-2-yl]-N-[(piperidin-4-yl) methyl] methanamine (MMV019918) and analogues. *Eur. J. Med. Chem.* 150, 698–718. doi: 10.1016/j.ejmech.2018.03.024

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 18, 463–477. doi: 10.1038/s41573-019-0024-5

Van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

Vert, J. P., and Jacob, L. (2008). Machine learning for *in silico* virtual screening and chemical genomics: new strategies. *Comb. Chem. High Throughput Screen* 11, 677–685. doi: 10.2174/138620708785739899

Wackerly, D., Mendenhall, W., and Scheaffer, R. L. (2014). *Mathematical Statistics With Applications*. Belmont, NC: BROOKS/COLE Cengage Learning.

Wan, F., Hong, L., Xiao, A., Jiang, T., and Zeng, J. (2019). NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics* 35, 104–111. doi: 10.1093/bioinformatics/bty543

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi: 10.1038/nmeth.2810

Wang, F., Yang, W., and Hu, X. (2019). Discovery of high affinity receptors for dityrosine through inverse virtual screening and docking and molecular dynamics. *Int. J. Mol. Sci.* 20:115. doi: 10.3390/ijms20010115

Wang, K., Sun, J., Zhou, S., Wan, C., Qin, S., Li, C., et al. (2013). Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity. *PLoS Comput. Biol.* 9:e1003315. doi: 10.1371/annotation/958d4c23-4f1e-4579-b6ef-8ae1f828b1dd

Wang, R., Fang, X., Lu, Y., and Wang, S. (2004). The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *J. Med. Chem.* 47, 2977–2980. doi: 10.1021/jm030580l

Wang, R., Fang, X., Lu, Y., Yang, C.-Y., and Wang, S. (2005). The PDBbind database: methodologies and updates. *J. Med. Chem.* 48, 4111–4119. doi: 10.1021/jm048957q

Wang, W., Yang, S., Zhang, X., and Li, J. (2014). Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* 30, 2923–2930. doi: 10.1093/bioinformatics/btu403

Weiland, G. A., and Molinoff, P. B. (1981). Quantitative analysis of drug-receptor interactions: I. Determination of kinetic and equilibrium properties. *Life Sci.* 29, 313–330. doi: 10.1016/0024-3205(81)90324-6

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* 28, 31–36. doi: 10.1021/ci00057a005

Weininger, D. (1990). SMILES. 3. DEPICT. Graphical depiction of chemical structures. *J. Chem. Inf. Model.* 30, 237–243. doi: 10.1021/ci00067a005

Westbrook, J., Feng, Z., Chen, L., Yang, H., and Berman, H. M. (2003). The Protein Data Bank and structural genomics. *Nucleic Acids Res.* 31, 489–491. doi: 10.1093/nar/gkg068

Willmott, C. J., Matsuura, K., and Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. *Atmos. Environ.* 43, 749–752. doi: 10.1016/j.atmosenv.2008.10.005

Wu, B., Liu, Z., Wang, S., Hu, B.-G., and Ji, Q. (2014). "Multi-label learning with missing labels," in *2014 22nd International Conference on Pattern Recognition*, 1964–1968. doi: 10.1109/ICPR.2014.343

Wu, H. (2009). Global stability analysis of a general class of discontinuous neural networks with linear growth activation functions. *Inf. Sci.* 179, 3432–3441. doi: 10.1016/j.ins.2009.06.006

Xu, L., Wang, Z., Shen, Z., Wang, Y., and Chen, E. (2014). "Learning low-rank label correlations for multi-label classification with missing labels," in *2014 IEEE International Conference on Data Mining*, 1067–1072. doi: 10.1109/ICDM.2014.125

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. (2008). Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24, i232–i240. doi: 10.1093/bioinformatics/btn162

Yu, H.-F., Jain, P., Kar, P., and Dhillon, I. (2014). "Large-scale multi-label learning with missing labels," in *International Conference on Machine Learning*, 593–601.

Zhu, S., Bing, J., Min, X., Lin, C., and Zeng, X. (2018). Prediction of drug-gene interaction by using Metapath2vec. *Front. Genet.* 9:248. doi: 10.3389/fgene.2018.00248

Zhu, S., Okuno, Y., Tsujimoto, G., and Mamitsuka, H. (2005). A probabilistic model for mining implicit 'chemical compound-gene' relations from literature. *Bioinformatics* 21, ii245–ii251. doi: 10.1093/bioinformatics/bti1141

Zong, N., Kim, H., Ngo, V., and Harismendy, O. (2017). Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations. *Bioinformatics* 33, 2337–2344. doi: 10.1093/bioinformatics/btx160

Zong, N., Wong, R. S. N., and Ngo, V. (2019). Tripartite network-based repurposing method using deep learning to compute similarities for drug-target prediction. *Methods Mol. Biol.* 1903, 317–328. doi: 10.1007/978-1-4939-8955-3_19

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Thafar, Raies, Albaradei, Essack and Bajic. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 21 November 2019
doi: 10.3389/fchem.2019.00773

Integrative Multi-Kinase Approach for the Identification of Potent Antiplasmodial Hits

*Marilia N. N. Lima[1], Gustavo C. Cassiano[2], Kaira C. P. Tomaz[2], Arthur C. Silva[1], Bruna K. P. Sousa[1], Leticia T. Ferreira[2], Tatyana A. Tavella[2], Juliana Calit[3], Daniel Y. Bargieri[3], Bruno J. Neves[4], Fabio T. M. Costa[2] and Carolina Horta Andrade[1,2]**

[1] LabMol—Laboratory for Molecular Modeling and Drug Design, Faculty of Pharmacy, Federal University of Goiás, Goiânia, Brazil, [2] Laboratory of Tropical Diseases—Prof. Dr. Luiz Jacintho da Silva, Department of Genetics, Evolution, Microbiology and Immunology, Institute of Biology, University of Campinas (UNICAMP), Campinas, Brazil, [3] Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil, [4] Laboratory of Cheminformatics, University Center of Anápolis/UniEVANGELICA, Anápolis, Brazil

Malaria is a tropical infectious disease that affects over 219 million people worldwide. Due to the constant emergence of parasitic resistance to the current antimalarial drugs, the discovery of new antimalarial drugs is a global health priority. Multi-target drug discovery is a promising and innovative strategy for drug discovery and it is currently regarded as one of the best strategies to face drug resistance. Aiming to identify new multi-target antimalarial drug candidates, we developed an integrative computational approach to select multi-kinase inhibitors for *Plasmodium falciparum* calcium-dependent protein kinases 1 and 4 (CDPK1 and CDPK4) and protein kinase 6 (PK6). For this purpose, we developed and validated shape-based and machine learning models to prioritize compounds for experimental evaluation. Then, we applied the best models for virtual screening of a large commercial database of drug-like molecules. Ten computational hits were experimentally evaluated against asexual blood stages of both sensitive and multi-drug resistant *P. falciparum* strains. Among them, LabMol-171, LabMol-172, and LabMol-181 showed potent antiplasmodial activity at nanomolar concentrations ($EC_{50} \leq 700$ nM) and selectivity indices >15 folds. In addition, LabMol-171 and LabMol-181 showed good *in vitro* inhibition of *P. berghei* ookinete formation and therefore represent promising transmission-blocking scaffolds. Finally, docking studies with protein kinases CDPK1, CDPK4, and PK6 showed structural insights for further hit-to-lead optimization studies.

Keywords: malaria, shape-based, machine learning, virtual screening, *Plasmodium falciparum*, multi-target

OPEN ACCESS

Edited by:
Jose L. Medina-Franco, National Autonomous University of Mexico, Mexico

Reviewed by:
Marco A. Loza-Mejía, Universidad La Salle, Mexico
Antonio Romo-Mancillas, Universidad Autónoma de Querétaro, Mexico

***Correspondence:**
Carolina Horta Andrade
carolina@ufg.br

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *16 August 2019*
Accepted: *25 October 2019*
Published: *21 November 2019*

Citation:
Lima MNN, Cassiano GC, Tomaz KCP, Silva AC, Sousa BKP, Ferreira LT, Tavella TA, Calit J, Bargieri DY, Neves BJ, Costa FTM and Andrade CH (2019) Integrative Multi-Kinase Approach for the Identification of Potent Antiplasmodial Hits. Front. Chem. 7:773. doi: 10.3389/fchem.2019.00773

INTRODUCTION

Malaria is a serious infectious disease that affects 219 million people worldwide and kills over 435,000 patients annually, especially pregnant women and children in Sub-Saharan Africa (WHO, 2018). The disease is transmitted to humans through the bites of infected female *Anopheles* mosquitoes and caused by *Plasmodium* genus parasites (Ashley et al., 2018). Among them, *P. falciparum* is the most devastating species responsible for severe form of malaria and deaths (WHO, 2017).

Current control and eradication demands a combination of drugs with different mechanisms of action. Despite of compelling investment for controlling and eliminating this infectious disease, resistant parasite strains have been reported to all major antimalarial drugs (Wu et al., 1996; Triglia et al., 1998; Srivastava et al., 1999; Wellems and Plowe, 2001), including front-line artemisinin-based combination therapies (Rogers et al., 2009; Witkowski et al., 2013; Ashley et al., 2014). All these aspects highlight the urgent need for the discovery of new antimalarial drugs by identifying molecules with novel mechanisms of action and efficient against resistant parasite strains (Burrows et al., 2017).

The complete genome sequencing of *P. falciparum* (Gardner et al., 2002) has provided new and valuable information on its biological pathways, identifying potentially relevant biological targets for therapeutic intervention. In this context, protein kinases have been investigated because of their importance in several essential signaling pathways, e.g., homeostasis, apoptosis and cell division (Lucet et al., 2012; Bullard et al., 2013). Kinases catalyze the transfer of phosphate groups from ATP to specific substrates. These enzymes share a high degree of sequence and structural homology between the ATP binding sites, making them potential targets to be grouped and inhibited simultaneously by a single molecule. This mechanism, known as multi-kinase inhibition (MKI), provides a synergistic effect responsible for increasing the effectiveness of the kinase inhibitors, and consequently preventing the emergence of parasite resistance (Garuti et al., 2015). On the other hand, promiscuity is the main challenge in parasitic MKI design, which requires selective inhibitors unable to interact with host protein (Davies et al., 2000; Bain et al., 2003, 2007). However, the vast phylogenetic distance between Apicomplexans and humans (Ward et al., 2004) makes possible the development of multi-target and selective antimalarial candidates.

Calcium-Dependent Protein Kinases (CDPKs), a kinase family of plants and some alveolates, absent in metazoans, have been considered as one of the main effectors of calcium signaling, demonstrating a pronounced importance in apicomplexans, controlling a range of events in the parasite life cycle (Nagamune et al., 2008). *Pf*CDPK1 is expressed in all *Plasmodium* life stages (Sebastian et al., 2012), being essential for the sexual stage of the parasite (Jebiwott et al., 2013; Bansal et al., 2018). Meanwhile, *Pf*CDPK4 regulates cell cycle progression in the male gametocyte (Billker et al., 2004) and, together with Protein Kinase G, is activated during hepatocytes invasion by sporozoite (Govindasamy et al., 2016). Protein Kinase 6 of *P. falciparum* (*Pf*PK6), classified as Cyclin-Dependent Kinase (Chakrabarti et al., 1993), appears to be located in the cytoplasm and nucleus, mainly expressed in trophozoite, schizonts and segmenters stages (Bracchi-Ricard et al., 2000). The low identity between *Pf*PK6 and human Cyclin-Dependent Kinase 2 brings out PK6 as a potential antimalarial target. Its numerous variations in the active site amino acids can be exploited to design selective plasmodial inhibitors (Waters and Geyer, 2003). Therefore, the structural dissimilarities between human kinases and *Plasmodium*-specific kinases, such as CDPK1 and CDPK4 and PK6, turn these enzymes attractive targets for development of new multi-target antimalarial therapies (Lucet et al., 2012; Crowther et al., 2016). Recently, Crowther and colleagues (Crowther et al., 2016) reported an experimental screening of ∼14,000 cell-active compounds against *Pf*CDPK1 and *Pf*CDPK4, mitogen-associated protein kinase 2, PK6, and protein kinase 7. They found potent inhibitors (IC_{50} <1 μM) for multiple kinases simultaneously, with low cytotoxicity to human, bypassing the challenging of MKI promiscuity. Thus, the availability of the whole dataset of compounds with data for kinase inhibition allowed us to generate and validate robust and predictive shape-based models, that were integrated with machine learning (ML) models for a virtual screening workflow aiming to prioritize compounds to be experimentally evaluated *in vitro* against asexual blood stages of both sensitive and multi-drug resistant *P. falciparum*, and against sexual stages of *P. berghei*, as well as in mammalian cells. This integrative analysis allowed us to identify new potential and selective antiplasmodial hits.

Abbreviations: MKI, multi-kinase inhibitors; CDPK1, calcium-dependent protein kinases 1; CDPK4, calcium-dependent protein kinases 4; PK6, protein kinase 6; ML, machine learning; ROC, receiver operating characteristic curve; AUC, area under the ROC curve, BEDROC, Boltzmann-enhanced discrimination of ROC; EF, enrichment factor; SE, sensitivity; SP, specificity; VS, virtual screening; CCR, correct classification rate; PPV, positive predictive value; NPV, negative predictive value; and SI, selectivity index.

MATERIALS AND METHODS

The overall study design is shown in **Figure 1**. Briefly, we followed the successive steps: (I) dataset collection, curation, and integration of compounds with activity against CDPK1, CDPK4, PK6, and asexual-blood stages of *P. falciparum*; (II) development of shape-based models for CDPK1, CDPK4, and PK6, and machine learning models for *P. falciparum*; (III) virtual screening of ChemBridge database (∼1 million compounds); and (IV) experimental validation of prioritized compounds against asexual blood stage of *P. falciparum* (sensitive and multi-drug resistant strains), sexual blood stages of *P. berghei* and cytotoxicity in mammalian cells.

Computational

The whole project was built envisioning best practices of ML modeling (Tropsha, 2010; Cherkasov et al., 2014).

Data Integration and Curation

In this study, five datasets extracted from the PubChem Bioassay database (Wang et al., 2012) were explored to build shape-based models and ML models. All datasets were carefully standardized according to the protocol described by Fourches et al. (2010, 2015, 2016). Thus, explicit hydrogens were added; counter ions, inorganic salts, polymers, mixtures, and organometallic compounds were removed; and specific chemotypes (aromatic, nitro groups and others) were normalized using ChemAxon Standardizer (v. 6.1, ChemAxon, Budapest, Hungary, www.chemaxon.com). Then, a go/no-go criteria of 1 μM for the progression of *P. falciparum* kinase inhibitors and antiplasmodial hits (Katsuno et al., 2015) was used as activity threshold to distinguish active vs. inactive compounds. Furthermore, we

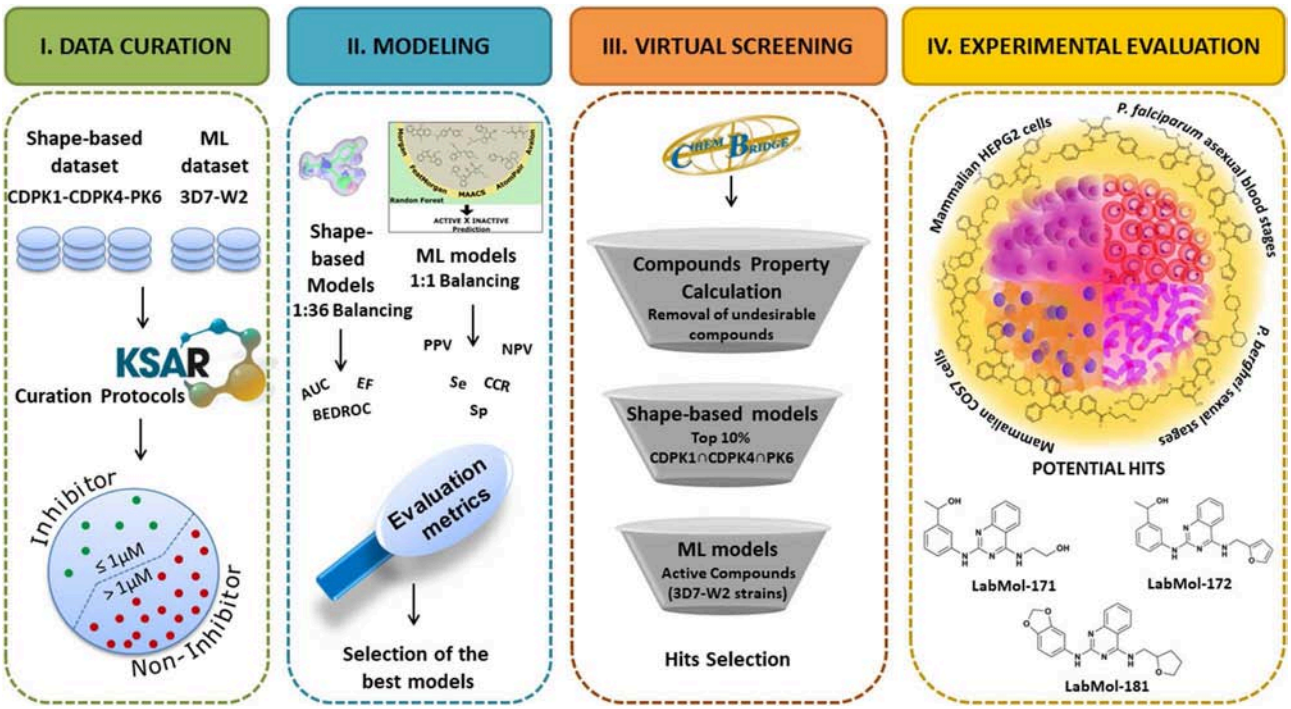


FIGURE 1 | Overall pipeline designed for the identification of new antiplasmodial compounds.

performed the analysis and exclusion of duplicates as follows: (a) if the reported outcomes of the duplicates were the same, one entry was retained in the dataset and the other excluded, and (b) if duplicates presented discordance in biological activity, both entries were excluded from dataset. A brief description of the datasets is presented below.

- CDPK1: 181 active compounds with IC_{50} $\leq 1\,\mu M$ and 13,270 inactive compounds with IC_{50} $>1\,\mu M$ (National Center for Biotechnology Information, 2016a);
- CDPK4: 55 active compounds with IC_{50} $\leq 1\,\mu M$ and 13,396 inactive compounds with IC_{50} $>1\,\mu M$ (National Center for Biotechnology Information, 2016b);
- PK6: 65 active compounds with IC_{50} $\leq 1\,\mu M$ and 13,386 inactive compounds with IC_{50} $>1\,\mu M$ (National Center for Biotechnology Information, 2016c);
- *P. falciparum* 3D7 (drug-susceptible) strain: 3,497 active compounds with EC_{50} $\leq 1\,\mu M$ and 4,376 inactive compounds with EC_{50} $>1\,\mu M$ (National Center for Biotechnology Information, 2009a, 2010a, 2011a, 2013);
- *P. falciparum* W2 (drug-resistant) strain: 3,637 active compounds with EC_{50} $\leq 1\,\mu M$ and 3,766 inactive compounds with EC_{50} $>1\,\mu M$ (National Center for Biotechnology Information, 2009b, 2010b, 2011b, 2012).

The inhibitory activity against each kinase was considered proportional to ATP consumed, as determined from measurements of residual [ATP] with the luciferase-based assay. So, all active compounds used to build shape-based models are inhibitors of ATP binding site (Crowther et al., 2016). All datasets generated for this study are included in the manuscript and the **Supplementary Files**.

Shape-Based Models

The shape-based models were built to distinguish active vs. inactive compounds for *P. falciparum* CDPK1, CDPK4, and PK6. Initially, the curated datasets were balanced by linear under-sampling method obeying a proportion of 1:36, aiming to reproduce the chemical space of an HTS, which contain more non-inhibitors. Then, 200 conformations were generated for each compound using OMEGA v.2.5.1.4 software (OMEGA 2.5.1.4: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com) (Hawkins et al., 2010), while the protonation states at neutral pH and AM1-BCC charges (Jakalian et al., 2002) were estimated using QUACPAC v.1.7.0.2 (QUACPAC 1.7.0.2: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com). To create the shape-based models, the most potent compounds against each kinase (see details in **Table S1**) were loaded into ROCS software v.3.2.2.2 (ROCS 3.2.2.2: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com) (Hawkins et al., 2007) and used as query compound. Then, the output conformations of active and inactive compounds were aligned by a solid-body optimization process that maximizes the overlap volume with queries, and ranked according to *Reference Tversky Combo* scoring function (Hawkins et al., 2007). Finally, the predictive performance of the shape-based models was assessed using the following metrics: Receiver Operating Characteristic (ROC) curve, Area Under the ROC Curve (AUC), Boltzmann-Enhanced Discrimination of

ROC (BEDROC) and Enrichment Factor (EF). These statistic metrics are calculated by the following equations:

$$AUC = \sum_i [(SE_{i+1})\,(SP_{i+1} - SP_i)] \quad (1)$$

$$BEDROC = RIE \times \frac{R_a sinh\left(\frac{\alpha}{2}\right)}{cosh\left(\frac{\alpha}{2}\right) - cosh\left(\frac{\alpha}{2} - \alpha R_a\right)} + \frac{1}{1 - e^{\alpha(1-R_a)}} \approx \frac{RIE}{\alpha} + \frac{1}{1 - e^{\alpha}}, if\ \alpha R_a \ll 1\ and\ \alpha \neq 0 \quad (2)$$

$$EF^{x\%} = \frac{Hits^{x\%}_{selected}/N^{x\%}_{selected}}{Hits_{total}/N_{total}} \quad (3)$$

Here, *SE* denotes sensitivity and *SP* specificity, RIE robust initial enhancement, R_a ratio of actives in the list.

The ROC curve provides a graphical representation of a predictor's behavior by plotting the true (Braga and Andrade, 2013; Neves et al., 2016) positive rate [sensitivity (SE)] against the 1 minus false positive rate [1—Specificity (SP)]. See SE e SP equations in ML section. The ideal predictive model would yield a point in the upper left corner of the ROC plot, representing 100% *SE* and *SP*. The AUC is the probability that a model will rank an active compound higher than a randomly chosen inactive. The EF shows how many times the shape-based models retrieved active compounds when compared with random selection (Braga and Andrade, 2013). Lastly, BEDROC uses an exponential decay function to favor models that pile up active compounds near the top of the rank-ordered list from the virtual screening (Huang and Wong, 2016).

Machine Learning Models

Binary ML models were built to distinguish active vs. inactive compounds for *P. falciparum*. The curated datasets for *P. falciparum* 3D7 and W2 strains were balanced in a proportion of 1 active:1 inactive. For this, the original chemical space of each library was maintained through linear under-sampling method based on k-nearest neighbors distances of each inactive to all active. ML models were built using an in-house workflow, implemented in KNIME (Berthold et al., 2009) including many modules as multiple machine learning methods, performance metrics, applicability domain, and Y-randomization test. Five molecular fingerprints implemented in RDKit (v.2.4.0) (http://www.rdkit.org) were used: (i) Morgan and (ii) FeatMorgan fingerprints, generated using radius of 2 and bit vector of 1,024 bits (Morgan, 1965; Rogers and Hahn, 2010); (iii) Molecular ACCess System (MACCS) structural keys (Dill et al., 1981; Anderson, 1984; Durant et al., 2002); (iv) AtomPair fingerprint with bit vector of 1,024 bits and path length ranging between 1 and 10 (Carhart et al., 1985); and (v) Avalon fingerprint with bit vector of 1,024 bits (Gedeck et al., 2006). The Random Forest method was the chosen algorithm to generate the models and to produce the final prediction based on combination of each decision tree (Breiman, 2001; Svetnik et al., 2003).

Moreover, for ML models' robustness estimation, 5-fold external cross-validation was performed. In this method, each dataset is randomly and equally divided into five subsets. Then, one of them is outwardly maintained as external set and the remaining four establish the modeling set. This procedure is repeated five times, allowing each subset to be used once as external validation set. The performance and robustness of ML models were assessed through statistic metrics such as: sensitivity (SE), specificity (SP), Correct Classification Rate (CCR), Positive Predictive Value (PPV), and Negative Predictive Value (NPV). These statistic metrics are calculated by the following equations:

$$SE = \frac{TP}{TP + FN} \quad (4)$$

$$SP = \frac{TN}{TN + FP} \quad (5)$$

$$CCR = \frac{SE + SP}{2} \quad (6)$$

$$PPV = \frac{TP}{TP + FP} \quad (7)$$

$$NPV = \frac{TN}{TN + FN} \quad (8)$$

Here, TP and TN correspond respectively to the number of true positives and true negatives. FP and FN represent, respectively, the number of false positives and false negatives.

In addition, 10 rounds of Y-randomization were conducted to evaluate whether the correlation between structure and activity occurred by chance. To measure the reliability of developed ML models, the Applicability Domain (AD) was estimated using Euclidean distances between each external compound, obtained by 5-fold cross-validation procedure, and their respective nearest neighbor in modeling set. These distances were related to the pre-defined AD threshold level. Toward a pre-defined distance threshold, the distance superior to this threshold were considered unreliable (Zhang et al., 2006).

In this study, we defined AD as:

$$D_T = \bar{y} + Z\sigma \quad (9)$$

Here, DT is a distance threshold, $\bar{y}$ is the average Euclidean distance of the k nearest neighbors of each compound of the training set, σ represents the standard deviation of the Euclidean distances and Z is an arbitrary parameter to control the level of significance. We set the default value of 0.5 for Z.

Consensus modeling was done combining the best ML models of each fingerprint type with Random Forest machine learning method. This approach was adopted with the aim to capture the different chemical information provided by each fingerprint, enriching the prediction during virtual screening and minimize individual model's error. Each individual model was applied to predict the activity of selected compounds after passing through shape-based screening filter. For this purpose, five models for 3D7 strain and five models for W2 strain were employed in separate runs. This way, when a model predicted a compound as active, a value of 0.2 was given, thus the final value of probability to be active was ranging from 0 to 1. Only compounds inside AD

and predicted as active at least in three models (probability ≥ 0.6) of both strains were picked up.

Virtual Screening

Developed shape-based and ML models were used for VS of ∼1.1 million compounds available on ChemBridge database (http://www.chembridge.com/) aiming to identify new potential kinases inhibitors with antiplasmodial activity. Prior to screening, the database was filtered using Veber (Veber et al., 2002) and Lipinski's rules (Lipinski et al., 2001) to prioritize drug-like molecules, using FILTER (OMEGA 2.5.1.4: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com). Subsequently, molecules were filtered by shape-based models developed for CDPK1, CDPK4, and PK6. Then, the common compounds between the top 10% of each kinase list had their antiplasmodial activity predicted by consensus ML models developed for 3D7 and W2 strains. The compounds prediction were recognized if it were found within the AD of more than 50% of all models used in consensus prediction. Finally, the selected virtual hits were purchased and submitted to *in vitro* experimental evaluation.

Homology Modeling

The amino acid sequence of *P. falciparum* CDPK1 and PK6 were not available on the Protein Data Bank at the time this work was conducted. Consequently, homology models were built by comparing the *P. falciparum* primary sequences with sequences of homolog proteins (templates) whose 3D structures were publicly available. Initially, the sequences of *P. falciparum* kinases were extracted from the UniProt database (Apweiler, 2004) and used as target for homology modeling in the SWISS-MODEL webserver (Bordoli et al., 2009; Biasini et al., 2014). Then, the tailored models were structurally optimized in GalaxyWEB server (Ko et al., 2012). Finally, overall stereochemical and geometrical quality of refined models were investigated using MolProbity server (Chen et al., 2010).

Docking

Chemical structures of antiplasmodial hits were imported to Maestro v. 10.7.015 (Schrödinger, LLC, New York, NY, 2016) and prepared using LigPrep (Schrödinger, LLC). In parallel, the 3D structures of *P. falciparum* CDPK1, CDPK4, and PK6 were prepared using the Protein Preparation Wizard available on Maestro workspace (Schrödinger LLC) as follows: bond orders and formal charges were adjusted; hydrogen atoms were added to the proteins; and protonation state of polar amino acids were predicted by PROPKA (Schrödinger, LLC) (Søndergaard et al., 2011) at neutral pHs. Before docking studies, grids were established to each protein ruled by a box space of $10 \times 10 \times 10$ Å^3, and fixing the box on the geometrical center of ATP-binding site using the receptor grid generation panel of the Glide (Schrödinger, LLC) (Friesner et al., 2004). Finally, molecular docking calculations were carried out using Glide Extra Precision (XP) mode and constraints into hinge region. The docking poses of each virtual hit were submitted to Prime (Schrödinger, LLC) for rescoring using the Molecular Mechanics/Generalized Born Surface Area (MMGBSA) approach with default conditions.

Experimental

Plasmodium Culture

Parasite cultures (3D7 and Dd2 strains) were maintained in O^+ human erythrocytes in RPMI 1640 medium supplemented with 0.05 mg/mL gentamycin, 38.4 mM HEPES, 0.2% sodium bicarbonate, and 10% O^+ human serum as described before (Trager and Jensen, 1976). To achieve a synchronic culture in the ring stage, two consecutive treatments at 48 h intervals with a 5% solution of D-sorbitol were done (Lambros and Vanderberg, 1979).

Determination of *Plasmodium* Growth Inhibition by SYBR Green I

Synchronized ring-stage (> 90%) infected erythrocytes were dispensed in duplicate into 96-well plates (0.5% parasitemia, 1% hematocrit) and incubated in dose response format with test compounds for 72 h. Chloroquine was used as an antimalarial control and uninfected erythrocytes as negative control. Then, *in vitro* susceptibility of parasite to tested drug was measured by SYBR Green (Hartwig et al., 2013). Following incubation, the plates were frozen and thawed, and 100 μL of the culture were transferred to a new black 96-well plate containing 100 μL of lysis buffer (20 mM Tris, 5 mM EDTA, 0.008% wt/vol saponin, 0.08% vol/vol Triton X-100, and 0.4 μL/mL of SYBR Green). After 1 h, the fluorescence was measured at 490 nm excitation and 540 nm emission (CLARIOstar, Labtech BMG). The results were compared with control cultures with no drugs. The EC_{50} was calculated by plotting the Log doses vs. Inhibition (expressed as a percentage relative to the control) in Prism 6 (GraphPad Software Inc.). Each test was performed at least three independent experiments.

Cytotoxicity Assay

The cytotoxicity was evaluated using two different lineages of mammalian cells: fibroblast-like cell lines derived from monkey kidney tissue (COS7 cells) and human hepatoma cell line (HEPG2). The cells were grown in 75 cm^2 flasks containing DMEM medium supplemented with 10% fetal bovine serum and 0.05 mg/mL gentamicin under a 5% CO_2 atmosphere at 37°C. After harvest of cells, 100 μL aliquots were distributed in 96-well plates (1×10^4 cells per well) and incubated until adhesion (∼12 h). The compounds at various concentrations (100−0.048 μM) were placed in the wells in duplicate and incubated for 72 h. The cell viability analysis were done by the MMT reduction method (3-[4,5-dimethyl-thiazol-2-yl]-2,5-diphenyltetrazolium chloride), after the incubation period (Mosmann, 1983). The optical density was determined at 570 nm (CLARIOstar, Labtech BMG) and the 50% cytotoxicity concentrations (CC_{50}) were expressed as the percent viability relative to the control (untreated cells). The selectivity index of the compounds was determined through the ratio of the CC_{50} of both cytotoxicity results (COS7 and HEPG2 cells) and EC_{50} 3D7, separately. Experiments were performed at least three times.

Inhibition of *P. berghei* Sexual Stage Progression

Balb/c mice were infected intraperitoneally with the *P. berghei* Ookluc line (Calit et al., 2018). Four to five days after infection,

TABLE 1 | Validation of shape-based models using different queries.

Kinase	Model	AUC	TOP 1%		TOP 5%		TOP 10%	
			EF	BEDROC	EF	BEDROC	EF	BEDROC
CDPK1	I*	0.81	22.10	0.63	10.06	0.50	5.41	0.52
	II	0.83	20.99	0.58	7.40	0.41	4.74	0.45
	III	0.69	0.55	0.02	2.20	0.09	2.21	0.16
CDPK4	IV*	0.77	3.64	0.13	3.64	0.17	3.27	0.24
	V	0.72	3.64	0.13	3.15	0.15	3.27	0.23
	VI	0.74	3.64	0.13	3.39	0.15	3.27	0.23
PK6	VII*	0.95	26.15	0.77	13.64	0.68	8.46	0.74
	VIII	0.94	26.15	0.75	13.54	0.66	8.62	0.72
	IX	0.93	24.62	0.65	13.54	0.64	8.31	0.72

*AUC, area under the ROC curve; EF, Enrichmenent Factor; BEDROC, Boltzmann-Enhanced Discrimination of ROC. *Selected model.*

the infected blood was collected by cardiac puncture and 4 μL were seeded to a volume of 80 μL of ookinete medium (Blagborough et al., 2013), at 21°C, containing 10 μM of compounds or DMSO vehicle control. After 24h incubation at 21°C, the nLuc substrate (Nano-Glo, Promega) was added to each well 1:1 (v:v) and incubated for 5 min at 37°C. The luciferase activity was measured using a plate luminometer SpectraMax i3; Molecular Devices and the % of conversion inhibition were calculated relative to the luciferase activity in the control assays. This assay was approved by the Ethics Committee (protocol number 132/2014-CEUA) of the Institute of Biomedical Sciences—University of São Paulo.

RESULTS AND DISCUSSION

Shape-Based Models

Shape-based models were built to distinguish active vs. inactive compounds for *P. falciparum* CDPK1, CDPK4, and PK6. Initially, the chemical structures of most potent inhibitors of each protein kinase were used as queries to develop shape-based models (**Table S1**). Molecular conformations of queries were selected according to energy minimization. Subsequently, the ability of the models to differentiate between the active and inactive compounds was inspected. Details of model performance are shown in **Table 1**. As observed, all models led to AUC values ranging between 0.69 and 0.95.

Model I showed the best statistical performance for CDPK1, with EF values of 22.10, 10.06, and 5.41; and BEDROC values of 0.63, 0.50, and 0.52 at the top 1, 5, and 10% of the ranked database, respectively. The model IV showed the best statistical performance for CDPK4, with EF values of 3.64, 3.64, and 3.27; and BEDROC values of 0.13, 0.17, and 0.24 at the top 1, 5, and 10% of the ranked database, respectively. Finally, the model VII showed the best statistical performance for PK6, with EF values of 26.15, 13.64, and 8.46; and BEDROC values of 0.77, 0.68, and 0.74 at the top 1, 5, and 10% of the ranked database, respectively. These results indicated that our shape-based models were statistically robust and therefore would be considered for a subsequent virtual screening study.

TABLE 2 | Summarized statistical characteristics of ML models accessed by 5-fold cross validation.

Model	CCR	SE	SP	PPV	NPV	Coverage
***P. falciparum* 3D7 strain**						
Avalon	0.75	0.72	0.78	0.76	0.73	0.99
MACCS	0.74	0.73	0.75	0.74	0.73	1.00
Morgan	0.75	0.69	0.80	0.78	0.72	0.99
FeatMorgan	0.75	0.71	0.79	0.77	0.73	0.99
AtomPair	0.73	0.70	0.77	0.75	0.72	0.99
Consensus	0.76	0.71	0.80	0.78	0.73	1.00
Consensus rigor	0.76	0.72	0.80	0.78	0.74	0.98
***P. falciparum* W2 strain**						
Avalon	0.71	0.67	0.75	0.73	0.70	0.99
MACCS	0.70	0.70	0.70	0.70	0.70	1.00
Morgan	0.71	0.66	0.76	0.73	0.69	0.99
FeatMorgan	0.71	0.68	0.74	0.72	0.70	0.99
AtomPair	0.70	0.69	0.71	0.70	0.69	0.99
Consensus	0.72	0.68	0.76	0.74	0.70	1.00
Consensus rigor	0.72	0.68	0.76	0.74	0.70	0.98

CCR, Correct Classification Rate; SE, Sensitivity; SP, Specificity; PPV, Positive Predictive Value; NPV, Negative Predictive Value.

ML Models

ML models were built to distinguish active vs. inactive compounds for *P. falciparum* sensitive (3D7) and resistant strains (W2). According to the statistical results of the 5-fold external cross-validation procedure, the combination of Avalon, MACCS, Morgan, FeatMorgan, AtomPair fingerprints with Random Forest algorithm led to predictive ML models, with CCR values ranging between 0.70 and 0.76. **Table 2** shows the detailed performances of the binary ML models.

The model built using Avalon (CCR = 0.75, SE = 0.72, SP = 0.78, PPV = 0.76, and NPV = 0.73) and Morgan (CCR = 0.75, SE = 0.69, SP = 0.80, PPV = 0.78, and NPV = 0.72) demonstrated the best performances among all other models

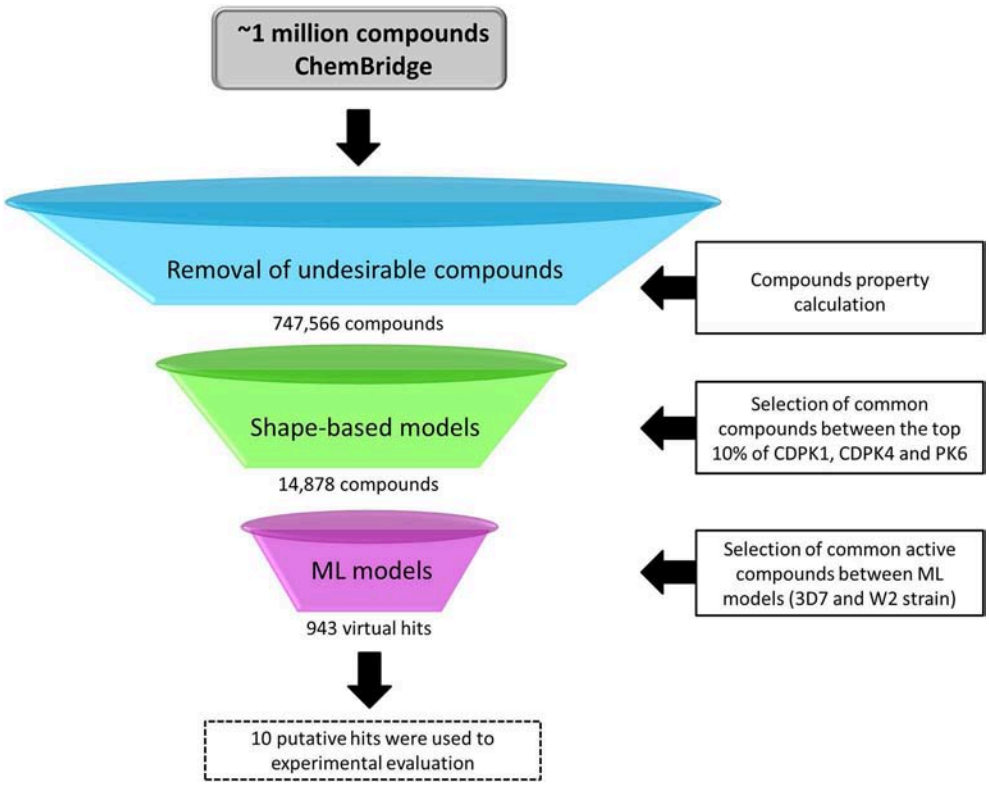


FIGURE 2 | General workflow of the virtual screening campaign.

developed for *P. falciparum* 3D7 strain. On the other hand, the best model developed for prediction activity against W2 strain was built using Avalon (CCR = 0.71, SE = 0.67, SP = 0.75, PPV = 0.73, NPV = 0.70), Morgan (CCR = 0.71, SE = 0.66, SP = 0.76, PPV = 0.73, NPV = 0.69), and FeatMorgan (CCR = 0.71, SE = 0.68, SP = 0.74, PPV = 0.72, NPV = 0.70). Subsequently, 10 rounds of Y-randomization were performed for each data set (**Table S2**). The results from this analysis (CCR, SE, SP values around 0.50) indicate that predictivity of our models was not due to chance correlation.

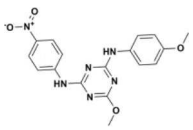
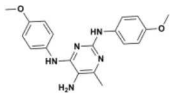
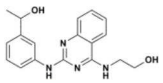
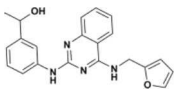
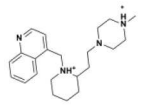
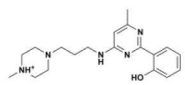
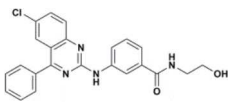
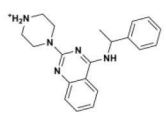
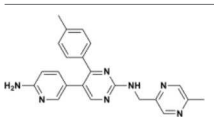
Virtual Screening

The virtual screening (VS) was carried out following the workflow presented in **Figure 2**. Initially, 1,091,088 compounds available on ChemBridge database were downloaded. Then, 747,566 molecules with probable oral bioavailability were prioritized using a drug-likeness filter. Then, conformers and AM1-BCC charges were generated for each molecule. The best shape-based models were used to prioritize potential *P. falciparum* multi-kinase inhibitors. Subsequently, the 14,878 common structures in top 10% scored list by shape-based filters were submitted to developed ML models for prediction of antiplasmodial activity against sensitive and resistant strains. In addition, the AD was determined in order to set "reliable" and "unreliable" predictions (Netzeva et al., 2005; Gadaleta et al., 2016). The predictions were considered reliable when the virtual hits are within the chemical space of compounds used to train ML models. At the end of this process, ten putative hits were selected for biological evaluation.

Experimental

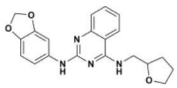
The ten virtual hits were evaluated *in vitro* against asexual blood stages of *P. falciparum* sensitive (3D7), and multi-drug-resistant (Dd2) strains. The EC_{50} for each compound (**Table 3**) indicate that three compounds (LabMol-171, LabMol-172 and LabMol-181) were potent at inhibiting the parasite growth showing activities in nanomolar range against both 3D7 and Dd2 strains. These results corroborate with go/no-go criteria established for the progression of *P. falciparum* kinase inhibitors and antiplasmodial hits in VS, since the three compounds showed $EC_{50} < 1\,\mu M$. The compound LabMol-181 (EC_{50} = 0.39 and 0.40 μM for 3D7 and Dd2, respectively) showed the most potent activity, when compared with reference drugs, chloroquine (EC_{50} = 0.02 and 0.15 μM for 3D7 and Dd2, respectively). Moreover, the three most active compounds (LabMol-171, LabMol-172 and LabMol-181) also have a common scaffold (quinazoline), varying groups at the R1 and R2 positions (**Figure 3**). In contrast, LabMol-175 (EC_{50} 3D7 > 5 μM) and LabMol-176 (EC_{50} 3D7 = 1.15 μM), which also display quinazoline scaffold, shown reduced inhibition activity against chloroquine-sensitive strain. This fact can be explained mainly by the presence of hydrophobic substituents in position R2 for both compounds, and an electron withdrawing group (Cl) attached to ring B in LabMol-175.

TABLE 3 | *In vitro* evaluation of selected virtual hits against asexual blood stage of *P. falciparum* 3D7 e Dd2 strains, cytotoxicity on mammalian cells (COS7, HEPG2), selectivity index and inhibition of ookinete formation of *P. berghei*.

Structure	EC_{50} 3D7 (μM)	EC_{50} Dd2 (μM)	CC_{50} COS7 (μM)	CC_{50} HEPG2 (μM)	SI*	SI**	% ookinete conversion inhibition (10 μM)
 LabMol-169	8.77 ± 2.20	3.40 ± 1.68	44.60 ± 3.73	19.95 ± 2.26	5.09	2.27	8.58 ± 7.62
 LabMol-170	1.14 ± 0.20	2.02 ± 0.36	7.59 ± 4.09	5.43 ± 2.39	6.66	4.76	11.24 ± 19.47
 LabMol-171	0.35 ± 0.08	0.70 ± 0.41	48.39 ± 14.04	19.31 ± 3.69	138.26	55.17	70.02 ± 22.16
 LabMol-172	0.43 ± 0.08	0.48 ± 0.17	12.11 ± 1.44	10.21 ± 5.53	28.16	23.74	8.59 ± 13.01
 LabMol-173	>5	–	–	–	–	–	0.00 ± 0.00
 LabMol-174	>5	–	–	–	–	–	21.41 ± 36.67
 LabMol-175	>5	–	–	–	–	–	12.87 ± 22.29
 LabMol-176	1.15 ± 0.26	1.71 ± 0.40	34.88 ± 4.06	5.63 ± 1.51	30.33	4.90	20.51 ± 19.45
 LabMol-177	>5	–	–	–	–	–	8.76 ± 15.18

(Continued)

TABLE 3 | Continued

Structure	EC_{50} 3D7 (μM)	EC_{50} Dd2 (μM)	CC_{50} COS7 (μM)	CC_{50} HEPG2 (μM)	SI*	SI**	% ookinete conversion inhibition (10 μM)
 LabMol-181	0.39 ± 0.06	0.40 ± 0.10	18.29 ± 1.92	6.13 ± 2.05	46.90	15.72	51.81 ± 23.16
Chloroquine	0.02 ± 0.01	0.15 ± 0.04	–	–	–	–	–

EC_{50} 3D7, half maximal effective concentration in 3D7 strain; EC_{50} Dd2, half maximal effective concentration in Dd2 strain; CC_{50} COS7, half maximal cytotoxic concentration on COS7 cell; CC_{50} HEPG2, half maximal cytotoxic concentration on HEPG2 cell; SI, selectivity index calculated between CC_{50} COS7 and EC_{50} 3D7 strain; SI**, selectivity index calculated between CC_{50} HEPG2 and EC_{50} 3D7 strain. The data are expressed as mean ± SD of three independent assays.*

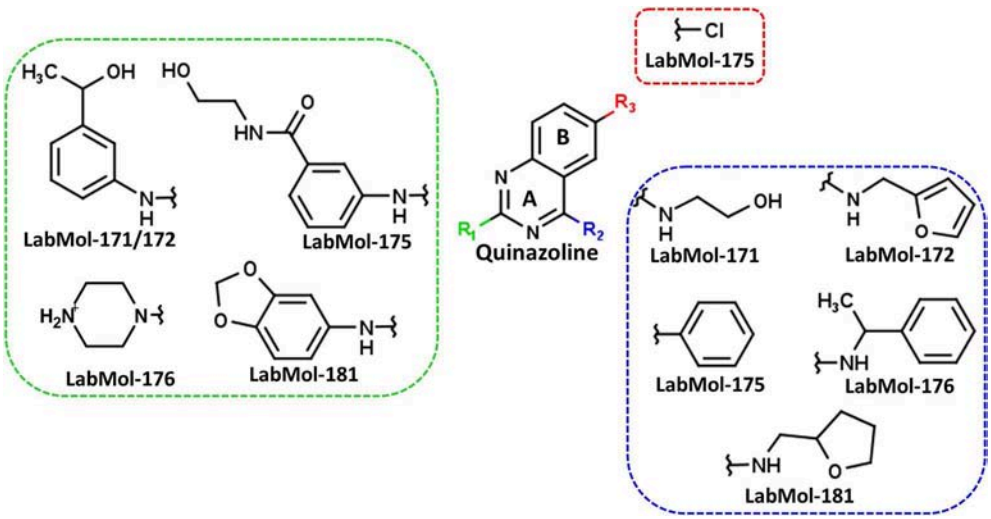


FIGURE 3 | Common quinazoline scaffold between the discovered hits. The dashed green, blue and red lines indicate the substituents in R1, R2, and R3 positions, respectively.

The selected compounds were also evaluated for their cytotoxicity against fibroblast-like cell lines derived from monkey kidney (COS-7 cells) and human hepatocytes (HEPG2 cells). With respect to selectivity, LabMol-171 and LabMol-172 showed the most promising results, since they showed selectivity index (SI) ranging between 23.74 and 138.26 (**Table 3**). It is worth noting that no compound showed cross-resistance with multi-drug resistant strain (Dd2 EC_{50}/3D7 $EC_{50} \leq 2$ for all compounds), thus suggesting a different mechanism of action from clinically established antimalarial drugs.

Previous reports have demonstrated that CDPK1 e CDPK4 have critical rule for parasite gametogenesis, displaying a potential target for development of transmission-blocking drugs (Billker et al., 2004; Bansal et al., 2018). Since CDPK1 and CDPK4 compose the present multi-target approach, we decided to evaluate the potential of these compounds in inhibiting the formation of ookinetes *in vitro*, using a recently described *in vitro* luciferase assay (Calit et al., 2018). LabMol-171 and LabMol-181, promising selected compounds in terms of selectivity and inhibition growing of asexual blood stages, also showed considerable inhibition at 10 μM (70.02 and 51.81%, respectively) during ookinete formation in comparison to control. These results demonstrate that these compounds are active against multiple parasite stages, comprising human treatment and transmission blocking to mosquitoes.

Rationalizing Anti-plasmodial Activity

Understanding the interaction pattern between the ligand and the protein target is essential for designing more potent and selective analogs. Here, molecular docking studies allowed us to rationalize the interaction the most potent hit with its associated protein targets.

As a crystal structure for docking execution was available only for *Pf*CDPK4 (PDB ID: 4QOX), the 3D structures of *Pf*CDPK1 and *Pf*PK6 were obtained by homology modeling process. The modeled and refined proteins were validated using MolProbity webserver (**Table S3**). This webserver encompass the metric clashscore (number of serious clashes per 1,000 atoms),

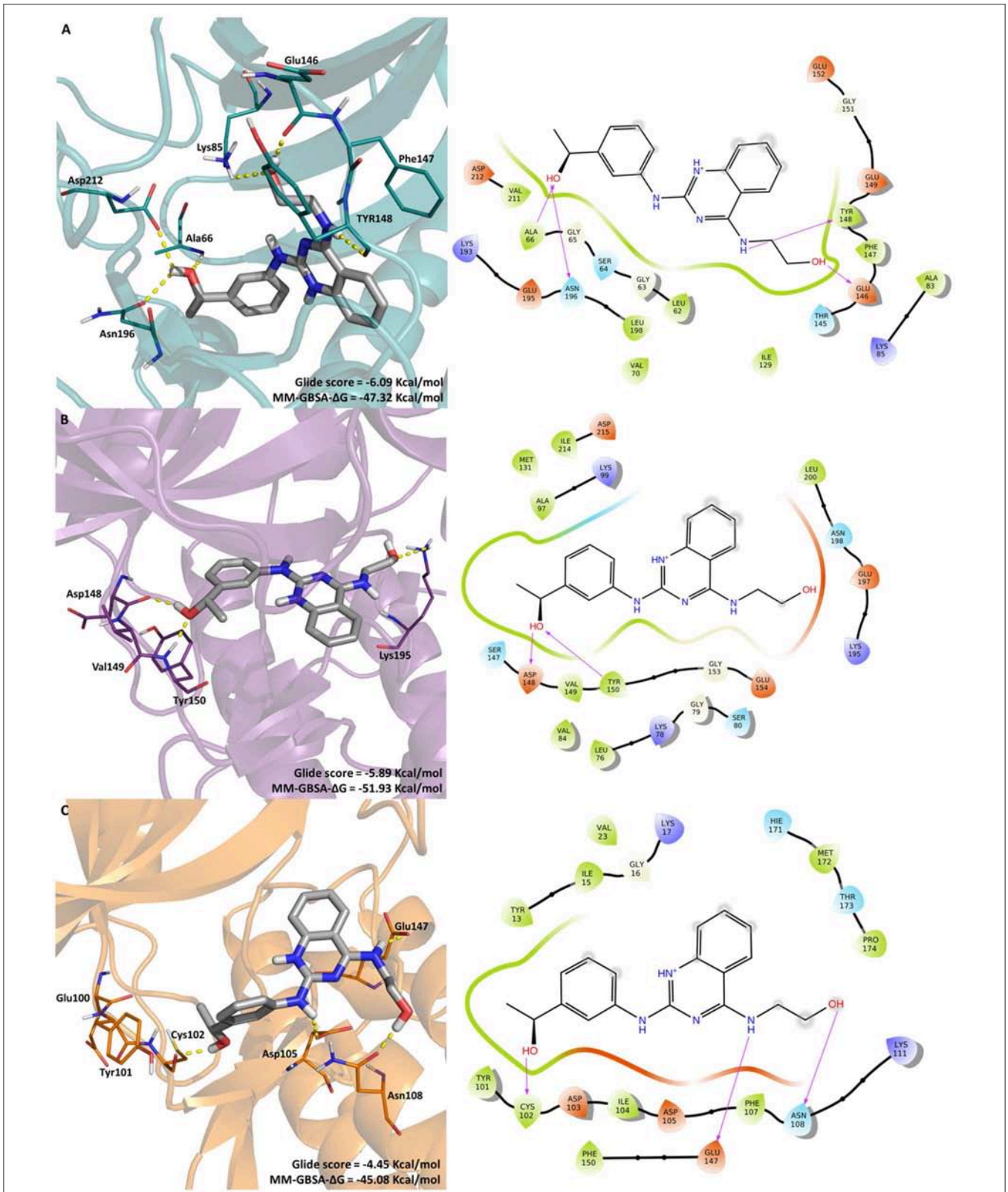


FIGURE 4 | Molecular interactions of LabMol-171 with **(A)** CDPK1(cyan), **(B)** CDPK4 (purple), and **(C)** PK6 (orange) obtained by docking. In 3D representation (left), hydrogen bonds are presented as yellow dashed lines, and the color code of oxygen, nitrogen and hydrogen atoms are red, blue, and white, respectively. The carbon atoms of LabMol-171 colored as gray. In 2D interaction diagrams (right) hydrogen bond are presented as magenta arrows.

which analyses steric overlap ≥0.4Å between non-bonded atoms that bring energy penalty; poor/favored rotamers, which evaluate the sidechain geometry conformation; outlier/favored Ramachandran, which evaluate protein backbone conformation by phi and psi backbone dihedrals; Molprobity score, which is represented by a number resulting from the combination of the clashscore, percentage of Ramanchandran not favored and percentage of bad side-chain rotamers, which reflects on a crystallographic resolution value; among others (Chen et al., 2010). After our investigation, we could conclude that clashscore and Molprobity score were within the desirable values, and 96.70–99.30% of the rotamers were in a favored state. Analyzes made for the values of Ramachandran pointed out that 97.25–98.30% of residues are within the favored region against 0.21–0.34% are classified as outliers. Thus, the overall stereochemistry and atoms conformation analysis displayed good quality of modeled kinases, approving them to use in docking studies.

The most promising compound, LabMol-171 (EC_{50} = 0.35 μM against 3D7 and SI = 138.26 on COS7 cell) was docked into the three protein kinases (*Pf*CDPK1, *Pf*CDPK4, and *Pf*PK6) to shed some light into the interaction pattern between the ligand and the proteins. A MM-GBSA calculation was performed in order to calculate the free energy of binding. **Figure 4** displays the interaction between the protein kinases and LabMol-171, the most promising experimental hit, Glide score and MMGBSA-ΔG values.

As we can see on **Figure 4**, the best free energy between LabMol-171 and protein kinases was obtained for calcium-dependent kinases. LabMol-171 could interact with CDPK4 (MM-GBSA-ΔG = −51.93 Kcal/mol) by hydrogen bonds at hinge region (Asp148, Tyr150) and the catalytic loop (Lys195). In relation to CDPK1 (MM-GBSA-ΔG = −47.32 Kcal/mol), hydrogen bonds were established with Lys85 and with residues belonging to the hinge region (Glu146, Tyr148), DFG motif (Asp212), catalytic loop (Asn196), and G-loop (Ala66). Aher and Roy (Aher and Roy, 2017) have showed the importance of some residues of CDPK1, including Val211, Tyr148, and Phe147 for *Pf*CDPK1 inhibitory activity.

For the docking results with PK6, a Cyclin-Dependent Kinase, LabMol-171 presented a lower Glide score (−4.45 Kcal/mol), showing high affinity with good values for free energy of binding (MM-GBSA-ΔG = −45.08 Kcal/mol). This kinase interacts with ligand in the hinge (Cys102) and catalytic loop regions (Glu147). Besides that, LabMol-171 was able to interact with Asp 105 and Asn108 of PK6.

Through our docking analysis, we could indicate that LabMol-171 could be a potential multi-kinase inhibitor, being able to interact mainly with hinge and catalytic loop region of these protein kinases. Besides that, previous studies have showed quinazoline scaffold inhibiting other molecular targets, as dihydrofolate reductase (Patel et al., 2019a,b) and prolyl-tRNA synthetase (Jain et al., 2017), besides kinases. So, prospective experimental target-fishing assays must be performed to understand the mechanism of action of quinazoline compounds in *Plasmodium*.

CONCLUSION

In this work, we developed robust and predictive shape-based and machine learning models, able to prioritize 10 promising hits as antimalarial candidates. Three compounds, LabMol-171, LabMol-172 and LabMol-181, reached activity in nanomolar concentration against *P. falciparum* strains, besides low cytotoxicity on mammalian cells. Moreover, these compounds did not show cross resistance with multi-drug resistant strain, suggesting a different mechanism of action. Besides that, LabMol-171 and LabMol-181 also showed considerable inhibition of ookinete formation in *P. berghei* standing out as powerful transmission blockers. Furthermore, a docking study shed some light into LabMol-171 interactions with CDPK1, CDPK4, and PK6 and suggests that this could be a potential MKI, being able to bind with hinge and catalytic loop regions of proposed kinases. In future studies, we aim to perform enzymatic assays against CDPK1, CDPK4 and PK6, and hit-to-lead optimization studies in order to reach new MKI antimalarial drugs, with transmission blocking activity.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

The animal study was reviewed and approved by Ethics Committee (protocol number 132/2014-CEUA) of the Institute of Biomedical Sciences—University of São Paulo.

AUTHOR CONTRIBUTIONS

Each author has contributed significantly to this work. ML contributed in the design, performing the computational experiments, and writing the paper. ML, BN, and CA conceived and designed the experiments. ML, AS, and BS performed the computational experiments. GC, KT, LF, TT, JC, DB, and FC performed the experimental assays. ML, AS, BS, GC, KT, LF, TT, JC, and BN analyzed the data. ML, BN, GC, and CA wrote the paper. All authors read, edited, and approved the final manuscript.

FUNDING

CA was supported by CNPq (grant 400760/2014-2) and FAPESP #2017/02353-9. FC was supported by FAPESP (Grants #2012/16525-2, #2017/18611-7, and 2018/07007-4). DB was supported by FAPESP (Grant #2013/13119-6), Instituto Serrapilheira (Grant #G-1709-16618), and CNPq (Grant 405996/2016-0). JC was supported by FAPESP (Fellowship #2018/24878-9). GC was supported by FAPESP (Fellowship 2015/20774-6). KT was supported by FAPESP (Fellowship 2018/05926-2). This study was financed in part

by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

ACKNOWLEDGMENTS

The authors thank Brazilian funding agencies, CNPq, CAPES, FAPESP, and FAPEG for financial support and fellowships. CA and FC are productivity fellows of CNPq. We are grateful to ChemAxon (https://chemaxon.com/) and OpenEye Scientific Software Inc. (https://www.eyesopen.com/) for providing academic license of their program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00773/full#supplementary-material

REFERENCES

Aher, R. B., and Roy, K. (2017). Exploring the structural requirements in multiple chemical scaffolds for the selective inhibition of *Plasmodium falciparum* calcium-dependent protein kinase-1 (PfCDPK-1) by 3D-pharmacophore modelling, and docking studies. *SAR QSAR Environ. Res.* 28, 390–414. doi: 10.1080/1062936X.2017.1326401

Anderson, S. (1984). Graphical representation of molecules and substructure-search queries in MACCStm. *J. Mol. Graph.* 2, 83–90. doi: 10.1016/0263-7855(84)80060-0

Apweiler, R. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* 32, 115D−119D. doi: 10.1093/nar/gkh131

Ashley, E. A., Dhorda, M., Fairhurst, R. M., Amaratunga, C., Lim, P., Suon, S., et al. (2014). Spread of artemisinin resistance in *Plasmodium falciparum* malaria. *N. Engl. J. Med.* 371, 411–423. doi: 10.1056/NEJMoa1314981

Ashley, E. A., Pyae Phyo, A., and Woodrow, C. J. (2018). Malaria. *Lancet* 391, 1608–1621. doi: 10.1016/S0140-6736(18)30324-6

Bain, J., McLauchlan, H., Elliott, M., and Cohen, P. (2003). The specificities of protein kinase inhibitors: an update. *Biochem. J.* 371, 199–204. doi: 10.1042/bj20021535

Bain, J., Plater, L., Elliott, M., Shpiro, N., Hastie, C. J., McLauchlan, H., et al. (2007). The selectivity of protein kinase inhibitors: a further update. *Biochem. J.* 408, 297–315. doi: 10.1042/BJ20070797

Bansal, A., Molina-Cruz, A., Brzostowski, J., Liu, P., Luo, Y., Gunalan, K., et al. (2018). PfCDPK1 is critical for malaria parasite gametogenesis and mosquito infection. *Proc. Natl. Acad. Sci. U.S.A.* 115, 774–779. doi: 10.1073/pnas.1715443115

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., et al. (2009). "KNIME - the Konstanz information miner," in *Data Analysis, Machine Learning and Applications* (Berlin; Heidelberg: Springer), 319–326. (SIGKDD Explor. Newsl. 11, 1 (November 2009), 26–31). doi: 10.1145/1656274.1656280

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., et al. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 42, W252–W258. doi: 10.1093/nar/gku340

Billker, O., Dechamps, S., Tewari, R., Wenig, G., Franke-Fayard, B., and Brinkmann, V. (2004). Calcium and a calcium-dependent protein kinase regulate gamete formation and mosquito transmission in a malaria parasite. *Cell* 117, 503–14. doi: 10.1016/S0092-8674(04)00449-0

Blagborough, A. M., Delves, M. J., Ramakrishnan, C., Lal, K., Butcher, G., and Sinden, R. E. (2013). Assessing transmission blockade in *Plasmodium* spp. *Methods Mol. Biol.* 923, 577–600. doi: 10.1007/978-1-62703-026-7_40

Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., and Schwede, T. (2009). Protein structure homology modeling using SWISS-MODEL workspace. *Nat. Protoc.* 4, 1–13. doi: 10.1038/nprot.2008.197

Bracchi-Ricard, V., Barik, S., Delvecchio, C., Doerig, C., Chakrabarti, R., and Chakrabarti, D. (2000). PfPK6, a novel cyclin-dependent kinase/mitogen-activated protein kinase-related protein kinase from *Plasmodium falciparum*. *Biochem. J.* 347(Pt 1), 255–63. doi: 10.1042/bj3470255

Braga, R. C., and Andrade, C. H. (2013). Assessing the performance of 3D pharmacophore models in virtual screening: how good are they? *Curr. Top. Med. Chem.* 13, 1127–1138. doi: 10.2174/15680266113130090010

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Bullard, K. M., DeLisle, R. K., and Keenan, S. M. (2013). Malarial kinases: novel targets for in silico approaches to drug discovery. *Methods Mol. Biol.* 993, 205–229. doi: 10.1007/978-1-62703-342-8_14

Burrows, J. N., Duparc, S., Gutteridge, W. E., Hooft van Huijsduijnen, R., Kaszubska, W., Macintyre, F., et al. (2017). New developments in anti-malarial target candidate and product profiles. *Malar. J.* 16, 26. doi: 10.1186/s12936-016-1675-x

Calit, J., Dobrescu, I., Gaitán, X. A., Borges, M. H., Ramos, M. S., Eastman, R. T., et al. (2018). Screening the pathogen box for molecules active against Plasmodium sexual stages using a new nanoluciferase-based transgenic line of *P. berghei* identifies transmission-blocking compounds. *Antimicrob. Agents Chemother.* 62:e01053-18. doi: 10.1128/AAC.01053-18

Carhart, R. E., Smith, D. H., and Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* 25, 64–73. doi: 10.1021/ci00046a002

Chakrabarti, D., Schuster, S. M., and Chakrabarti, R. (1993). Cloning and characterization of subunit genes of ribonucleotide reductase, a cell-cycle-regulated enzyme, from Plasmodium falciparum. *Proc. Natl. Acad. Sci. U.S.A.* 90, 12020–4. doi: 10.1073/pnas.90.24.12020

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., et al. (2010). MolProbity : all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. Sect. D Biol. Crystallogr.* 66, 12–21. doi: 10.1107/S0907444909042073

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., et al. (2014). QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* 57, 4977–5010. doi: 10.1021/jm4004285

Crowther, G. J., Hillesland, H. K., Keyloun, K. R., Reid, M. C., Lafuente-Monasterio, M. J., Ghidelli-Disse, S., et al. (2016). Biochemical screening of five protein kinases from plasmodium falciparum against 14,000 cell-active compounds. *PLoS ONE* 11:e0149996. doi: 10.1371/journal.pone.0149996

Davies, S. P., Reddy, H., Caivano, M., and Cohen, P. (2000). Specificity and mechanism of action of some commonly used protein kinase inhibitors. *Biochem. J.* 351, 95–105. doi: 10.1042/bj3510095

Dill, J. D., Hounshell, W. D., Marson, S., Peacock, S., and Wipke, W. T. (1981). "Search and retrieval using an automated molecular access system," in *182nd National Meeting of the American Chemical Society* (New York, NY), 23–28.

Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* 42, 1273–1280. doi: 10.1021/ci010132r

Fourches, D., Muratov, E., and Tropsha, A. (2010). Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* 50, 1189–204. doi: 10.1021/ci100176x

Fourches, D., Muratov, E., and Tropsha, A. (2015). Curation of chemogenomics data. *Nat. Chem. Biol.* 11, 535–535. doi: 10.1038/nchembio.1881

Fourches, D., Muratov, E., and Tropsha, A. (2016). *J. Chem. Inf. Model.* 56, 1243–1252. doi: 10.1021/acs.jcim.6b00129

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* 47, 1739–1749. doi: 10.1021/jm0306430

Gadaleta, D., Mangiatordi, G. F., Catto, M., Carotti, A., and Nicolotti, O. (2016). Applicability domain for QSAR models. *Int. J. Quant. Struct. Relationships* 1, 45–63. doi: 10.4018/IJQSPR.2016010102

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., et al. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419, 498–511. doi: 10.1038/nature01097

Garuti, L., Roberti, M., and Bottegoni, G. (2015). Multi-kinase inhibitors. *Curr. Med. Chem.* 22, 695–712. doi: 10.2174/0929867321666141216125528

Gedeck, P., Rohde, B., and Bartels, C. (2006). QSAR – how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J. Chem. Inf. Model.* 46, 1924–1936. doi: 10.1021/ci050413p

Govindasamy, K., Jebiwott, S., Jaijyan, D. K., Davidow, A., Ojo, K. K., Van Voorhis, W. C., et al. (2016). Invasion of hepatocytes by Plasmodium sporozoites requires cGMP-dependent protein kinase and calcium dependent protein kinase 4. *Mol. Microbiol.* 102, 349–363. doi: 10.1111/mmi.13466

Hartwig, C. L., Ahmed, A. O. A., Cooper, R. A., and Stedman, T. T. (2013). "SYBR Green I®-based parasite growth inhibition assay for measurement of antimalarial drug susceptibility in *Plasmodium falciparum*," in *Methods in Malaria Research*, eds K. Moll, A. Kaneko, A. Scherf, and M. Wahlgren (Glasgow: EVIMalaR), 122–129.

Hawkins, P. C. D., Skillman, A. G., and Nicholls, A. (2007). Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* 50, 74–82. doi: 10.1021/jm0603365

Hawkins, P. C. D., Skillman, A. G., Warren, G. L., Ellingson, B. A., and Stahl, M. T. (2010). Conformer generation with OMEGA: algorithm and validation using high quality structures from the protein databank and cambridge structural database. *J. Chem. Inf. Model.* 50, 572–584. doi: 10.1021/ci100031x

Huang, Z., and Wong, C. F. (2016). Inexpensive method for selecting receptor structures for virtual screening. *J. Chem. Inf. Model.* 56, 21–34. doi: 10.1021/acs.jcim.5b00299

Jain, V., Yogavel, M., Kikuchi, H., Oshima, Y., Hariguchi, N., Matsumoto, M., et al. (2017). Targeting prolyl-tRNA synthetase to accelerate drug discovery against malaria, leishmaniasis, toxoplasmosis, cryptosporidiosis, and coccidiosis. *Structure* 25, 1495–1505.e6. doi: 10.1016/j.str.2017.07.015

Jakalian, A., Jack, D. B., and Bayly, C. I. (2002). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* 23, 1623–1641. doi: 10.1002/jcc.10128

Jebiwott, S., Govindaswamy, K., Mbugua, A., and Bhanot, P. (2013). Plasmodium berghei calcium dependent protein kinase 1 is not required for host cell invasion. *PLoS ONE* 8:e79171. doi: 10.1371/journal.pone.0079171

Katsuno, K., Burrows, J. N., Duncan, K., van Huijsduijnen, R. H., Kaneko, T., Kita, K., et al. (2015). Hit and lead criteria in drug discovery for infectious diseases of the developing world. *Nat. Rev. Drug Discov.* 14, 751–758. doi: 10.1038/nrd4683

Ko, J., Park, H., Heo, L., and Seok, C. (2012). GalaxyWEB server for protein structure prediction and refinement. *Nucleic Acids Res.* 40, W294–W297. doi: 10.1093/nar/gks493

Lambros, C., and Vanderberg, J. P. (1979). Synchronization of *Plasmodium falciparum* erythrocytic stages in culture. *J. Parasitol.* 65:418. doi: 10.2307/3280287

Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 64, 3–26. doi: 10.1016/S0169-409X(00)00129-0

Lucet, I. S., Tobin, A., Drewry, D., Wilks, A. F., and Doerig, C. (2012). *Plasmodium* kinases as targets for new-generation antimalarials. *Future Med. Chem.* 4, 2295–2310. doi: 10.4155/fmc.12.183

Morgan, H. L. (1965). The generation of a unique machine description for chemical structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* 5, 107–113. doi: 10.1021/c160017a018

Mosmann, T. (1983). Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays. *J. Immunol. Methods* 65, 55–63. doi: 10.1016/0022-1759(83)90303-4

Nagamune, K., Moreno, S. N., Chini, E. N., and Sibley, L. D. (2008). Calcium regulation and signaling in apicomplexan parasites. *Subcell. Biochem.* 47, 70–81. doi: 10.1007/978-0-387-78267-6_5

National Center for Biotechnology Information (2009a). *PubChem BioAssay Database*. Available online at: https://pubchem.ncbi.nlm.nih.gov/bioassay/1828 (accessed November 27, 2018).

National Center for Biotechnology Information (2009b). *PubChem BioAssay Database*. Source=NCGC, AID=1883.

National Center for Biotechnology Information (2010a). *PubChem BioAssay Database*. Available online at: https://pubchem.ncbi.nlm.nih.gov/bioassay/449703 (accessed November, 27, 2018).

National Center for Biotechnology Information (2010b). *PubChem BioAssay Database*. Available online at: https://pubchem.ncbi.nlm.nih.gov/bioassay/449704 (accessed November 27, 2018).

National Center for Biotechnology Information (2011a). *PubChem BioAssay Database*. Available online at: https://pubchem.ncbi.nlm.nih.gov/bioassay/524790 (accessed November 27, 2018).

National Center for Biotechnology Information (2011b). *PubChem BioAssay Database*. Available online at: https://pubchem.ncbi.nlm.nih.gov/bioassay/524796 (accessed November 27, 2018).

National Center for Biotechnology Information (2012). *PubChem BioAssay Database*. Available online at: https://pubchem.ncbi.nlm.nih.gov/bioassay/606570 (accessed November 27, 2018).

National Center for Biotechnology Information (2013). *PubChem BioAssay Database*. Available online at: https://pubchem.ncbi.nlm.nih.gov/bioassay/660866 (accessed November 27, 2018).

National Center for Biotechnology Information (2016a). *PubChem BioAssay Database*. Available online at: https://pubchem.ncbi.nlm.nih.gov/bioassay/1159585 (accessed November 27, 2018).

National Center for Biotechnology Information (2016b). *PubChem BioAssay Database*. Available online at: https://pubchem.ncbi.nlm.nih.gov/bioassay/1159588 (accessed November 27, 2018).

National Center for Biotechnology Information (2016c). *PubChem BioAssay Database*. Available online at: https://pubchem.ncbi.nlm.nih.gov/bioassay/1159586 (accessed November 27, 2018).

Netzeva, T. I., Worth, A., Aldenberg, T., Benigni, R., Cronin, M. T. D., Gramatica, P., et al. (2005). Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern. Lab. Anim.* 33, 155–73. doi: 10.1177/026119290503300209

Neves, B. J. B. J., Muratov, E., Machado, R. B. R. B., Andrade, C. H. C. H., and Cravo, P. V. L. P. V. L. (2016). Modern approaches to accelerate discovery of new antischistosomal drugs. *Expert Opin. Drug Discov.* 11, 557–567. doi: 10.1080/17460441.2016.1178230

Patel, T. S., Bhatt, J. D., Dixit, R. B., Chudasama, C. J., Patel, B. D., and Dixit, B. C. (2019a). Design and synthesis of leucine-linked quinazoline-4(3H)-one-sulphonamide molecules distorting malarial reductase activity in the folate pathway. *Arch. Pharm.* 352:e1900099. doi: 10.1002/ardp.201900099

Patel, T. S., Bhatt, J. D., Dixit, R. B., Chudasama, C. J., Patel, B. D., and Dixit, B. C. (2019b). Green synthesis, biological evaluation, molecular docking studies and 3D-QSAR analysis of novel phenylalanine linked quinazoline-4(3H)-one-sulphonamide hybrid entities distorting the malarial reductase activity in folate pathway. *Bioorg. Med. Chem.* 27, 3574–3586. doi: 10.1016/j.bmc.2019.06.038

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi: 10.1021/ci100050t

Rogers, W. O., Sem, R., Tero, T., Chim, P., Lim, P., Muth, S., et al. (2009). Failure of artesunate-mefloquine combination therapy for uncomplicated *Plasmodium falciparum* malaria in southern Cambodia. *Malar. J.* 8:10. doi: 10.1186/1475-2875-8-10

Sebastian, S., Brochet, M., Collins, M. O., Schwach, F., Jones, M. L., Goulding, D., et al. (2012). A Plasmodium calcium-dependent protein kinase controls zygote development and transmission by translationally activating repressed mRNAs. *Cell Host Microbe* 12, 9–19. doi: 10.1016/j.chom.2012.05.014

Søndergaard, C. R., Olsson, M. H. M., Rostkowski, M., and Jensen, J. H. (2011). Improved treatment of ligands and coupling effects in empirical calculation and rationalization of p K a values. *J. Chem. Theory Comput.* 7, 2284–2295. doi: 10.1021/ct200133y

Srivastava, I. K., Morrisey, J. M., Darrouzet, E., Daldal, F., and Vaidya, A. B. (1999). Resistance mutations reveal the atovaquone-binding domain of cytochrome b in malaria parasites. *Mol. Microbiol.* 33, 704–711. doi: 10.1046/j.1365-2958.1999.01515.x

Svetnik, V., Liaw, A., Tong, C., Christopher Culberson, J., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for

compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 43, 1947–1958. doi: 10.1021/ci034160g

Trager, W., and Jensen, J. B. (1976). Human malaria parasites in continuous culture. *Science* 193, 673–5. doi: 10.1126/science.781840

Triglia, T., Wang, P., Sims, P. F. G., Hyde, J. E., and Cowman, A. F. (1998). Allelic exchange at the endogenous genomic locus in *Plasmodium falciparum* proves the role of dihydropteroate synthase in sulfadoxine-resistant malaria. *EMBO J.* 17, 3807–3815. doi: 10.1093/emboj/17.14.3807

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Mol. Inform.* 29, 476–488. doi: 10.1002/minf.201000061

Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W., and Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* 45, 2615–2623. doi: 10.1021/jm020017n

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., et al. (2012). PubChem's BioAssay database. *Nucleic Acids Res.* 40, D400–D412. doi: 10.1093/nar/gkr1132

Ward, P., Equinet, L., Packer, J., and Doerig, C. (2004). Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote. *BMC Genomics* 5:79. doi: 10.1186/1471-2164-5-79

Waters, N. C., and Geyer, J. A. (2003). Cyclin-dependent protein kinases as therapeutic drug targets for antimalarial drug development. *Expert Opin. Ther. Targets* 7, 7–17. doi: 10.1517/14728222.7.1.7

Wellems, T. E., and Plowe, C. V. (2001). Chloroquine-resistant malaria. *J. Infect. Dis.* 184, 770–776. doi: 10.1086/322858

WHO (2017). *World Malaria Report 2017. 196.* Genova.

WHO (2018). *World Malaria Report 2018.* Genova.

Witkowski, B., Khim, N., Chim, P., Kim, S., Ke, S., Kloeung, N., et al. (2013). Reduced artemisinin susceptibility of *Plasmodium falciparum* ring stages in Western Cambodia. *Antimicrob. Agents Chemother.* 57, 914–923. doi: 10.1128/AAC.01868-12

Wu, Y., Kirkman, L. A., and Wellems, T. E. (1996). Transformation of *Plasmodium falciparum* malaria parasites by homologous integration of plasmids that confer resistance to pyrimethamine. *Proc. Natl. Acad. Sci. U.S.A.* 93, 1130–1134. doi: 10.1073/pnas.93.3.1130

Zhang, S., Golbraikh, A., Oloff, S., Kohn, H., and Tropsha, A. (2006). A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *J. Chem. Inf. Model.* 46, 1984–1995. doi: 10.1021/ci060132x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Lima, Cassiano, Tomaz, Silva, Sousa, Ferreira, Tavella, Calit, Bargieri, Neves, Costa and Andrade. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 25 November 2019
doi: 10.3389/fchem.2019.00798

Structure-Based and Molecular Modeling Studies for the Discovery of Cyclic Imides as Reversible Cruzain Inhibitors With Potent Anti-*Trypanosoma cruzi* Activity

Rafael A. A. Ferreira[1†], Ivani Pauli[2†], Thiago S. Sampaio[1], Mariana L. de Souza[2], Leonardo L. G. Ferreira[2], Luma G. Magalhães[2], Celso de O. Rezende Jr.[1], Rafaela S. Ferreira[3], Renata Krogh[2], Luiz C. Dias[1] and Adriano D. Andricopulo[2*]*

[1] Instituto de Química, Universidade Estadual de Campinas, Campinas, Brazil, [2] Laboratório de Química Medicinal e Computacional, Centro de Pesquisa e Inovação em Biodiversidade e Fármacos, Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, Brazil, [3] Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

OPEN ACCESS

Edited by:
Jose L. Medina-Franco, National Autonomous University of Mexico, Mexico

Reviewed by:
Dario Ariel Estrin, University of Buenos Aires, Argentina
Gildardo Rivera, National Polytechnic Institute, Mexico

***Correspondence:**
Luiz C. Dias
ldias@unicamp.br
Adriano D. Andricopulo
aandrico@ifsc.usp.br

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: 01 October 2019
Accepted: 05 November 2019
Published: 25 November 2019

Citation:
Ferreira RAA, Pauli I, Sampaio TS, de Souza ML, Ferreira LLG, Magalhães LG, Rezende CdO Jr, Ferreira RS, Krogh R, Dias LC and Andricopulo AD (2019) Structure-Based and Molecular Modeling Studies for the Discovery of Cyclic Imides as Reversible Cruzain Inhibitors With Potent Anti-Trypanosoma cruzi Activity. Front. Chem. 7:798. doi: 10.3389/fchem.2019.00798

Chagas disease causes ∼10,000 deaths each year, mainly in Latin America, where it is endemic. The currently available chemotherapeutic agents are ineffective in the chronic stage of the disease, and the lack of pharmaceutical innovation for Chagas disease highlights the urgent need for the development of new drugs. The enzyme cruzain, the main cysteine protease of *Trypanosoma cruzi*, has been explored as a validated molecular target for drug discovery. Herein, the design, molecular modeling studies, synthesis, and biological evaluation of cyclic imides as cruzain inhibitors are described. Starting with a micromolar-range cruzain inhibitor (**3a**, $IC_{50} = 2.2\,\mu M$), this molecular optimization strategy resulted in the nanomolar-range inhibitor **10j** ($IC_{50} = 0.6\,\mu M$), which is highly active against *T. cruzi* intracellular amastigotes ($IC_{50} = 1.0\,\mu M$). Moreover, most compounds were selective toward *T. cruzi* over human fibroblasts, which were used as host cells, and are less toxic to hepatic cells than the marketed drug benznidazole. This study enabled the discovery of novel chemical diversity and established robust structure-activity relationships to guide the design of optimized cruzain inhibitors as new trypanocidal agents.

Keywords: Chagas disease, *Trypanosoma cruzi*, cruzain, SAR, medicinal chemistry, synthesis, inhibitors, molecular docking

INTRODUCTION

Caused by the protozoan *Trypanosoma cruzi* and endemic in 21 countries in Latin America, Chagas disease kills ∼10,000 people each year[1]. This neglected tropical disease has reached non-endemic regions, affecting 8 million people worldwide and putting another 25 million at risk of infection[1]. The USA shows the greatest burden among non-endemic countries, with ∼300,000 people estimated to be infected with *T. cruzi* (Pérez-Molina and Molina, 2018). In addition, Chagas disease is a major cause of infectious cardiomyopathy worldwide, contributing substantially to the global burden of cardiovascular disease (Bern, 2015; Cucunubá et al., 2016). Mortality and a

[1]https://www.who.int/chagas/en/

reduction in productivity of the affected populations significantly impact the economies of the endemic regions. These economic and social burdens can be translated in numbers that estimate losses of more than US $7.2 billion per year and ∼243,600 disability-adjusted life years (DALYs) due to Chagas disease (GBD DALYs and HALE Collaborators, 2016; Arnal et al., 2019).

Even more than a century after the discovery of Chagas disease by Brazilian physician Carlos Chagas in 1909, current chemotherapy for this condition relies on two drugs only—benznidazole (**BZ**) and nifurtimox (Dias et al., 2014). These nitroheterocyclic compounds, which were identified between the mid-1960s and 1970s, are effective only when administered during the acute stage of the disease, a limitation that leaves millions of chronic chagasic patients without appropriate treatment (Molina et al., 2015). Moreover, benznidazole and nifurtimox cause severe adverse effects in up to 40% of patients, leading to poor adherence to treatment (Pérez-Molina and Molina, 2018). These drawbacks highlight the urgent need for the development of effective and safe drugs for the therapy of Chagas disease (Olivera et al., 2015; Ferreira and Andricopulo, 2019).

The enzyme cruzain (EC 3.4.22.51) is the main cysteine protease of *T. cruzi* and has been explored as a validated molecular target in Chagas disease drug discovery (McKerrow, 1999; Jose Cazzulo et al., 2001). It is expressed throughout the life cycle of *T. cruzi* and is involved in critical biological processes such as the interaction with host cells, parasite reproduction and evasion from the host immunologic system (Engel et al., 1998; Ferreira and Andricopulo, 2017). Cruzain has been validated as a molecular target for Chagas disease drug discovery based on genetic studies of *T. cruzi* and the ability of cruzain inhibitors to decrease parasite burden *in vivo* (Doyle et al., 2011; Ndao et al., 2014). These studies have recently supported the design and identification of several classes of cruzain inhibitors, including vinyl sulfones, triazoles, pyrimidines, thiosemicarbazones, chalcones, nitroalkenes, and benzimidazoles (Rogers et al., 2012; Ferreira et al., 2014; Avelar et al., 2015; Espíndola et al., 2015; Neitz et al., 2015; Latorre et al., 2016). The vinyl sulfone K777, which is a covalent cruzain inhibitor, showed promising results in preclinical efficacy tests; however, toxicity-related drawbacks prevented the compound from progressing into advanced clinical development (Ndao et al., 2014). The poor safety profile of K777 was associated with the irreversible mode of action of the compound. Following the failure of K777, the pursuit of novel cruzain inhibitors has recently focused on the design of reversible ligands. These investigations, along with the available structural data of cruzain bound with small molecule ligands, have been key to promoting the discovery of novel classes of inhibitors with improved safety profiles. Moreover, these data have enabled the integration of experimental and computational approaches into robust structure-based drug design (SBDD) campaigns that have been key to identifying novel chemical diversity to be explored in Chagas disease drug discovery.

MATERIALS AND METHODS

Molecular Docking

The three-dimensional structures of the designed cruzain inhibitors were constructed using the standard geometric parameters of SYBYL-X 2.1 (Certara, Princeton, NJ). Each compound was energetically minimized using the Tripos force field (Clark et al., 1989) and Powell conjugate gradient algorithm (Powell, 1977) with a convergence criterion of 0.05 kcal/mol.Å and Gasteiger-Hückel charges (Gasteiger and Marsili, 1980). The designed imide derivatives were docked into the cruzain catalytic site using GOLD 5.3 (Cambridge Crystallographic Data Centre, Cambridge, UK) (Jones et al., 1997). The X-ray structure of cruzain (PDB 3KKU, 1.28 Å) (Ferreira et al., 2010) was prepared by removing the water molecules and adding hydrogen atoms. The active site residues Cys25 and His162 were maintained as negatively charged and protonated, respectively. A sphere with a 10 Å radius centered on the sulfur atom of Cys25 was settled as the binding site. Compounds were docked by applying the GoldScore scoring function with a search efficiency of 200%. Visual analysis of the molecular docking-derived binding conformations was carried out with PyMOL 1.3 (Schrödinger, New York, NY) (Lill and Danielson, 2011).

Pro-Cruzain Expression, Activation, and Purification

Cruzain was expressed and purified using a modified version of a previously published protocol (Ferreira et al., 2014). *Escherichia coli* (M15) was preinoculated in Luria Bertani (LB) medium with ampicillin (100 μg/mL) and kanamycin (50 μg/mL) and kept overnight (37°C, 200 rpm). The preinoculum was diluted 10-fold in fresh LB medium (1 L) supplemented with 0.5 M NaCl, 0.2% glucose, 1 mM betaine, 0.5 M sorbitol, 100 μg/mL ampicillin and 50 μg/mL kanamycin, and incubated (37°C, 200 rpm). When the optical density (OD_{600}) reached 0.9, the culture was incubated at 47°C for 20 min to induce the expression of chaperones. Next, cruzain expression was induced by adding 0.2 mM isopropyl β-D-thiogalactopyranoside (IPTG), and the culture was kept overnight at 20°C. The cells were harvested by centrifugation (5,000 rpm, 30 min, 4°C) and suspended in 50 mL of lysis buffer (300 mM NaCl, 50 mM Tris-HCl, 1.6 mg/mL lysozyme, pH 8.0). The cells were then lysed by sonication (12 cycles of 30 s), and the suspension was centrifuged (9,000 rpm, 30 min, 4°C). The supernatant was collected, cruzain was precipitated by the addition of 35% ammonium sulfate (2 h), and this suspension was centrifuged (9,000 rpm, 30 min, 4°C). The precipitate was resuspended in lysis buffer and dialyzed to remove the ammonium sulfate. The soluble fraction of cruzain was purified by metal affinity chromatography using a Ni–NTA resin (Qiagen, Hilden, Germany). Contaminant proteins were removed using a washing buffer (300 mM NaCl, 50 mM Tris-HCl, 10 mM imidazole, pH 8.0). Next, cruzain was eluted with an increasing gradient of imidazole (25, 50, 75, 100, and 250 mM). The fractions containing cruzain were dialyzed in 0.1 M acetate buffer (1.5 L, pH 5.5) and concentrated to 0.5 mg/mL for subsequent activation.

Pro-cruzain was incubated in activation buffer (100 mM sodium acetate, 10 mM EDTA, 5 mM DTT, and 1 M NaCl, pH 5.5) at 37°C (Mott et al., 2010). The activation process was monitored at 30 min intervals by the cleavage of the substrate Z-Phe-Arg-AMC and confirmed by SDS-PAGE. Next, the mature enzyme was diluted 20-fold in binding buffer (20 mM sodium phosphate, 150 mM NaCl, pH 7.2) and incubated overnight with thiopropyl Sepharose 6B resin (GE Healthcare Life Sciences, Pittsburgh, PA) at 4°C. Cruzain was eluted using binding buffer supplemented with 20 mM DTT. Fractions containing the enzyme were stored at −80°C in 0.1 M sodium acetate (pH 5.5).

Cruzain Inhibition Assays

The catalytic activity of cruzain was monitored by cleavage of the fluorogenic substrate Z-Phe-Arg-aminomethyl coumarin (Z-Phe-Arg-AMC) as previously described (Ferreira et al., 2014). The enzyme kinetics assays were performed in 0.1 M sodium acetate buffer (pH 5.5) with 5 mM dithiothreitol (DTT) and 0.01% Triton X-100. The final concentrations of cruzain and substrate ($K_m = 1.6\,\mu M$) were 1.5 nM and 5.0 μM, respectively, except in the tests for the mechanism of inhibition, in which different substrate concentrations were used. The enzyme reaction was monitored for 5 min at 30°C in 96-well black flat bottom plates, and the activity was calculated based on the initial rates compared with a control (DMSO). Wavelengths of 355 nm for excitation and 460 nm for emission were used. IC_{50} values were independently determined by determining rate measurements for at least six inhibitor concentrations, each evaluated in triplicate. To determine the mechanism of inhibition, eight substrate concentrations and four inhibitor concentrations were used, each in triplicate. The mechanism of inhibition was determined by Lineweaver-Burk plots. SigmaPlot 10.0 (Systat Software Inc., Erkrath, Germany) was used to determine the IC_{50} values.

T. cruzi Intracellular Amastigote Assays

In vitro assays against *T. cruzi* were performed as described previously (Ferreira et al., 2014). The *T. cruzi* Tulahuen strain, which expresses the *E. coli* β-galactosidase gene *lacZ* (Buckner et al., 1996), was provided by Frederick S. Buckner (University of Washington, Seattle, WA). Stock solutions of the synthesized compounds were prepared in 100% DMSO. Epimastigotes were grown in liver infusion tryptone (LIT) supplemented with 10% fetal calf serum (FCS), penicillin and streptomycin (28°C). Metacyclogenesis from epimastigotes to trypomastigotes was induced by incubation of the epimastigotes in Grace's insect medium (Sigma-Aldrich, St. Louis, MO) supplemented with 10% FCS (28°C). HFF-1 human fibroblasts were seeded at 2×10^3/well (80 μL) in 96-well tissue culture plates in RPMI 1640 without phenol red supplemented with 10% FCS and incubated overnight (37°C, 5% CO_2). Next, trypomastigotes were added at 1.0×10^4/well (20 μL), and the plates were incubated (37°C, 5% CO_2). After 24 h, 3-fold serial dilutions (50 μL) of the synthesized compounds were added at concentrations ranging from 0.1 to 100 μM, and the plates were incubated (37°C, 5% CO_2). Each compound concentration was evaluated in triplicate. All plates included **BZ** (Sigma-Aldrich) as a positive control and untreated wells (100% growth) as a negative control. After 120 h, 50 μL of chlorophenol red β-D-galactopyranoside (CPRG, Sigma-Aldrich) and IGEPAL CA-630 (Sigma-Aldrich) (0.1%) was added to each well. The absorbance was measured at 570 nm in an automated microplate reader, and the data were transferred to SigmaPlot 10.0 (Systat Software Inc.) for IC_{50} value calculation.

Cytotoxicity in HFF-1 and HepG2 Cell Lines

The cytotoxicities of the compounds against HFF-1 fibroblasts and HepG2 hepatocytes were evaluated using the MTS tetrazolium assay (Promega, Madison, WI) (Barltrop et al., 1991). HFF-1 fibroblasts were seeded at 2×10^3/well in 96-well culture plates in RPMI 1640 without phenol red supplemented with 10% FCS and incubated overnight (37°C, 5% CO_2). HepG2 hepatocytes were seeded at 6×10^3/well in 96-well culture plates in DMEM (Cultilab, Campinas SP) supplemented with 10% FCS and incubated overnight (37°C, 5% CO_2). The compounds were added in 3-fold serial dilutions, and the plates were incubated at 37°C with 5% CO_2. Each compound concentration was evaluated in triplicate. All plates included doxorubicin (Sigma-Aldrich, St. Louis, MO) as a positive control and untreated wells (100% growth) as a negative control. After 72 h, 20 μL of MTS was added, and the plates were incubated for 4 h. The absorbance was measured at 490 nm in an automated microplate reader, and the data were transferred to SigmaPlot 10.0 (Systat Software Inc., Erkrath, Germany) for IC_{50} value calculation. The percent of nonviable cells was determined and compared to the negative control wells (100% growth).

Chemistry

All reactions were performed under an argon atmosphere with dry solvents and magnetic stirring unless stated otherwise. Dichloromethane (DCM) and triethylamine (Et_3N) were distilled from CaH_2. Tetrahydrofuran (THF) was distilled from sodium/benzophenone. Dimethyl formamide (DMF) was purchased from Sigma-Aldrich (anhydrous) and used without further purification. Yields refer to homogeneous materials obtained after purification of the reaction products by flash column chromatography using silica gel (200–400 mesh) or recrystallization. Analytical thin-layer chromatography was performed on silica-gel 60 and GF (5–40 μm thickness) plates, and visualization was accomplished using UV light, basic potassium permanganate staining or ninhydrin solution followed by heating. ^1H and proton-decoupled ^{13}C NMR spectra were acquired in $CDCl_3$, CD_3OD, or d_6-DMSO at 250 MHz (^1H) and 62.5 MHz (^{13}C) (Bruker DPX250), at 400 MHz (^1H) and 100 MHz (^{13}C) (Bruker Avance 400), at 500 MHz (^1H) and 125 MHz (^{13}C) (Varian Inova 500), or at 600 MHz (^1H) and 150 MHz (^{13}C) (Bruker Avance 600). Chemical shifts (δ) are reported in ppm using residual undeuterated solvent as an internal standard ($CDCl_3$ at 7.26 ppm, CD_3OD at 3.31 ppm, d_6-DMSO at 2.50 ppm, and TMS at 0.00 ppm for ^1H NMR spectra and $CDCl_3$ at 77.16 ppm, CD_3OD at 49.0 ppm, d_6-DMSO at 39.52 ppm for ^{13}C NMR spectra). Multiplicity data are reported as follows: s = singlet, d = doublet, t = triplet, q = quartet, br s = broad singlet, dd = doublet of doublets, dt = doublet of triplets, ddd = doublet of doublet of doublets and m = multiplet. The

multiplicity is followed by the coupling constant(s) in Hz and integration. High-resolution mass spectrometry (HRMS) was measured using electrospray ionization (ESI) (Waters Xevo Q-TOF, Thermo LTQ-FT Ultra, or Thermo Q-Exactive) or electron ionization (EI) (GCT Premier, Waters).

General Procedures for the Preparation of Alcohols 2a-m

Method A: The aniline derivative containing the appropriate substituents was dissolved in dry dichloromethane (DCM) under an argon atmosphere with magnetic stirring. The solution was cooled to 0°C followed by the slow addition of acetoxyacetyl chloride (1.1 equivalents). The mixture was stirred for 5 min before the addition of zinc mesh 20 (0.1 equivalents) and removal of the ice bath, allowing the reaction to reach room temperature. The reaction mixture was monitored by thin layer chromatography (TLC), and after 30 min, total conversion was observed. Half of the initial volume of DCM was added to dilute the reaction mixture, and the zinc was removed by filtration. The supernatant was washed with an aqueous sodium bicarbonate solution. The organic phase was dried over magnesium sulfate ($MgSO_4$) and concentrated to dryness under vacuum. The product was purified by column chromatography on silica gel using a mixture of ethyl acetate/hexanes (50%) as the eluent. The obtained ester was dissolved in methanol at room temperature followed by the addition of solid potassium carbonate (K_2CO_3). The reaction mixture was stirred for 1 h and then quenched by the addition of ethyl acetate. The mixture was washed with aqueous saturated ammonium chloride (NH_4Cl) solution and brine. The resulting organic phase was dried over $MgSO_4$ and concentrated under reduced pressure. Purification was performed by column chromatography on silica gel using a mixture of ethyl acetate/hexanes (50%) as the eluent.

Method B: The aniline derivative containing the appropriate substituents was dissolved in dry DCM under an argon atmosphere with magnetic stirring. To the resulting solution, dry triethylamine (1.1 equivalents) and acetoxyacetyl chloride (1.1 equivalents) were added and stirring was maintained for 2 h to monitor the conversion by TLC. The mixture was then washed with saturated aqueous NH_4Cl solution and brine solution. The organic phase was dried over $MgSO_4$ and concentrated to dryness under vacuum. The obtained ester was dissolved in methanol at room temperature followed by the addition of solid potassium carbonate (K_2CO_3). The reaction mixture was stirred for 1 h and then quenched by the addition of ethyl acetate. The mixture was washed with aqueous saturated ammonium chloride (NH_4Cl) solution and brine. The resulting organic phase was dried over $MgSO_4$ and concentrated under reduced pressure. Purification was performed by column chromatography on silica gel using a mixture of ethyl acetate/hexanes (50%) as the eluent.

General Procedures for Production of Carboxylic Acids 1 and 9a-n

Method C: The appropriate cyclic anhydride and amino acid (1.1 equivalents) were dissolved in glacial acetic acid under an argon atmosphere. The mixture was magnetically stirred overnight followed by 4 h of reflux periodically monitored by TLC. After consumption of the cyclic anhydride, the reaction pot was cooled to room temperature. Acetic acid was removed in a rotary evaporator, and ice-cold water was added to the resulting slurry to generate a precipitate. Concentrated hydrochloric acid (HCl) was added, and the solid was collected by filtration, followed by high vacuum drying.

Method D: The appropriate cyclic anhydride and amino acid (1.1 equivalents) were added to a reaction flask containing toluene. To this suspension, TEA (0.1 equivalents) was added and a Dean-Stark apparatus was coupled to the system. The mixture refluxed for 4 h and cooled to room temperature. Toluene was removed in a rotary evaporator, and the residue was dissolved in ethyl acetate and washed twice with 1 M HCl. The organic phase was extracted with saturated aqueous sodium bicarbonate solution. The aqueous phase was acidified with HCl solution to pH 5.0 and extracted with ethyl acetate. The combined organic phases were dried over $MgSO_4$, concentrated under vacuum and used in the following step without further purification.

Method E: Leucine was dissolved in a 1:1 mixture of hydrobromic acid and water in an ice bath. Sodium nitrite (6.5 M aqueous solution) was added dropwise. The ice bath was removed, and the solution stirred at room temperature for 2.5 h. The acidic residues were removed under reduced pressure, and the remaining solution was extracted with diethyl ether 3 times. The combined organic phases were dried over $MgSO_4$ and concentrated under vacuum.

The residue was dissolved in DCM under an argon atmosphere. The solution was cooled in an ice bath and benzyl alcohol (1.1 equivalents), 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC) (1.1 equivalents) and 4-dimethylaminopyridine (DMAP) (0.1 equivalents) were added. The ice bath was removed, and the mixture stirred for 5 h. Then, the reaction mixture was washed with saturated aqueous NH_4Cl solution and brine. The organic phase was dried over $MgSO_4$ and concentrated under vacuum. The product was purified by column chromatography on silica gel using a mixture of ethyl acetate/hexanes (50%) as the eluent.

A solution of the pure ester was prepared in acetonitrile under an argon atmosphere, and the appropriate cyclic amine was added (2.0 equivalents). To the resulting solution, cesium carbonate (1.1 equivalents) was added followed by stirring for 2 h. The solids were removed by filtration, and the solvent was removed under reduced pressure followed by redissolution in DCM and washing with water. The organic phase was dried over $MgSO_4$ and concentrated to dryness under vacuum. The product was purified by column chromatography on silica gel using a gradient from 0 to 4% methanol in DCM as the eluent.

The ester was dissolved in methanol, and 10% Pd/C was added (0.1 equivalents). The mixture stirred at room temperature under a hydrogen atmosphere (1 atm) for 30 min. The catalyst was removed by filtration over Celite®, and the solvent removed under reduced pressure to generate the pure carboxylic acid.

General Procedure for the Preparation of Compounds 3a-m and 10a-n

Method F: A solution of carboxylic acid in dry DCM was prepared, and the alcohol was added (1.1 equivalents). The solution was cooled in an ice bath followed by the addition of EDC (1.2 equivalents) and DMAP (0.1 equivalents). The

solution was allowed to warm to room temperature and stirred for 4 h while monitoring periodically by TLC. The reaction mixture was washed with saturated aqueous NH_4Cl solution and brine. The organic phase was dried over $MgSO_4$ and concentrated under reduced pressure. The product was purified by column chromatography on silica gel using a mixture of ethyl acetate/hexanes (50%) as the eluent.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl (2S)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3a)

Methods A and C followed by F

Very viscous liquid, 90% ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.26 (sl, 1H); 7.83 (d, J = 2.5 Hz, 1H); 7.68 (dd, J = 9.0, 2.5 Hz, 1H); 6.90 (d, J = 9.0 Hz, 1H); 5.90 (m, 2H); 5.04 (d, J = 15.6 Hz, 1H); 4.88 (dd, J = 11.0, 4.4 Hz, 1H); 4.40 (d, J = 15.6 Hz, 1H); 3.89 (s, 3H); 3.23 (m, 2H); 2.65 (m, 2H); 2.27 (m, 2H); 2.08 (m, 1H); 1.94 (m, 1H); 1.41 (m, 1H); 0.93 (d, J = 6.6 Hz, 6H), ^{13}C NMR (125 MHz, $CDCl_3$) δ 180.4; 180.3; 168.1; 164.7; 151.9; 131.1; 127.8; 127.7; 122.3; 122.2; 119.4; 112.1; 63.4; 56.3; 51.0; 39.3; 39.2; 37.2; 24.7; 23.5; 23.4; 23.0; 20.9.

HRMS (ESI-Orbitrap): $[M+H]^+$ Calculated for $C_{23}H_{27}O_6N_2Cl$ 463.1617; found 463.1636.

2-((4-methoxyphenyl)amino)-2-oxoethyl (2S)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3b)

Methods B and C followed by F

Very viscous liquid, 88% ^{1}H NMR (250 MHz, DMSO) δ 9.78 (sl, 1H); 7.46 (d, J = 9.0 Hz, 1H); 6.89 (d, J = 9.0 Hz, 1H); 5.85 (m, 1H); 4.79 (dd, J = 11.1, 4.4 Hz, 1H); 4.63 (m, 2H); 3.72 (s, 3H); 3.24 (m, 2H); 2.39 (m, 2H); 2.20 (m, 2H); 1.99 (m, 1H); 1.77 (m, 1H); 1.34 (m, 1H); 0.84 (m, 6H), ^{13}C NMR (62.5 MHz, DMSO) δ 179.6; 179.5; 168.5; 164.3; 155.5; 131.3; 127.4; 120.4; 113.9; 63.4; 55.2; 50.2; 36.0; 24.1; 23.2; 23.0; 20.7.

HRMS (ESI-Orbitrap): $[M+H]^+$ Calculated for $C_{23}H_{28}O_6N_2$ 429.2026; found 429.2039.

2-((3-chlorophenyl)amino)-2-oxoethyl (2S)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3c)

Methods B and C followed by F

Very viscous liquid, 91% ^{1}H NMR (250 MHz, DMSO) δ 10.15 (m, 1H); 7.77 (m, 1H); 7.44 (m, 1H); 7.33 (m, 1H); 7.11 (m, 1H); 5.84 (m, 2H); 5.03 (m, 1H); 4.67 (m, 2H); 3,59 (m, 2H); 3.23 (m, 1H); 2.43 (m, 3H); 2.19 (m, 2H); 1.97 (m, 1H); 1.75 (m, 1H); 1.34 (m, 1H); 0.83 (m, 6H). ^{13}C NMR (62.5 MHz, DMSO) δ 179.6; 179.5; 179.4; 173.3; 173.2; 172.4; 169.2; 168.6; 165.7; 165.2; 139.8; 139.7; 133.2; 130.5; 127.5; 127.4; 125.0; 124.9; 123.4; 123.3; 118.8; 117.7; 63.4; 62.7; 52.3; 51.6; 51.5; 50.3; 50.2; 38.4; 36.1; 36.0; 25.3; 24.2; 24.1; 23.2; 23.1; 23.0; 20.7.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{22}H_{25}O_5N_2ClNa$ 455.1350; found 455.1362.

2-oxo-2-(phenylamino)ethyl (2S)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3d)

Methods B and C followed by F

Very viscous liquid, 78% ^{1}H NMR (400 MHz, DMSO) δ 9.97 (s, 1H); 7.55 (m, 2H); 7.32 (m, 2H); 7.07 (m, 1H); 5.85 (m,2H); 4.79 (dd, J = 11.2, 4.4 Hz, 1H); 4.67 (m, 2H); 3.25 (m, 2H); 2.39 (m, 2H); 2.20 (dd, J = 14.9, 7.2 Hz, 2H); 2.03 (m, 1H); 1.77 (ddd, J = 14.2, 10.1, 4.5 Hz, 1H); 1.35 (m, 1H); 0.84 (dd, J = 7.6, 6.6 Hz, 6H). ^{13}C NMR (100 MHz, DMSO) δ 179.6; 168.5; 164.8; 138.2; 128.8; 127.4; 123.7; 119.3; 63.4; 50.2; 38.4; 36.0; 24.1; 23.2; 23.0; 20.7.

2-((4-fluorophenyl)amino)-2-oxoethyl (2S)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3e)

Methods B and C followed by F

Very viscous liquid, 82% ^{1}H NMR (250 MHz, DMSO) δ 10.02 (s, 1H); 7.56 (dd, J = 8.6, 5.0 Hz, 2H); 7.15 (t, J = 8.6 Hz, 2H); 5.84 (m, 2H); 4.79 (dd, J = 10.9, 4.3 Hz, 1H); 4.65 (m, 2H); 3.24 (m, 2H); 2.43 (m, 2H); 2.20 (m, 2H); 2.02 (m, 1H); 1.76 (m, 1H); 1.35 (m, 1H); 0.84 (m, 6H).

HRMS (ESI-Orbitrap): $[M+H]^+$ Calculated for $C_{22}H_{25}O_5N_2F$ 417.1826; found 417.1817.

2-((4-bromophenyl)amino)-2-oxoethyl (2S)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3f)

Methods B and C followed by F

Very viscous liquid, 84% ^{1}H NMR (250 MHz, DMSO) δ 10.13 (s, 1H); 7.53 (m, 4H); 5.84 (m, 2H); 4.78 (dd, J = 11.1, 4.5 Hz, 1H); 4.66 (m, 2H); 3.27 (m, 2H); 2.39 (m, 2H); 2.19 (m, 2H); 2.02 (m, 1H); 1.76 (m, 1H); 1.33 (m, 1H); 0.83 (m, 6H). ^{13}C NMR (62.5 MHz, DMSO) δ 179.7; 179.6; 168.6; 165.0; 137.7; 131.7; 127.6; 127.5; 121.3; 115.4; 63.4; 50.3; 36.0; 24.1; 23.2; 23.1; 20.8.

HRMS (ESI-Orbitrap): $[M+K]^+$ Calculated for $C_{22}H_{25}O_5N_2BrK$ 515.0577; found 515.0584.

2-((4-chlorophenyl)amino)-2-oxoethyl (2S)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3g)

Methods B and C followed by F

Very viscous liquid, 75% ^{1}H NMR (500 MHz, DMSO) δ 10.13 (s, 1H); 7.59 (m, 2H); 7.37 (m, 2H); 5.84 (m, 2H); 4.79 (dd, J = 11.2, 4.5 Hz, 1H); 4.67 (m, 2H); 3.24 (m, 2H); 2.39 (m, 2H); 2.20 (dd, J = 14.9, 7.4 Hz, 2H); 2.02 (m, 1H); 1.76 (ddd, J = 14.2, 10.1, 4.5 Hz, 1H); 1.35 (m, 1H); 0.84 (dd, J = 9.8, 6.7 Hz, 6H). ^{13}C NMR (125 MHz, DMSO) δ 179.6; 179.5; 168.5; 165.0; 137.2; 128.8; 127.5; 127.4; 127.3; 120.9; 63.4; 50.2; 38.5; 36.0; 24.1; 23.2; 23.0; 20.7.

HRMS (ESI-Orbitrap): $[M+H]^+$ Calculated for $C_{22}H_{26}O_5N_2Cl$ 433.1530; found 433.1507.

2-((4-hydroxyphenyl)amino)-2-oxoethyl (2S)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3h)

Methods B and C followed by F

Very viscous liquid, 84% ^{1}H NMR (250 MHz, DMSO) δ 10.05 (s, 1H); 7.57 (d, J = 8.8 Hz, 1H); 6.98 (d, J = 8.8 Hz, 1H); 5.80 (m, 2H); 4.88 (m, 1H); 4.66 (m, 1H); 3.24 (m, 1H); 2.40 (m, 2H); 2.19 (m, 2H); 1.98 (m, 1H); 1.77 (m, 1H); 1.35 (m, 1H); 0.85 (m, 6H).

2-(naphthalen-2-ylamino)-2-oxoethyl (2S)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3i)

Methods B and C followed by F

Very viscous liquid, 68% ^{1}H NMR (250 MHz, DMSO) δ 8.43 (s, 2H); 7.80 (m, 4H); 7.45 (m, 2H); 5.91 (m, 2H); 5.13 (d, J = 15.5 Hz, 1H); 4.92 (dd, J = 10.8, 4.5 Hz, 1H); 4.47 (d, J = 15.5 Hz, 1H); 3.23 (m, 2H); 2.68 (m, 2H); 2.22 (m, 3H); 1.97 (m, 1H); 0.95 (d, J = 6.5 Hz, 6H).

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{26}H_{28}O_5N_2Na$ 471.18959; found 471.18888.

2-((4-fluoro-3-nitrophenyl)amino)-2-oxoethyl (2S)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3j)

Methods A and C followed by F

Very viscous liquid, 84% ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.72 (sl, 1H); 8.56 (dd, J = 6.6, 2.7 Hz, 1H); 8.20 (m, 1H); 7.27 (m, 1H); 5.92 (m, 2H); 5.07 (d, J = 15.8 Hz, 1H); 4.90 (dd, J = 10.7, 4.6 Hz, 1H); 4.46 (d, J = 15.7 Hz, 1H); 3.26 (m, 2H); 2.67 (m, 2H); 2.31 (m, 2H); 1.99 (m, 2H); 1.41 (m, 1H); 0.93 (d, J = 6.4 Hz, 6H), ^{13}C NMR (100 MHz, $CDCl_3$) δ 180.7; 180.5; 153.1; 150.5; 137.1; 137.0; 134.3; 134.2; 127.8; 127.6; 126.7; 126.6; 118.7; 118.5; 117.00; 116.97; 63.3; 51.1; 39.4; 39.3; 37.4; 29.6; 24.7; 23.6; 23.5; 23.0; 21.0.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{22}H_{24}O_7N_3FNa$ 484.14095; found 484.14483.

2-((4-((tert-butoxycarbonyl)(methyl)amino)phenyl)amino)-2-oxoethyl (2S)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3k)

Methods A and C followed by F

Very viscous liquid, 90% ^{1}H NMR (500 MHz, $CDCl_3$) δ 7.74 (d, J = 8.5 Hz, 2H); 7.21 (d, J = 8.5 Hz); 5.90 (m, 2H); 5.05 (d, J = 15.9 Hz, 1H); 4.88 (dd, J = 11.0, 4.3 Hz, 1H); 4.41 (d, J = 15.3 Hz, 1H); 3.24 (s, 3H); 2.65 (m, 2H); 2.27 (m, 2H); 2.10 (m, 1H); 1.95 (m, 1H); 1.45 (s, 9H); 0.92 (d, J = 6.71 Hz, 6H), ^{13}C NMR (125 MHz, $CDCl_3$) δ 180.4; 180.3; 168.1; 164.8; 154.8; 140.2; 134.7; 127.72; 127.68; 125.9; 120.1; 80.2; 63.4; 51.0; 39.3; 39.2; 37.3; 37.2; 28.3; 24.7; 23.6; 23.5; 23.0; 20.9.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{28}H_{37}O_7N_3Na$ 550.25237; found 550.25177.

2-((4-nitrophenyl)amino)-2-oxoethyl (2S)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3l)

Methods A and C followed by F

Very viscous liquid, 89% ^{1}H NMR (600 MHz, $CDCl_3$) δ 8.72 (sl, 1H); 8.24 (m, 2H); 8.03 (m, 2H); 5.91 (m, 2H); 5.10 (d, J = 15.8 Hz, 1H); 4.90 (dd, J = 11.0, 4.4Hz, 1H); 3.25 (m, 2H); 2.66 (m, 2H); 2.29 (m, 2H); 2.06 (m, 1H); 1.96 (m, 1H); 1.41 (m, 1H); 0.93 (d, J = 6.6 Hz, 6H), ^{13}C NMR (150 MHz, $CDCl_3$) δ 180.6; 180.5; 168.1; 165.7; 143.9; 143.4; 127.8; 127.7; 124.9; 119.6; 63.4; 51.1; 39.37; 39.35; 37.4; 24.7; 23.6; 23.5; 23.0; 21.0.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{22}H_{25}O_7N_3Na$ 466.15902; found 466.15749.

2-((4-methoxy-2-nitrophenyl)amino)-2-oxoethyl (2S)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3m)

Methods A and C followed by F

Very viscous liquid, 98% ^{1}H NMR (500 MHz, $CDCl_3$) δ 10.51 (sl, 1H); 8.57 (d, J = 9.15 Hz, 1H); 7.66 (d, J = 2.9 Hz, 1H); 7.24 (dd, J = 9.3, 3.0 Hz, 1H); 5.92 (m, 2H); 4.95 (dd, J = 11.3, 4.4 Hz, 1H); 4.77 (m, 2H); 3.87 (s, 3H); 3.31 (m, 1H); 3.18 (m, 1H); 2.62 (m, 2H); 2.26 (m, 2H); 2.16 (m, 1H); 1.97 (m, 1H); 1.39 (m, 1H); 0.93 (dd, J = 6.6, 4.5 Hz, 6H), ^{13}C NMR (125 MHz, $CDCl_3$) δ 179.5; 179.43; 168.0; 165.0; 155.5; 137.7; 127.7; 127.6; 126.6; 124.0; 122.8; 108.8; 64.0; 55.8; 50.9; 39.2; 38.9; 36.7; 24.6; 23.6; 23.4; 23.0; 20.77.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{23}H_{28}O_8N_3$ 474.19045; found 474.19131.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl (S)-2-(1,3-dioxoisoindolin-2-yl)-4-methylpentanoate (10a)

Methods A and C followed by F

Very viscous liquid, 82% ^{1}H NMR (250 MHz, DMSO) δ 10.04 (sl, 1H); 7.92 (m, 4H); 7.69 (d, J = 2.4 Hz, 1H); 7.37 (dd, J = 8.9, 2.4 Hz, 1H); 7.10 (d, J = 9.0 Hz, 1H); 5.04 (dd, J = 11.2, 4.4 Hz, 1H); 4.72 (m, 2H); 3.81 (s, 3H); 2.24 (m, 1H); 1.93 (ddd, J = 14.0, 9.9, 4.5 Hz, 1H); 1.49 (m, 1H); 0.89 (m, 6H). ^{13}C NMR (62.5 MHz, DMSO) δ 169.1; 167.2; 164.6; 150.7; 135.0; 131.9; 131.0; 123.5; 121.0; 120.6; 119.2; 112.9; 63.4; 56.1; 49.9; 36.6; 24.5; 22.9; 20.8.

HRMS (ESI-Orbitrap): $[M+H]^+$ Calculated for $C_{23}H_{24}O_6N_2Cl$ 459.1323; found 459.1355.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl (2S)-2-(1,3-dioxooctahydro-2H-isoindol-2-yl)-4-methylpentanoate (10b)

Methods A and C followed by F

Very viscous liquid, 91% ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.32 (sl, 1H); 7.83 (d, J = 2.5 Hz, 1H); 7.68 (dd, J = 8.9, 2.6 Hz, 1H); 6.90 (d, J = 8.9 Hz, 1H); 5.08 (d, J = 15.6 Hz, 1H); 4.90 (dd, J = 10.9, 4.3 Hz, 1H); 4.45 (d, J = 15.6 Hz, 1H); 3.89 (s, 3H); 2.12 (ddd, J = 14.5, 10.7, 4.3 Hz, 1H); 1.98 (m, 2H); 1.85 (m, 2H); 1.77 (m, 1H); 1.47 (m, 4H); 1.36 (m, 1H); 0.96 (m, 6H). ^{13}C NMR (100 MHz, $CDCl_3$) δ 180.3; 179.4; 168.4; 164.7; 152.0; 131.1; 122.3; 119.5; 112.1; 63.5; 56.3; 50.5; 40.0; 39.8; 37.4; 25.0; 24.0; 23.3; 23.1; 21.6; 21.1.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{23}H_{29}O_6N_2ClNa$ 487.19118; found 487.16025.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl (S)-2-(2,5-dioxo-2,5-dihydro-1H-pyrrol-1-yl)-4-methylpentanoate (10c)

Methods A and D followed by F

Very viscous liquid, 46% ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.12 (sl, 1H); 7.76 (d, J = 2.6 Hz, 1H); 7.62 (dd, J = 8.9, 2.6 Hz, 1H); 6.90 (d, J = 8.9 Hz, 1H); 6.84 (s, 2H); 5.04 (d, J = 15.5 Hz, 1H); 4.91 (dd, J = 11.4, 4.4 Hz, 1H); 4.47 (d, J = 15.5 Hz, 1H); 3.89 (s, 3H); 2.18 (ddd, J = 14.3, 11.2, 4.3 Hz, 1H); 1.98 (ddd, J = 14.3, 10.0, 4.3 Hz, 1H); 1.47 (m, 1H); 0.96 (dd, J = 6.6, 1.2 Hz, 6H). ^{13}C NMR (100 MHz, $CDCl_3$) δ 170.4; 168.4; 164.6; 152.1; 134.6; 130.9; 122.5; 122.4; 63.6; 56.3; 50.4; 38.0; 24.9; 23.0; 21.0.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl (S)-2-(2,5-dioxopyrrolidin-1-yl)-4-methylpentanoate (10d)

Methods A and D followed by F

Very viscous liquid, 72% ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.30 (sl, 1H); 7.79 (d, J = 2.6 Hz, 1H); 7.62 (dd, J = 8.9, 2.7 Hz, 1H); 6.89 (d, J = 8.9 Hz, 1H); 5.01 (d, J = 15.4 Hz, 1H); 4.94 (dd, J = 10.5, 4.8 Hz, 1H); 4.49 (d, J = 15.5 Hz, 1H); 3.88 (s, 3H); 2.86 (m, 4H); 2.05 (m, 2H); 1.45 (m, 1H); 0.96 (dd, J = 6.7, 4.6 Hz, 6H), ^{13}C NMR (100 MHz, $CDCl_3$) δ 177.2; 168.2; 164.7; 152.0; 131.0; 122.3; 119.6; 112.1; 63.4; 56.3; 50.9; 37.3; 28.2; 24.9; 22.9; 21.1.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{19}H_{23}O_6N_2ClNa$ 433.11423; found 433.11424.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl (S)-4-methyl-2-(3-methyl-2,5-dioxo-2,5-dihydro-1H-pyrrol-1-yl)pentanoate (10e)

Methods A and D followed by F

Very viscous liquid, 88% ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.19 (sl, 1H); 7.78 (d, J = 2.5 Hz, 1H); 7.65 (dd, J = 8.9, 2.6 Hz, 1H); 6.90 (d, J = 9.0 Hz, 1H); 6.46 (d, J = 1.9 Hz, 1H); 5.04 (d, J = 15.6 Hz, 1H); 4.89 (dd, J = 11.4, 4.3 Hz, 1H); 4.45 (d, J = 15.6 Hz, 1H); 3.89 (s, 3H); 2.17 (m, 4H); 1.96 (m, 1H); 1.47 (m, 1H); 0.95 (dd, J = 6.7, 1.7 Hz, 6H), ^{13}C NMR (125 MHz, $CDCl_3$) δ 171.6; 170.5; 168.6; 164.6; 152.0; 146.3; 130.9; 127.8; 122.4; 122.3; 119.5; 112.1; 63.5; 56.3; 50.5; 38.1; 24.9; 23.0; 20.9; 11.2.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{20}H_{23}O_6N_2ClNa$ 445.11369; found 445.11261.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl (S)-2-(3,4-dichloro-2,5-dioxo-2,5-dihydro-1H-pyrrol-1-yl)-4-methylpentanoate (10f)

Methods A and D followed by F

Very viscous liquid, 27% ^{1}H NMR (500 MHz, $CDCl_3$) δ 7.85 (sl, 1H); 7.74 (d, J = 2.5 Hz, 1H); 7.59 (dd, J = 9.0, 2.5 Hz, 1H); 6.91 (d, J = 9.0 Hz, 1H); 5.02 (d, J = 15.4, 1H); 4.98 (dd, J = 11.4, 4.5 Hz, 1H); 4.52 (d, J = 15.4 Hz, 1H); 3.90 (s, 3H); 2.21 (m, 1H); 2.00 (m, 1H); 1.51 (m, 1H); 0.97 (dd, J = 6.5, 4.3 Hz, 6H), ^{13}C NMR (125 MHz, $CDCl_3$) δ 167.6; 164.1; 163.0; 152.3; 133.9; 130.6; 122.6; 122.4; 119.6; 112.2; 63.9; 56.4; 51.9; 37.9; 25.0; 23.0; 20.9.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{19}H_{19}O_6N_2Cl_3Na$ 499.02009; found 499.02014.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl 2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)acetate (10g)

Methods A and C followed by F

Very viscous liquid, 80% ^{1}H NMR (500 MHz, DMSO) δ 12.40 (sl, 1H); 9.89 (s, 1H); 7.75 (d, J = 2.6 Hz, 1H); 7.43 (dd, J = 8.9, 2.6 Hz, 1H); 7.11 (d, J = 9.0 Hz, 1H); 5.64 (m, 2H); 4.62 (m, 2H); 3.82 (s, 3H); 3.07 (m, 2H); 2.46 (m, 2H); 2.33 (m, 2H), ^{13}C NMR (125 MHz, DMSO) δ 174.6; 172.6; 165.3; 150.7; 132.0; 125.3; 124.9; 121.1; 120.6; 119.3; 112.9; 62.5; 56.2; 38.8; 38.6; 25.6; 25.2.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl (2S)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-3-phenylpropanoate (10h)

Methods A and D followed by F

Very viscous liquid, 87% ^{1}H NMR (250 MHz, $CDCl_3$) δ 8.26 (sl, 1H); 7.83 (d, J = 2.7 Hz, 1H); 7.71 (dd, J = 8.9, 2.6 Hz, 1H); 7.22 (m, 5H); 6.92 (d, J = 8.8 Hz, 1H); 5.70 (m, 2H); 5.12 (m, 2H); 4.42 (d, J = 15.5 Hz, 1H); 3.90 (s, 3H); 3.56 (m, 1H); 3.36 (dd, J = 14.3, 11.3 Hz, 1H); 2.99 (m, 2H); 2.48 (ddd, J = 15.8, 6.0, 2.9 Hz, 1H); 2.22 (m, 3H).

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{26}H_{25}O_6N_2ClNa$ 519.12988; found 519.12918.

$[\alpha]_D^{20}$ −24,0 (1, MeOH).

HPLC flow 3.5 mL/min, polarity hexane:ethyl acetate (60:40). Rt 16.76 min, area 97.10%; Rt 20.09 min, area 2.90%.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl 2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-3-phenylpropanoate (10h$_{rac}$)

Methods A and D followed by F

Very viscous liquid, 81% ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.29 (sl, 1H); 7.84 (d, J = 2.2 Hz, 1H); 7.69 (dd, J = 8.8, 2.4 Hz, 1H); 7.24 (m, 3H); 7.13 (d, J = 7.1 Hz, 2H); 6.92 (d, J = 9.0 Hz, 1H); 5.70 (m, 2H); 5.12 (m, 2H); 4.43 (d, J = 15.6 Hz, 1H); 3.90 (s, 3H); 3.56 (dd, J = 14.3, 5.2 Hz, 1H); 3.35 (dd, J = 14.2, 11.5 Hz, 1H); 3.04 (m, 1H); 2.93 (m, 1H); 2.48 (ddd, J = 15.7, 6.05, 2.6 Hz, 1H); 2.29 (ddd, J = 15.8, 5.9, 3.0 Hz, 1H); 2.17 (m, 2H), ^{13}C NMR (125 MHz, $CDCl_3$) δ 180.2; 179.8; 167.3; 164.6; 152.0; 135.3; 131.0; 129.0; 128.6; 127.4; 127.21; 127.17; 122.33; 122.28; 119.5; 112.1; 63.4; 56.3; 53.2; 39.0; 38.9; 34.3; 23.11; 23.07.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{26}H_{25}O_6N_2ClNa$ 519.12988; found 519.12965.

$[\alpha]_D^{20}$ 0 (1, MeOH).

HPLC flow 3.5 mL/min, polarity hexane:ethyl acetate (60:40). Rt 17.16 min, area 50.19%; Rt 20.49 min, area 49.81%.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl (2R)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-3-(naphthalen-2-yl)propanoate (10i)

Methods A and D followed by F

Very viscous liquid, 43% (for 2 steps) ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.29 (sl, 1H); 7.85 (d, J = 2.7 Hz, 1H); 7.77 (m, 3H); 7.71 (dd, J = 8.9, 2.6 Hz, 1H); 7.54 (m, 1H); 7.46 (m, 2H); 7.29 (dd, J = 8.4, 1.8 Hz, 1H); 6.93 (d, J = 8.9 Hz, 1H); 5.48 (m, 2H); 5.26 (dd, J = 11.4, 5.4 Hz, 1H); 5.14 (d, J = 15.4 Hz, 1H); 4.44 (d, J = 15.4 Hz, 1H); 3.72 (dd, J = 14.3, 5.3 Hz, 1H); 3.54 (dd, J = 14.4, 11.4 Hz, 1H); 3.00 (ddd, J = 9.0, 7.6, 3.0 Hz, 1H); 2.87 (ddd, J = 9.0, 7.8; 2.7 Hz, 1H); 2.46 (ddd, J = 15.8, 6.4, 2.6 Hz, 1H); 2.22 (ddd, J = 15.8, 6.4, 3.0 Hz, 1H); 2.08 (m, 2H), ^{13}C NMR (100 MHz, $CDCl_3$) δ 180.2; 180.0; 167.4; 164.7; 152.1; 133.3;

132.8; 132.5; 131.1; 128.5; 128.1; 127.7; 127.4; 127.2; 127.0; 126.6; 126.3; 125.9; 122.5; 122.4; 119.6; 112.2; 63.5; 56.4; 53.2; 39.0; 38.9; 34.6; 23.1.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{30}H_{27}O_6N_2ClNa$ 569.14553; found 569.14428.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl (S)-2-(2,5-dioxo-2,5-dihydro-1H-pyrrol-1-yl)-3-phenylpropanoate (10j)

Methods A and D followed by F

Very viscous liquid, 23% ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.17 (sl, 1H); 7.76 (d, J = 2.6 Hz, 1H); 7.66 (dd, J = 8.9, 2.6 Hz, 1H); 7.25 (m, 3H); 7.13 (m, 2H); 6.92 (d, J = 8.9 Hz, 1H); 6.69 (s, 2H); 5.10 (m, 2H); 4.50 (d, J = 15.4 Hz, 1H); 3.90 (s, 3H); 3.57 (dd, J = 14.2, 5.2 Hz, 1H); 3.38 (dd, J = 14.2, 11.2 Hz, 1H), ^{13}C NMR (125 MHz, $CDCl_3$) δ 170.1; 167.6; 164.5; 152.2; 135.5; 134.3; 130.9; 128.9; 128.8; 127.4; 122.5; 119.7; 112.2; 63.7; 56.4; 53.2; 35.4.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{22}H_{19}O_6N_2ClNa$ 465.08239; found 465.08244.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl (S)-4-methyl-2-(pyrrolidin-1-yl)pentanoate (10k)

Methods A and E followed by F

Very viscous liquid, 93% ^{1}H NMR (500 MHz, MeOD) δ 8.66 (sl, 1H); 7.57 (d, J = 2.7 Hz, 1H); 7.46 (dd, J = 8.8, 2.5 Hz, 1H); 6.88 (d, J = 9.0 Hz, 1H); 4.73 (m, 2H); 3.88 (s, 1H); 3.35 (dd, J = 9.9, 4.6 Hz, 1H); 2.73 (m, 2H); 2.63 (m, 2H); 1.80 (m, 5H); 1.58 (m, 2H); 0.96 (dd, J = 8.5, 6.1 Hz, 6H), ^{13}C NMR (125 MHz, MeOD) δ 171.3; 165.3; 152.2; 130.6; 122.7; 122.4; 120.0; 112.1; 65.1; 62.7; 56.3; 50.6; 39.0; 25.4; 23.4; 23.3; 21.9.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl (S)-4-methyl-2-(piperidin-1-yl)pentanoate (10l)

Methods A and E followed by F

Very viscous liquid, 87% ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.05 (sl, 1H); 7.56 (d, J = 2.7 Hz, 1H); 7.42 (dd, J = 8.9, 2.6 Hz, 1H); 6.90 (d, J = 8.8 Hz, 1H); 4.83 (d, J = 15.6 Hz, 1H); 4.63 (d, J = 15.6 Hz, 1H); 3.82 (s, 3H); 3.37 (dd, J = 8.8, 5.7 Hz, 1H); 2.59 (m, 4H); 1.60 (m, 7H); 1.44 (m, 2H); 0.94 (dd, J = 9.1, 6.3 Hz, 6H), ^{13}C NMR (125 MHz, $CDCl_3$) δ 170.8; 165.3; 152.4; 130.2; 123.0; 122.6; 120.1; 112.2; 66.5; 62.4; 56.4; 50.9; 36.6; 26.4; 25.3; 24.4; 22.8; 22.4.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{20}H_{29}O_4N_2ClNa$ 419.17081; found 419.17050.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl (S)-4-methyl-2-morpholinopentanoate (10m)

Methods A and E followed by F

Very viscous liquid, 75% ^{1}H NMR (500 MHz, $CDCl_3$) δ 7.96 (sl, 1H); 7.56 (d, J = 2.6 Hz, 1H); 7.38 (dd, J = 8.8, 2.6 Hz, 1H); 6.89 (d, J = 8.9 Hz, 1H); 4.76 (d, J = 15.3 Hz, 1H); 4.68 (d, J = 15.3 Hz, 1H); 3.88 (s, 3H); 3.68 (m, 4H); 3.40 (t, J = 7.3 Hz, 1H); 2.67 (m, 4H); 1.65 (m, 3H); 0.96 (dd, J = 8.9, 6.1 Hz, 6H), ^{13}C NMR (125 MHz, $CDCl_3$) δ 170.8; 165.0; 152.4; 130.1; 122.8; 122.6; 120.0; 112.2; 67.2; 65.8; 62.6; 56.3; 49.8; 36.8; 25.0; 22.6; 22.4.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{19}H_{27}O_5N_2ClNa$ 421.15007; found 421.14966.

tert-butyl (S)-4-(1-(2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethoxy)-4-methyl-1-oxopentan-2-yl)piperazine-1-carboxylate (10n)

Methods A and E followed by F

Very viscous liquid, 69% ^{1}H NMR (500 MHz, $CDCl_3$) δ 7.86 (sl, 1H); 7.56 (d, J = 2.4 Hz, 1H); 7.37 (dd, J = 8.9, 2.4 Hz, 1H); 6.89 (d, J = 9.0 Hz, 1H); 4.74 (d, J = 15.3 Hz, 1H); 4.67 (d, J = 15.3 Hz, 1H); 3.89 (s, 3H); 3.43 (m, 5H); 2.62 (m, 4H); 1.65 (m, 3H); 1.44 (s, 9H); 0.95 (dd, J = 12.3, 5.9 Hz, 6H), ^{13}C NMR (125 MHz, $CDCl_3$) δ 170.7; 164.9; 154.3; 152.4; 130.1; 122.8; 122.7; 120.0; 112.3; 79.8; 65.5; 62.6; 56.4; 49.3; 37.1; 28.4; 25.0; 22.5.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{24}H_{36}O_6N_3ClNa$ 520.21848; found 520.21832.

General Procedure for the Preparation of Compounds 3k' and 10o

The esters, obtained through methods A and D or E followed by F, were dissolved in dry DCM, and a large excess of HCl (4 M solution in dioxane) was added. The resulting solution was stirred at room temperature for 4 h. The solvent mixture was removed under reduced pressure, and the residue was dissolved in DCM followed by crystallization from hexane.

2-((4-(methylamino)phenyl)amino)-2-oxoethyl 2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3k')

Very viscous liquid, 75% ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.04 (sl, 1H); 7.53 (d, J =8.8 Hz, 2H); 6.60 (d, J = 8.8 Hz, 2H); 5.89 (m, 2H); 5.02 (d, J = 15.3 Hz, 1H); 4.87 (dd, J = 11.1, 4.3 Hz, 1H); 4.40 (d, J = 15.4 Hz, 1H); 3.72 (sl, 1H); 3.20 (m, 2H); 2.83 (s, 3H); 2.64 (m, 2H); 2.25 (m, 2H); 2.11 (m, 1H); 1.94 (m, 1H); 1.40 (m, 1H); 0.92 (d, J = 6.6 Hz, 6H), ^{13}C NMR (125 MHz, $CDCl_3$) δ 180.3; 180.1; 168.0; 164.3; 146.6; 127.7; 121.7; 112.5; 63.5; 51.1; 39.2; 37.1; 31.0; 24.8; 23.55; 23.48; 23.0; 20.9.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{23}H_{29}O_5N_3FNa$ 450.19994; found 450.19913.

2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl (S)-4-methyl-2-(piperazin-1-yl)pentanoate hydrochloride (10o)

Amorphous solid, 96% ^{1}H NMR (500 MHz, MeOD) δ 7.67 (m, 1H); 7.38 (m, 1H); 7.02 (m, 1H); 4.83 (m, 2H); 3.98 (m, 1H); 3.85 (s, 3H); 3.78 (m, 1H); 3.46 (m, 6H); 3.27 (m, 1H); 1.82 (m, 3H); 1.00 (dd, J = 11.1, 6.2 Hz, 6H), ^{13}C NMR (125 MHz, $CDCl_3$) δ 172.3; 170.5; 167.3; 167.2; 156.2; 153.9; 132.7; 123.9; 123.4; 121.4; 114.1; 74.4; 68.3; 66.9; 64.5; 64.4; 57.2; 47.8; 44.5; 44.0; 41.0; 38.2; 26.4; 26.1; 23.4; 23.1; 22.4; 22.2.

HRMS (ESI-Orbitrap): $[M+H]^+$ Calculated for $C_{19}H_{28}O_4N_3Cl$ 398.18466; found 398.18384.

General Procedure for the Preparation of Compounds 3l' and 3m'

The esters, obtained through methods A and D followed by F, were dissolved in methanol, followed by the addition of the catalyst Pd/C (10 wt%, 0.1 equivalents). The mixture stirred at room temperature under a hydrogen atmosphere (1 atm) for 4 h and then filtered through Celite®. The filtrate was concentrated

in vacuum. The product was purified by column chromatography on silica gel using a mixture of ethyl acetate/hexanes (50%) as the eluent.

2-((4-aminophenyl)amino)-2-oxoethyl 2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3l')

Pale amorphous solid, 92% ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.06 (sl, 1H); 7.51 (d, J = 8.7 Hz, 2H); 6.67 (d, J = 8.7 Hz, 2H); 5.89 (m, 2H); 5.03 (d, J = 15.4 Hz, 1H) 4.87 (dd, J = 11.0, 4.3 Hz, 1H); 4.40 (d, J = 15.4 Hz, 1H); 3.65 (sl, 2H); 3.20 (m, 2H); 2.64 (m, 2H); 2.26 (m, 2H); 2.10 (m, 1H); 1.94 (m, 1H); 1.40 (m, 1H); 0.92 (d, J = 6.6 Hz, 6H). ^{13}C NMR (125 MHz, $CDCl_3$) δ 180.4; 180.2; 168.1; 164.4; 143.4; 128.9; 127.7; 121.7; 115.3; 63.5; 51.1; 39.3; 37.1; 24.8; 23.58; 23.50; 23.1; 21.0.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{22}H_{27}O_5N_3Na$ 436.18484; found 436.18421.

2-((2-amino-4-methoxyphenyl)amino)-2-oxoethyl 2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (3m')

Light brown amorphous solid, 94% ^{1}H NMR (500 MHz, $CDCl_3$) δ 7.90 (sl, 1H); 7.08 (d, J = 8.5 Hz, 1H); 6.34 (m, 2H); 5.90 (m, 2H); 4.98 (d, J = 15.6 Hz, 1H); 4.87 (dd, J = 11.0, 4.3 Hz, 1H); 4.54 (d J = 15.4 Hz, 1H); 3.75 (s, 3H); 3.18 (m, 2H); 2.61 (m, 2H); 2.23 (m, 2H); 2.07 (m, 1H); 1.93 (m, 1H); 1.37 (m, 1H); 0.91 (dd, J = 6.6, 3.4 Hz, 6H). ^{13}C NMR (125 MHz, $CDCl_3$) δ 180.12; 180.09; 168.5; 165.6; 159.3; 142.8; 127.71; 127.71; 127.26; 115.5; 104.5; 102.6; 63.8; 55.3; 51.0; 39.21; 39.15; 37.1; 24.7; 23.50; 23.48; 23.0; 20.9.

HRMS (ESI-Orbitrap): $[M+Na]^+$ Calculated for $C_{23}H_{30}O_6N_3$ 444.21627; found 444.21738.

Procedure for the Preparation of N-(2-((3-chloro-4-methoxyphenyl)amino)-2-oxoethyl)-2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanamide (5)

2-(1,3-Dioxo-1,3,3a,4,7,7a-hexahydro-2*H*-isoindol-2-yl)-4-methylpentanoic acid **(1)** was dissolved in dichloromethane at room temperature. Oxalyl chloride (1.5 equivalents) and 0.1 mL of dimethyl formamide (DMF) were added to this solution as the catalyst. After 2 h, the solvent was removed under vacuum. The residue was dissolved in dichloromethane, and glycine (1.0 equivalents) was added, followed by the addition of triethylamine (2.2 equivalents).

The isolated product from the previous step was dissolved in dichloromethane at room temperature. Oxalyl chloride (1.5 equivalents) and 0.1 mL dimethyl formamide (DMF) were added to this solution as the catalyst. After 2 h, the solvent was removed under vacuum. The residue was dissolved in dichloromethane, and 3-chloro-4-methoxyaniline (1.0 equivalents) was added, followed by the addition of triethylamine (2.2 equivalents). The reaction stirred for an additional 2 h. The reaction mixture was washed with water and saturated aqueous NH_4Cl solution. The organic phase was dried over $MgSO_4$ and the solvent was removed under reduced pressure. The product was purified by column chromatography on silica gel using a mixture of ethyl acetate/hexanes (50%) as the eluent.

Amorphous solid, 78% ^{1}H NMR (250 MHz, DMSO) δ 9.68 (sl, 1H); 7.64 (d, J = 2.4 Hz, 1H); 7.39 (dd, J = 8.9, 2.4 Hz, 1H); 7.09 (d, J = 9.0 Hz, 1H); 5.86 (m, 2H) 4.65 (dd, J = 11.3, 4.3 Hz, 1H); 3.81 (sl, 3H); 3.18 (m, 2H); 2.39 (m, 2H); 2.11 (m, 3H); 1.79 (m, 1H); 1.29 (m, 1H); 0.83 (m, 6H). ^{13}C NMR (62.5 MHz, DMSO) δ 180.1; 179.8; 166.8; 150.8; 132.1; 127.7; 127.6; 122.0; 120.4; 120.2; 112.7; 56.2; 52.4; 36.2; 24.3; 23.4; 23.3; 23.1; 20.6.

Procedure for the Preparation of 2-((3-chloro-4-methoxyphenyl)amino)ethyl 2-(1,3-dioxo-1,3,3a,4,7,7a-hexahydro-2H-isoindol-2-yl)-4-methylpentanoate (7)

A solution of compound **1** in dry DCM was prepared, and alcohol **6** was added (1.1 equivalents). The solution was cooled in an ice bath followed by the addition of EDC (1.2 equivalents) and DMAP (0.1 equivalents). The solution was allowed to warm to room temperature and stirred for 4 h while monitoring periodically by TLC. The reaction mixture was washed with saturated aqueous NH_4Cl solution and brine. The organic phase was dried over $MgSO_4$ and concentrated under reduced pressure. The product was purified by column chromatography on silica gel using a mixture of ethyl acetate/hexanes (50%) as the eluent.

Pale oil, 62% ^{1}H NMR (500 MHz, $CDCl_3$) δ 6.82 (d, J = 8.7 Hz, 1H); 6.70 (d, J = 2.8 Hz, 1H); 6.51 (dd, J = 8.8, 2.8 Hz, 1H); 5.89 (m, 1H); 5.82 (m, 1H); 4.76 (dd, J = 11.3, 4.3 Hz, 1H); 4.43 (ddd, J = 11.1, 7.1, 3.9 Hz, 1H); 4.18 (m, 1H); 3.82 (m, 3H); 3.30 (m, 2H); 3.10 (m, 2H); 2.60 (m, 2H); 2.23 (m, 2H); 2.09 (m, 1H); 1.84 (m, 1H); 1.35 (m, 1H); 0.89 (dd, J = 8.8, 6.5 Hz, 6H). ^{13}C NMR (125 MHz, $CDCl_3$) δ 179.64; 179.61; 169.1; 147.6; 142.5; 127.7; 127.6; 123.4; 115.2; 114.2; 112.1; 64.0; 57.0; 51.2; 43.1; 39.0; 38.9; 36.7; 24.8; 23.43; 23.42; 23.1; 21.0. HRMS (ESI-Orbitrap) m/z [M + Na] calculated for $C_{23}H_{29}O_5N_2ClNa$ 471.16572; found 471.16522.

RESULTS AND DISCUSSION

In this work, we report the discovery of a series of reversible cruzain inhibitors showing promising trypanocidal activity and low toxicity. The imide derivative **3a** (**Figure 1A**), a reversible cruzain inhibitor that was previously identified from a virtual high-throughput screening (HTS) combined approach (Ferreira et al., 2010), was taken as the initial hit for molecular optimization. The SBDD strategy relied on the molecular docking-predicted binding mode of **3a** within the catalytic site of cruzain (**Figure 1B**). The enzyme-inhibitor interactions rely mainly on a hydrogen bonding network. The secondary amide oxygen fills the S3 subsite and interacts with the main chain nitrogen of Gly66. The secondary amide nitrogen in turn forms a hydrogen bond with the main chain carbonyl of Asp161, which lies at the interface between the S2 and S1' subsites. The ester carbonyl projects into the S1 subsite where it engages in a hydrogen bond with the side chain nitrogen of Gln19. One of the imide carbonyl oxygens forms

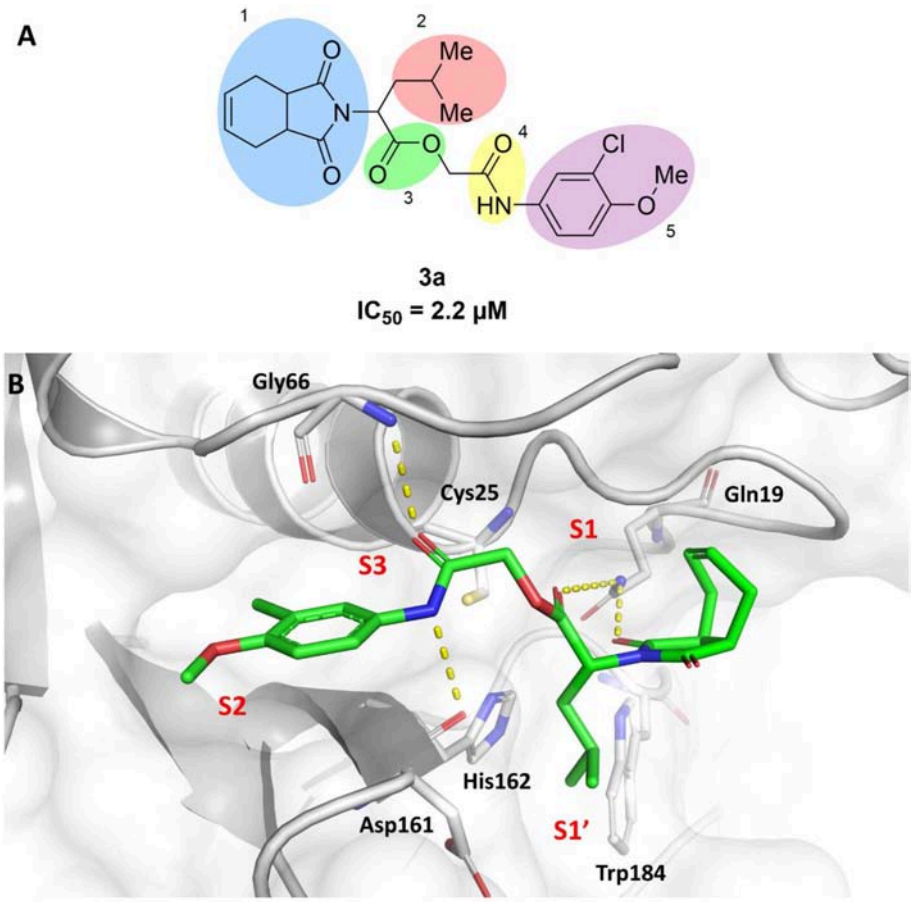


FIGURE 1 | (A) Compound **3a** was used as the initial hit for the design of imide derivatives as novel cruzain inhibitors with anti-T. cruzi activity. **(B)** Molecular docking-predicted binding conformation of compound **3a** within the active site of cruzain. Cruzain (PDB 3KKU, 1.28 Å) is depicted in cartoon and surface representations. Binding site residues (carbon in gray) and compound **3a** (carbon in green) are shown as sticks. Hydrogen bonds are shown as dashed lines. Cruzain subsites are labeled as S1, S1', S2, and S3.

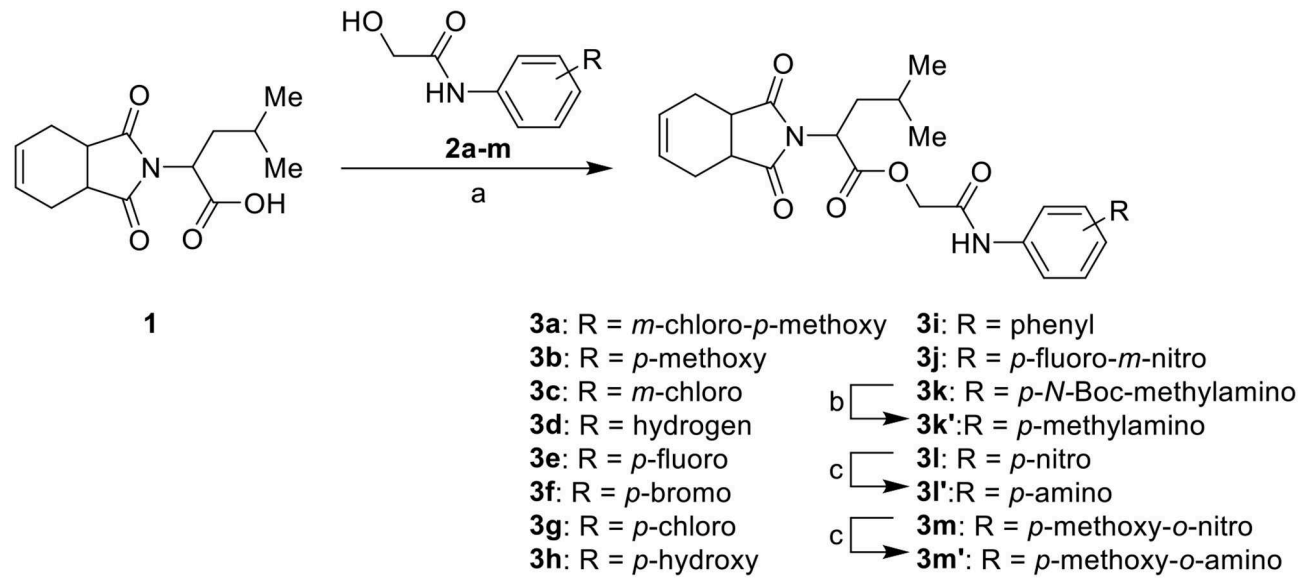
a hydrogen bond with the side chain nitrogen of Gln19 at the S1 subsite. The 3-chloro-4-methoxyphenyl ring projects into the S2 subsite and the isobutyl fragment lies in S1'. Based on these findings, compound **3a** was divided into five fragments that underwent the molecular modifications that eventually led to the series of cruzain inhibitors described herein. The data gathered from the designed analogs were used to further improve the potency against both cruzain and *T. cruzi* and disclose the structure-activity relationships (SAR) enclosed in this dataset.

Synthesis of Imide Derivatives

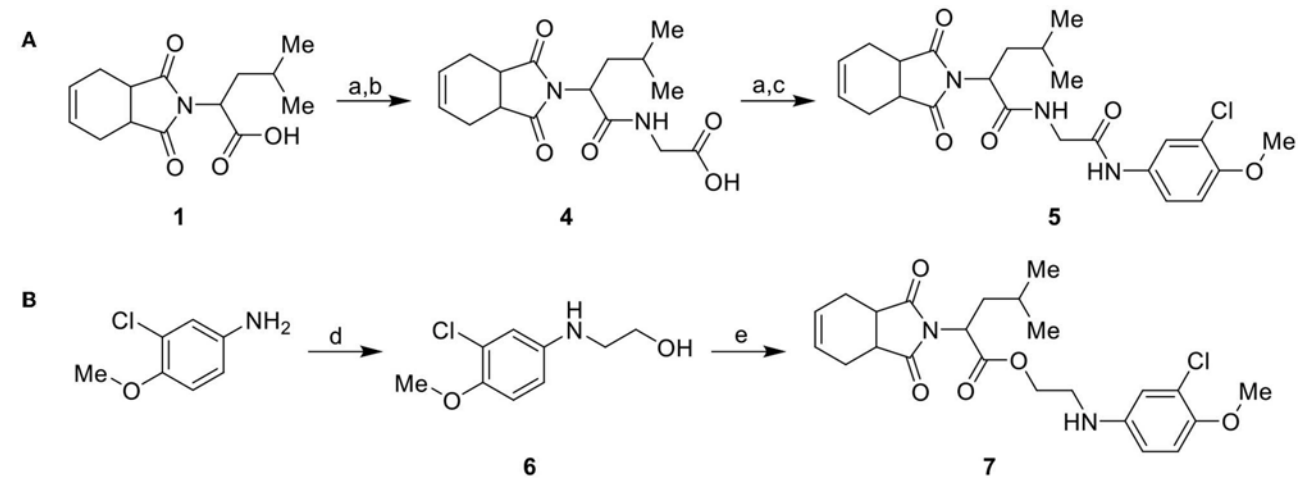
The synthesis of the initial hit (**3a**) and analogs with modifications to the aromatic moiety is outlined in **Scheme 1**. The aromatic ring was manipulated by introducing electron withdrawing and donating groups and hydrogen bond donors and acceptors. Carboxylic acid **1** was coupled under Steglich-like conditions (Pande et al., 2014) with alcohols **2a-m** furnishing the intended esters **3a-m**. Ester **3k'** was obtained through deprotection of **3k** under acidic conditions, and compounds **3l'** and **3m'** were generated from the reduction of **3l** and **3m**, respectively.

Considering the low stability of the ester group in acidic and basic media, we synthesized a more stable analog with an amide group, as illustrated in **Scheme 2**. Initially, carboxylic acid intermediate **1** was transformed into an acyl chloride via reaction with oxalyl chloride. The acyl chloride was further used in a coupling reaction with glycine, producing intermediate **4**. Intermediate **4** was converted into an acyl chloride, using the same conditions previously mentioned, and coupled with 3-chloro-4-methoxyaniline to generate the target compound **5** (**Scheme 2A**). Compound **7** (**Scheme 2B**) was synthesized from 3-chloro-4-methoxyaniline with ethyl carbonate catalyzed by molecular sieves (Kinage et al., 2011) to produce **6**, followed by coupling with **1** using Steglich conditions.

Further analogs of compound **3a** were designed to explore modifications to the imide nucleus and the hydrophobic isobutyl region (R^1). The synthesis was performed according to **Scheme 3**

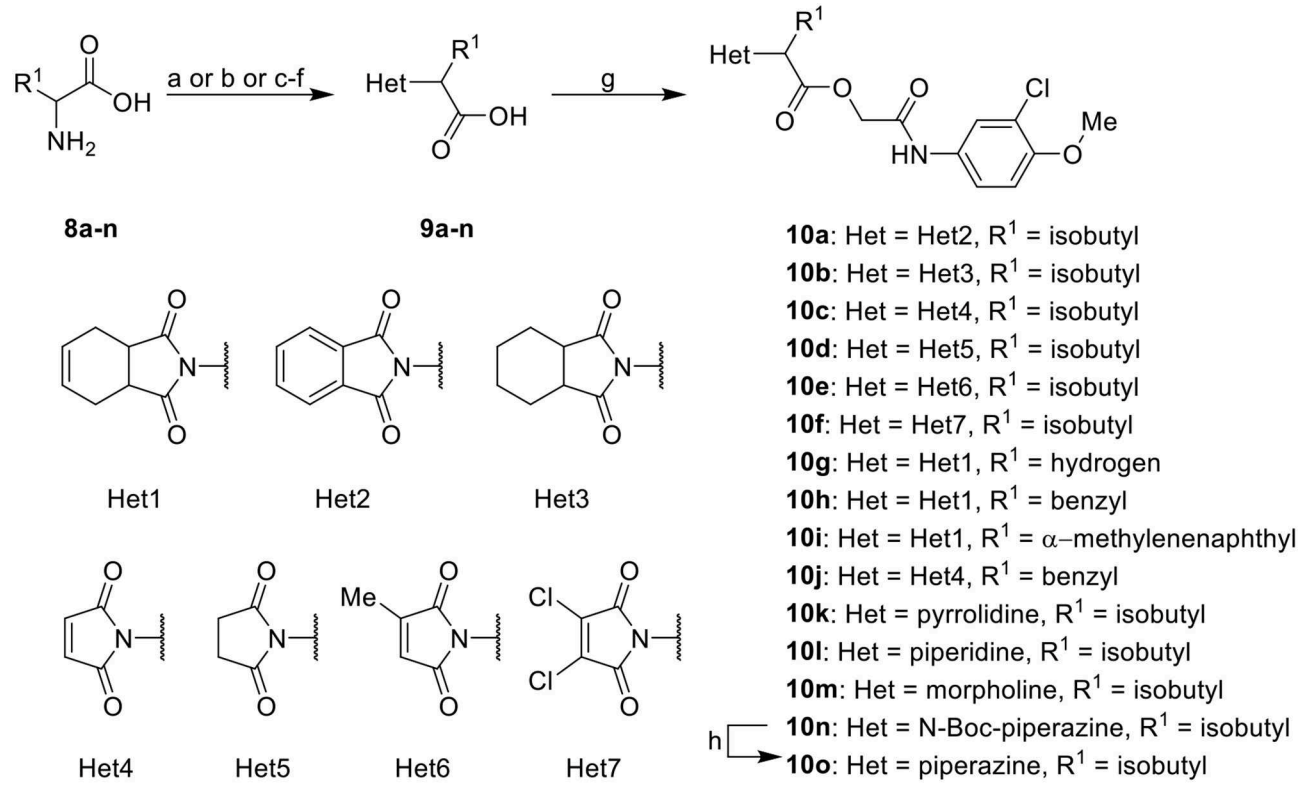


SCHEME 1 | Synthesis of initial hit **3a** and analogs with modifications to the aromatic ring. Reagents and conditions: (a) CH_2Cl_2, **1**, EDC, DMAP (cat), rt, 4 h; (b) CH_2Cl_2, HCl (4 M in 1,4-dioxane), rt, 4 h; (c) MeOH, H_2 1 atm, 10% Pd/C, rt, 30 min.



SCHEME 2 | Synthesis of analogs of the initial hit **3a** with modifications to the ester and amide functional groups. **(A)** Synthesis of compound **5** by coupling intermediate **4** with 3-chloro-4-methoxyaniline. **(B)** Synthesis of compound **7** by coupling compoud **6** with **1**. Reagents and conditions: (a) CH_2Cl_2, oxalyl chloride, DMF (cat), rt, 2 h; (b) CH_2Cl_2, glycine, Et_3N, rt, 2 h; (c) CH_2Cl_2, 3-chloro-4-methoxyaniline, Et_3N, rt, 2 h; (d) DMF, ethylene carbonate, molecular sieve 4 Å (cat), 150°C, 4 d; (e) CH_2Cl_2, **1**, EDC, DMAP (cat), rt, 4 h.

to evaluate the effects of changing the stereochemical and electronic features on the biological activity of the target compounds. Intermediates **9a-c** and **9g** were prepared according to a previously described procedure (Faghihi, 2008). A mixture of readily available *L*-amino acids and the corresponding cyclic anhydride in acetic acid stirred overnight, followed by 4 h of reflux to obtain carboxylic acids **9a-c** and **9g**. The synthesis of compounds **9d-f** and **9h-j** was carried out by refluxing in toluene with triethylamine as the catalyst. The stereocenter of the carboxylic acids was observed to have the absolute configuration of *R* (Pande et al., 2014). The amino acid side chains gave rise to hydrophobic fragments. Diazotization of *L*-leucine (Badiola et al., 2014) followed by bromoacid formation and subsequent nucleophilic substitution with the appropriate amines generated the amine nucleus in intermediates **9k-n**, presumably with an inverted configuration of the stereogenic center (*S*) due to inversion during the nucleophilic substitution reaction. Intermediates **9a-n** were coupled under Steglich conditions (Neises and Steglich, 1978) to produce the intended esters **10a-n**. Compound **10o** was produced by deprotection of the N-Boc analog (**10n**) under acidic conditions.



SCHEME 3 | Synthesis of analogs of compound **3a** with changes to the imide and isobutyl moieties. Reagents and conditions: (a) (for compounds **9a-c** and **9g**) AcOH, cyclic anhydride (correspondent to Het), rt, o.n., then reflux, 4 h; (b) (for compounds **9d-f**, **9h-j**) toluene, cyclic anhydride (corresponding to Het), Et_3N (cat), reflux, 3 h; (c) (for compounds **9k-n**) H_2O, HBr (48% v/v in H_2O), $NaNO_2$, 0°C to rt, 3 h; (d) CH_2Cl_2, BnOH, EDC, DMAP (cat), rt, 4 h; (e) MeCN, amine (corresponding to Het), Cs_2CO_3, rt – 60°C, 30 min – o.n.; (f) MeOH, H_2 1 atm, 10% Pd/C, rt, 30 min; (g) CH_2Cl_2, **2a**, EDC, DMAP (cat), rt, 4 h; (h) CH_2Cl_2, HCl (4 M in 1,4-dioxane), rt, 4 h.

Identification of Imide Derivatives as Cruzain Inhibitors

The activities of inhibitor **3a** (IC_{50} = 2.2 μM) and its analogs having modifications to the imide moiety against cruzain are shown in **Figure 2**. The incorporation of an aromatic ring into the imide ring to form an isoindoline-1,3-dione core in compound **10a** reduced the percent inhibition from 88 to 39% at an inhibitor concentration of 100 μM. In fact, compound **10a** assumes a completely different binding conformation, with the imide projecting into the S2 subsite. This causes the loss of key interactions with the enzyme as predicted by the molecular docking simulations (**Figure S1A**). The coupling of a saturated ring to the imide resulted in a 2-fold decrease in the potency of compound **10b** (IC_{50} = 4.2 μM) over **3a**. The addition of unsaturation to the unsubstituted imide yielded compound **10c** (IC_{50} = 2.3 μM), which was equipotent with respect to inhibitor **3a**. In fact, compound **10c** preserves the enzyme-inhibitor interactions observed for **3a**, however, the imide carbonyl oxygen is predicted to interact with the side chain nitrogen of Trp184 instead of Gln19. The predicted binding modes of analogs **10b** and **10c** are shown in **Figures S1B,C**, respectively). Withdrawing the unsaturation of the imide produced compound **10d** (IC_{50} = 12.0 μM), which was 6-fold less active than **3a**.

Following the positive result observed for compound **10c**, we designed compounds **10e** and **10f**, which have the imide nucleus incorporating an unsaturation but with two different substitution patterns. Compound **10e**, with a methyl substituent, was equipotent (IC_{50} = 2.5 μM) relative to **10c**, whereas compound **10f**, featuring two chlorine atoms as substituents, resulted in a 6-fold increase in potency (IC_{50} = 0.60 μM). Replacing the imide core with an amine (**10k**, **10l**, **10m**, **10n**, and **10o**) led to a significant drop in the activity of the compounds, highlighting the essentiality of the carbonyl groups. In fact, **Figure 1B** shows a hydrogen bond between one of the carbonyl groups and the side chain nitrogen of Gln19 at the S1 subsite. Replacing the imide group with other ring systems, such as pyrrolidine, piperidine, and morpholine, abrogated the biological activity of this series. The predicted binding conformations of **10k** and **10l** (pyrrolidine and piperidine derivatives, respectively) show that the lack of the imide carbonyl groups changes the binding pattern of these compounds compared to that observed for the active analogs. As shown in **Figure S2**, the isobutyl and cyclic amines do not interact with

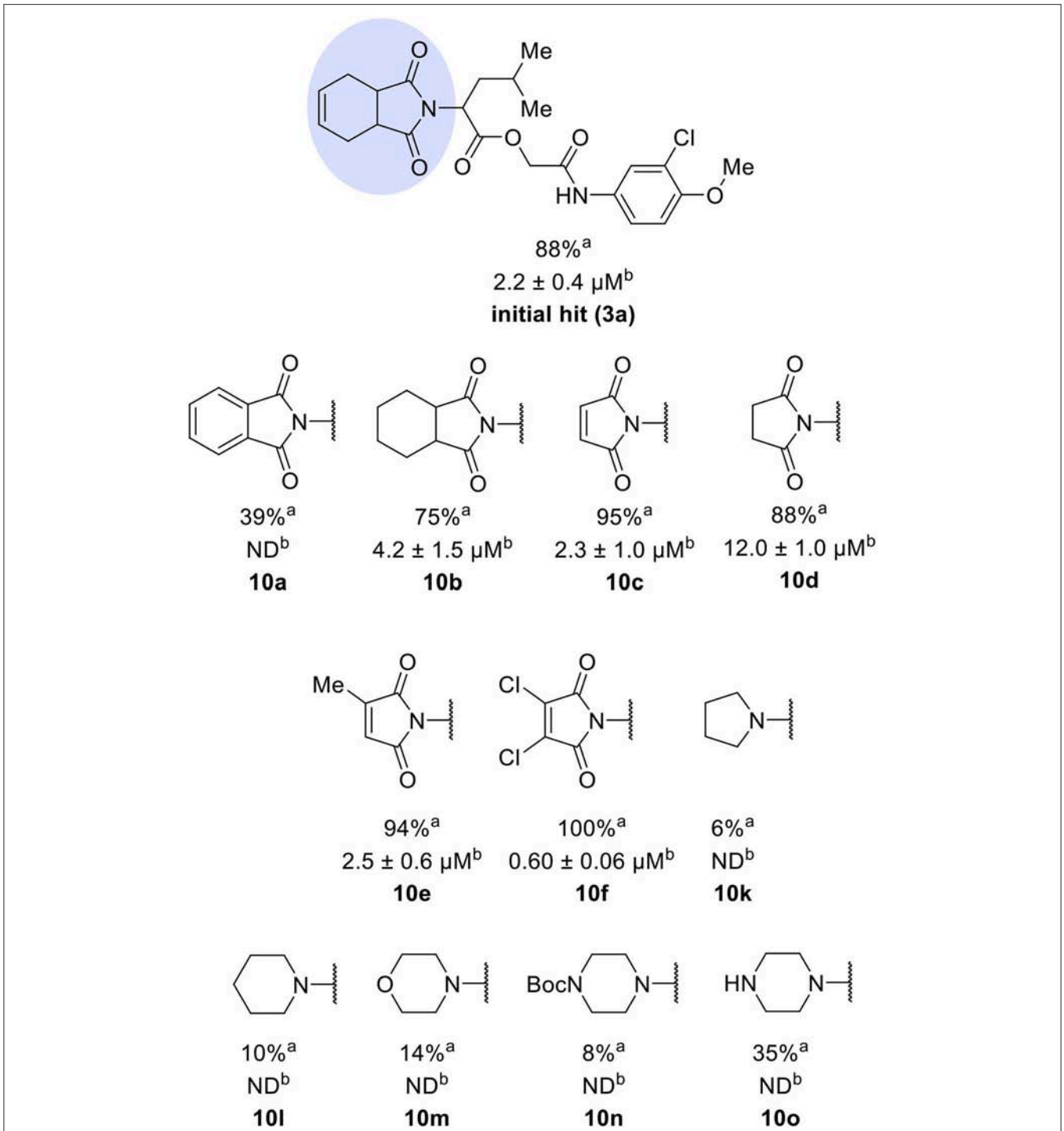


FIGURE 2 | Activity of the series of analogs featuring modifications in the imide nucleus against cruzain. [a]Percent inhibition against cruzain at 100 μM corresponding to the mean of three measurements. [b]IC_{50} values against cruzain, which were independently determined by obtaining rate measurements in triplicate for at least six inhibitor concentrations. The IC_{50} values represent the mean ± SD of three individual experiments. ND, IC_{50} not determined.

the S1' and S1 subsites as expected, and become exposed to the solvent.

The molecular optimization strategy also involved modifications to the hydrophobic (isobutyl) fragment of compound **3a** (**Figure 3A**). Removing the isobutyl group to obtain a methylene as the linker between the ester and the imide rendered the resulting compound inactive (**10g**). Bulky hydrophobic groups proved to be essential for interaction with

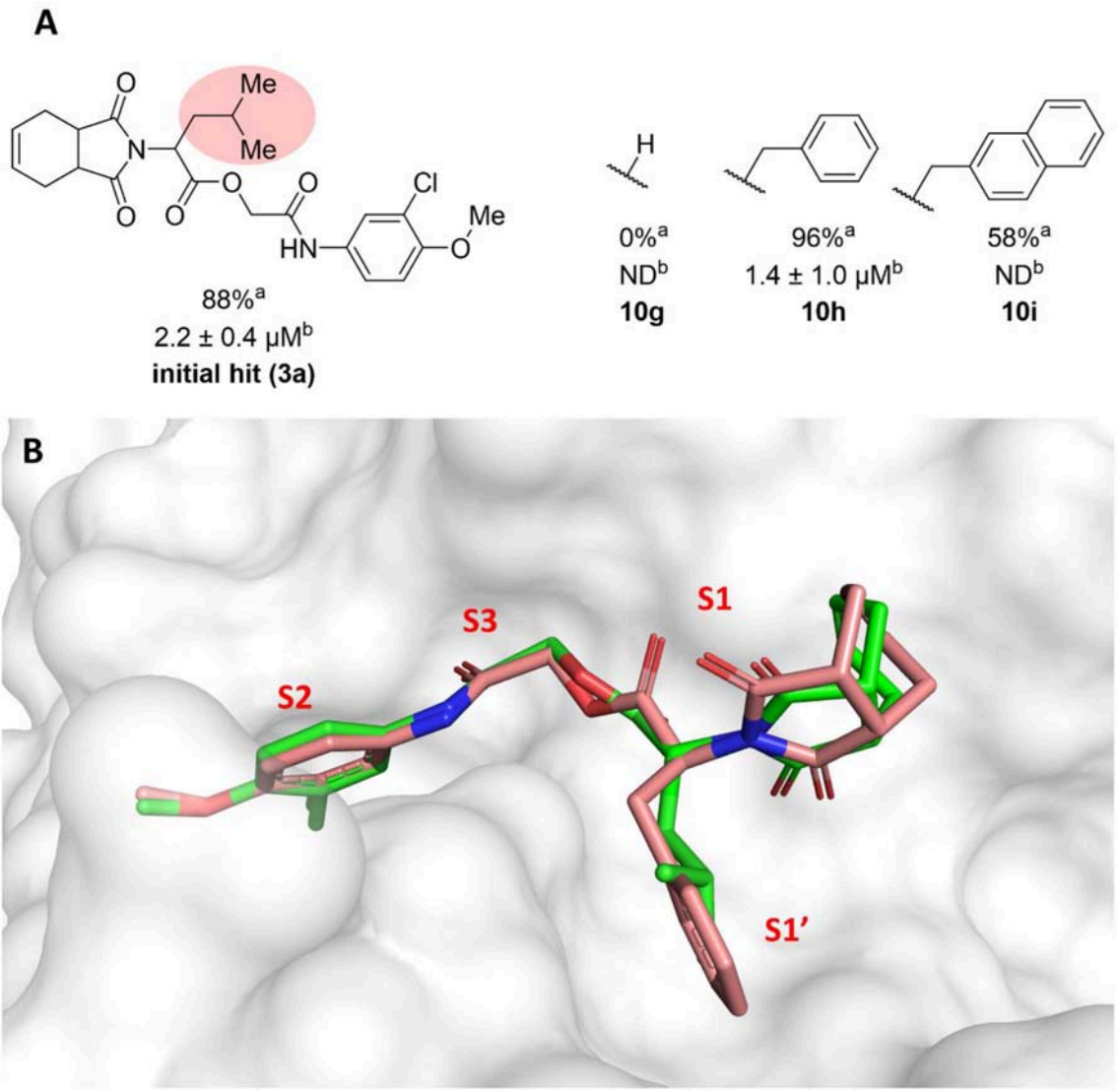


FIGURE 3 | (A) Results for modifications of the isobutyl fragment. [a]Percent inhibition against cruzain at 100 μM corresponding to the mean of three measurements. [b]IC_{50} values against cruzain, which were independently determined by obtaining rate measurements in triplicate for at least six inhibitor concentrations. The IC_{50} values correspond to the mean ± SD of three individual experiments. ND, IC_{50} not determined. **(B)** Superposition of the molecular docking-predicted binding conformation of compounds **3a** (carbon in green) and compound **10 h** (carbon in salmon). Cruzain (PDB 3KKU, 1.28 Å) is depicted in surface representation. Inhibitors **3a** and **10 h** are shown as sticks. Cruzain subsites are labeled as S1, S1', S2, and S3.

the S1' subsite. Replacing the isobutyl with a methylene led to a completely different binding pattern from that observed for the active compounds. The docking algorithm was unable to place the imide into S1 and the aromatic ring into S2, which rendered compound **10g** inactive (**Figure S3A**). Replacement of the isobutyl by the bulkier and planar benzyl group increased the potency by 2-fold (**10h**, $IC_{50} = 1.4\,\mu M$) compared to compound **3a**. This activity improvement can be reasoned to be a result of the better complementarity of the benzyl group with the S1' subsite of the enzyme (**Figure 3B**). Replacing the benzyl with the bulkier methylene naphthyl moiety led to compound **10i**, which was significantly less active, suggesting the ideal volume of a benzyl ring to occupy the S1' subsite. In fact, the naphthyl group of analog **10i** projects into the S2 subsite, which can be the driving force for the percent inhibition value (58%) determined for this compound since S2 usually accommodates bulky hydrophobic groups (**Figure S3B**). However, this modification caused the imide to lose its interaction with S1, which is a key driving force for the activity of this series.

To probe the relevance of the ester and amide groups for the activity of cruzain inhibitors, compounds **5** and **7** were prepared (**Figure 4A**). Replacing the ester with an amide, resulting in compound **5**, caused a reduction in the percent inhibition from 88 to 38%. Converting the amide carbonyl

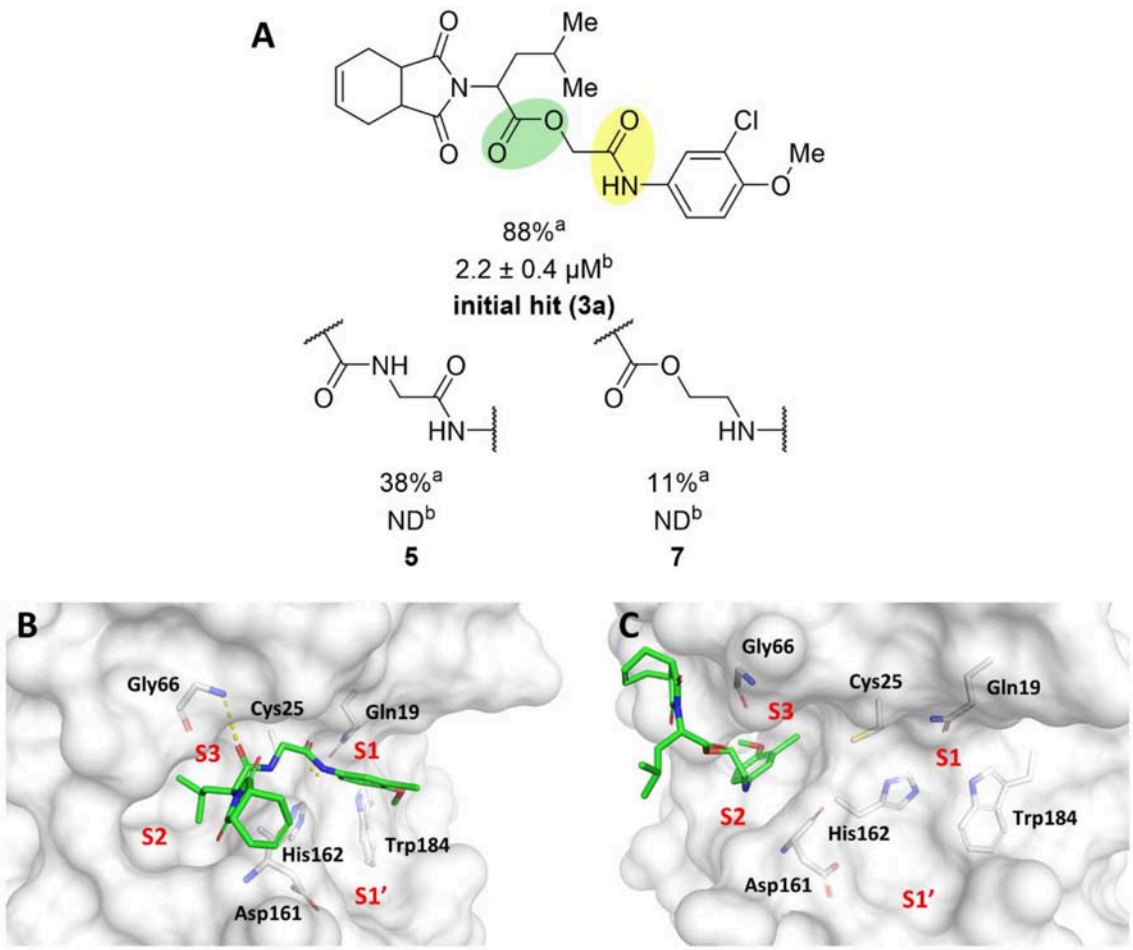


FIGURE 4 | (A) Results for the modifications to the ester and secondary amide functions. [a]Percent inhibition against cruzain at 100 μM corresponding to the mean of three measurements. [b]IC_{50} values against cruzain, which were independently determined by obtaining rate measurements in triplicate for at least six inhibitor concentrations. The IC_{50} values correspond to the mean ± SD of three individual experiments. ND: IC_{50} not determined. **(B)** Molecular docking predicted binding mode of compound **5**. **(C)** Predicted binding mode of compound **7**. Cruzain (PDB 3KKU, 1.28 Å) is depicted in surface representation. Binding site residues (carbon in gray) and inhibitors (carbon in green) are shown as sticks. Hydrogen bonds are shown as dashed lines. Cruzain subsites are labeled as S1, S1', S2, and S3.

into a methylene decreased the biological activity even further, as observed for compound **7** (11% inhibition). As illustrated in **Figure 1B**, the amide oxygen of compound **3a** is predicted to form a hydrogen bond with the main chain nitrogen of Gly66 and the ester carbonyl interacts with Gln19. The loss of these interactions may be one of the reasons for the significant activity drop observed for compound **7**. In fact, modifying the ester and amide groups significantly changed the binding mode of both compounds **5** and **7** (**Figures 4B,C**, respectively) compared to that of the active analogs. Compound **5**, which had the two carbonyl groups preserved, forms a hydrogen bond with Gly66 at S3, which can be reasoned as the cause of its higher percent inhibition value (38%) compared to that of analog **7** (11%). Notwithstanding, the imide of both compounds is exposed to the solvent and no interaction with S1' is observed, which are critical detrimental features of analogs **5** and **7**.

The leverage of the 3-chloro-4-methoxyphenyl moiety on cruzain activity was assessed by modifying the substitution pattern of the aromatic ring (**Figure 5**). Removing either chlorine or methoxyphenyl led to the less active compounds **3b** (IC_{50} = 13.9 μM) and **3c** (IC_{50} = 40.1 μM), respectively. Compound **3b** retains the key polar interactions with the enzyme, however, the lack of the chlorine prevents an optimal interaction with the S2 pocket, which is known to be essential for potent cruzain inhibition as reported previously (Ferreira et al., 2014) (**Figure S4A**). The effect of removing the methoxy group from the S2-interacting aromatic ring proved to be more detrimental to activity than the removal of the chlorine. In fact, compound **3c** was predicted to lose most of the key polar interactions with the enzyme (**Figure S4B**), which resulted in a poorly active compound (IC_{50} = 40.1 μM). The unsubstituted phenyl derivative (**3d**) was 8-fold less active (IC_{50} = 16.7 μM) than the parent 3-chloro-4-methoxyphenyl compound (**3a**). Additional

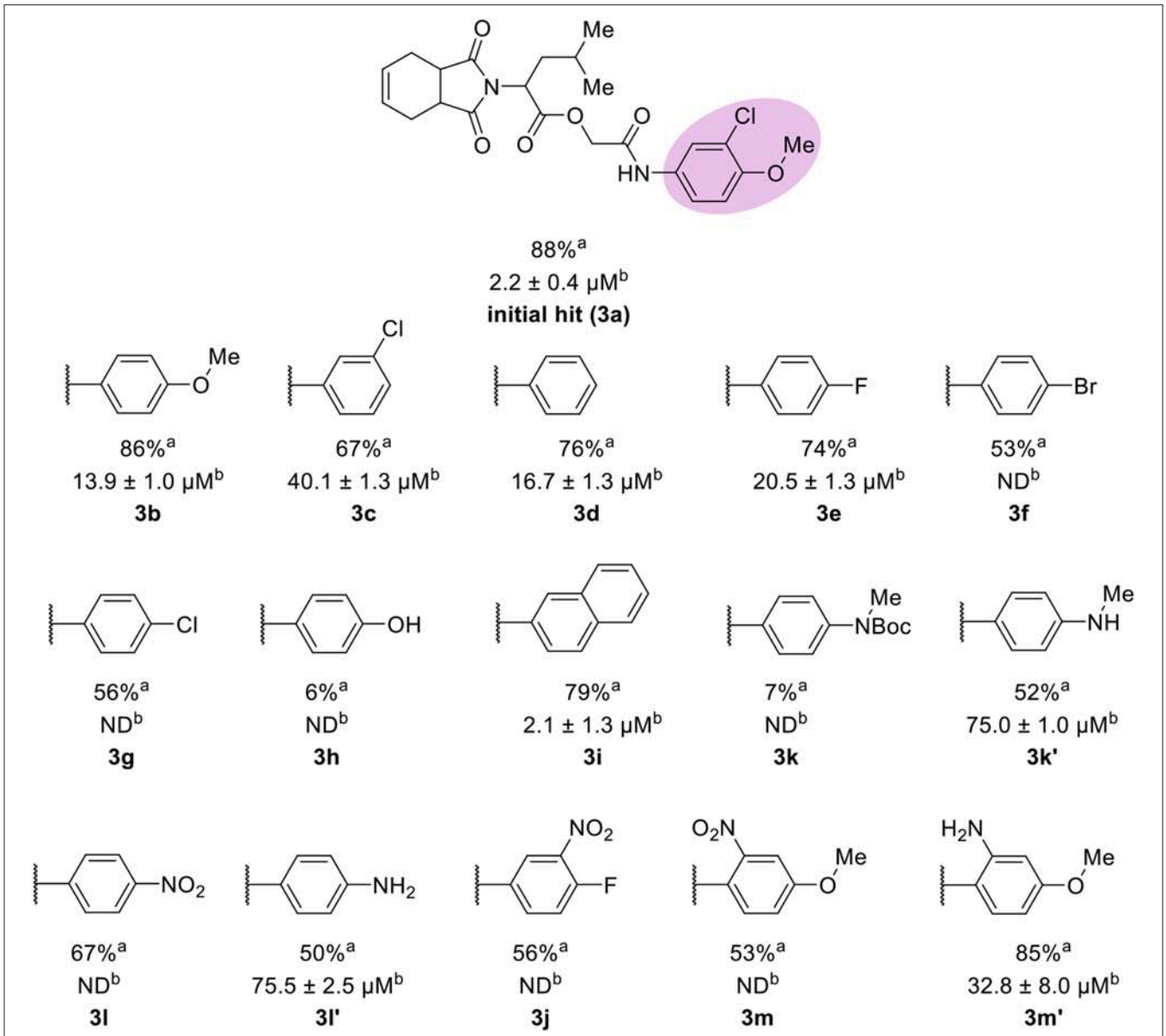
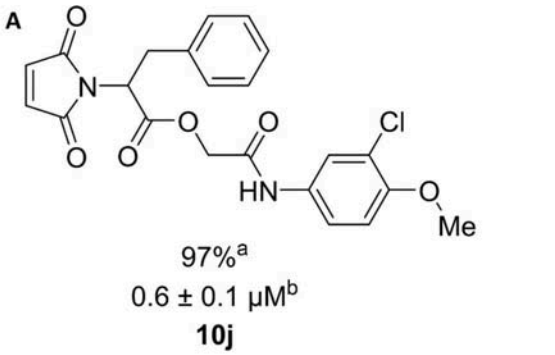


FIGURE 5 | Results for the modifications to the aromatic ring. [a]Percent inhibition against cruzain at 100 μM corresponding to the mean of three measurements. [b]IC_{50} values against cruzain, which were independently determined by obtaining rate measurements in triplicate for at least six inhibitor concentrations. The IC_{50} values correspond to the mean ± SD of three individual experiments. ND, IC_{50} not determined.

manipulations, such as introducing hydrogen bond donors and acceptors and replacing the chlorine or changing its position, led to poorly active or inactive compounds (**3e-h**, **3k**, **3k'**, **3l**, **3l'**, **3j**, **3m,** and **3m'**). **Figures S4** and **S5** show that a suboptimal interaction between the aromatic ring and the S2 subsite lead to the loss of most, if not all, polar contacts that proved to be important for activity. The exception was compound **3i**, with naphthyl replacing the phenyl as the S2-interacting aromatic system (**Figure S5D**). This is consistent with the binding mode of these compounds, which shows that the aromatic moiety interacts with the S2 subsite (**Figure 1B**). The S2 subsite, which is mostly composed of hydrophobic amino acids, can accommodate bulky groups, as shown in our previous work on benzimidazole derivatives (Ferreira et al., 2014). Such groups can therefore promote an optimum interaction with the S2 subsite, leading to more potent cruzain inhibitors.

Since all compounds described so far were obtained from enantiopure amino acids, we next evaluated how the absolute configuration influences the activity against cruzain. To this end, the racemic form of compound **10h** was prepared (**10h**$_{rac}$). The synthesis of **10h**$_{rac}$ was achieved by the same procedure presented in **Scheme 3**, using *DL*-phenylalanine

instead of *L*-phenylalanine. Chiral column HPLC and optical rotation analyses confirmed that all reactions did not lead to racemization when *L*-phenylalanine was used as the starting material. These analyses also confirmed that the racemic compound was obtained when *DL*-phenylalanine was used as the starting material. The racemate (**10h**$_{rac}$) exhibited an IC_{50} value of 1.16 μM against cruzain, while compound **10h** had an IC_{50} of 1.4 μM. The very close IC_{50} values obtained for both the racemate and the enantiopure compound demonstrated that the absolute configuration of the stereocenter is not relevant to the activity of this series of compounds.



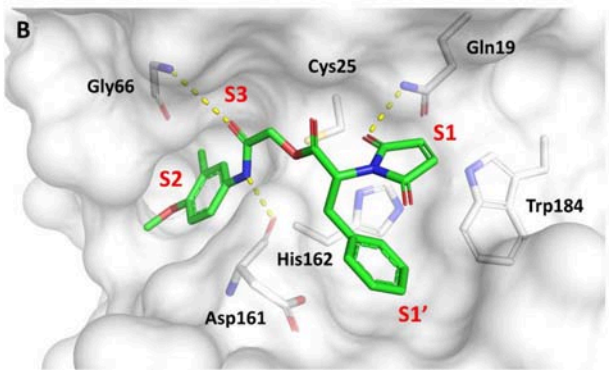


FIGURE 6 | (A) Modifications to the imide nucleus and hydrophobic fragment. [a]Percent inhibition against cruzain at 100 μM corresponding to the mean of three measurements. [b]IC_{50} values against cruzain, which were independently determined by obtaining rate measurements in triplicate for at least six inhibitor concentrations. The IC_{50} values correspond to the mean ± SD of three individual experiments. ND, IC_{50} not determined. **(B)** Molecular docking predicted binding mode of compound **10j**. Cruzain (PDB 3KKU, 1.28 Å) is depicted in surface representation. Binding site residues (carbon in gray) and inhibitors (carbon in green) are shown as sticks. Hydrogen bonds are shown as dashed lines. Cruzain subsites are labeled as S1, S1', S2, and S3.

In the next molecular optimization step, we combined the best fragments from all five regions of initial hit **3a**. This strategy produced the most active compound (**10j**, IC_{50} = 0.6 μM) among the synthesized analogs, which was 4-fold more active than compound **3a** (**Figure 6A**). Compound **10j** conserves the same binding mode of the other active compounds in this series, forming the key intermolecular interactions with the S1, S1', S2, and S3 subsites. The S1' and S2 subsites are optimally filled with the benzyl and 3-chloro-4-methoxyphenyl rings, respectively. A hydrogen bond is formed between the amide oxygen and Gly66, and the imide carbonyl interacts with Gln19 (**Figure 6B**).

The aim of this study was to identify new imide derivatives as competitive cruzain inhibitors. Considering the competitive nature of hit compound **3a**, the synthesized analogs were expected to follow the same behavior. Hence, we conducted further studies to establish the mechanism of inhibition of compounds **10h** and **3i**. The competitive mechanism of these compounds was corroborated by the Lineweaver–Burk plots shown in **Figure 7**. As expected for competitive inhibitors, the maximum velocity ($1/V_{max}$, intersections with the *y*-axis) remained unchanged with increasing inhibitor concentrations [I], while the apparent Michaelis–Menten constant ($K_M^{app} = -1/K_M$, intersections with the *x*-axis) increased with escalating [I].

Discovery of Novel Trypanocidal Agents

The biological activity of 22 compounds was evaluated against *T. cruzi* intracellular amastigotes along with **BZ** (**Tables 1–3**). Several compounds showed trypanocidal activity similar to or superior to that of **BZ** (IC_{50} = 3.0 μM). The data in **Table 1** show that the hit compound **3a**, despite being a potent cruzain

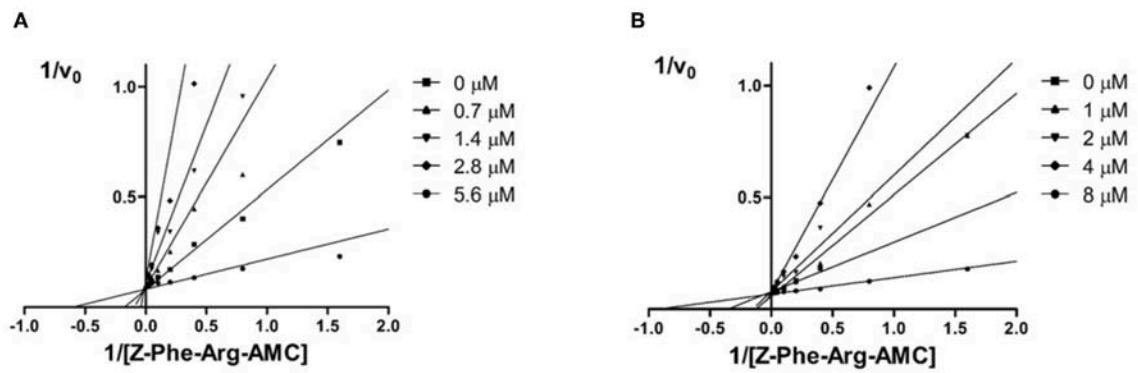
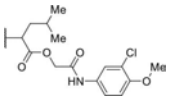


FIGURE 7 | Double reciprocal *Lineweaver-Burk* plots for compounds **10 h (A)** and **3i (B)**. Each curve denotes to a different inhibitor concentration.

TABLE 1 | Cruzain inhibition and trypanocidal activity of compounds with modifications to the imide nucleus.



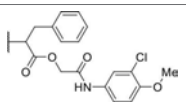
Compound	Structure	IC_{50} cruzain (μM)[a]	IC_{50} *T. cruzi* (μM)[b]
Initial hit (3a)		2.2 ± 0.4	>100
10b		4.2 ± 1.5	>100
10c		2.3 ± 1.0	0.9 ± 0.3
10d		12.0 ± 1.0	18.5 ± 3.1
10e		2.5 ± 0.6	1.2 ± 0.3
10f		0.6 ± 0.1	2.0 ± 0.7
10k		ND	>100
10l		ND	11.6 ± 2.4
10m		ND	1.8 ± 0.4
10n		ND	1.6 ± 0.5
10o		ND	1.7 ± 0.5
BZ	-	ND	3.0 ± 0.6

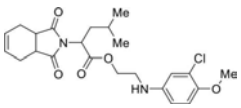
[a]IC_{50} values against cruzain were independently determined by obtaining rate measurements in triplicate for at least six inhibitor concentrations. The values represent the mean ± SD of three individual experiments. [b]IC_{50} values against T. cruzi represent the mean ± SD of two individual experiments. ND, not determined.

inhibitor, was not active against the parasite. The same behavior was observed for compound **10b**, which differs from **3a** only by the absence of a double bond in the ring coupled to the imide. Compounds having the imide with unsaturation and without the coupled six-membered ring showed promising trypanocidal activity (**10c**, **10e,** and **10f**, IC_{50} values of 0.9, 1.2 and 2 μM, respectively).

Compounds lacking the imide carbonyl groups (**10l**-**10o**), although inactive against cruzain, were active against *T. cruzi*. Compound **10k**, in which the imide was replaced with a pyrrolidine, was inactive against both cruzain and *T. cruzi*. Replacing pyrrolidine with piperidine, a larger heterocycle, resulted in compound **10l**, which was moderately active against the parasite (IC_{50} = 11.6 μM). Introducing an additional heteroatom in the 6-membered cycle, whether nitrogen or oxygen, generated remarkably active compounds (**10m**, **10n,** and **10o**, IC_{50} values of 1.8, 1.6, and 1.7, respectively). These results indicate that these cyclic amines exert their trypanocidal activity by modulating a molecular target other than cruzain.

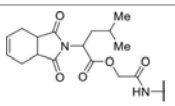
TABLE 2 | Cruzain inhibition and trypanocidal activity of compounds with modifications to the hydrophobic fragment and the amide function.

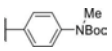
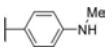
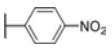
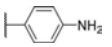
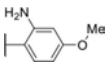


Compound	Structure	IC_{50} cruzain (μM)[a]	IC_{50} *T. cruzi* (μM)[b]
10h		1.4 ± 0.8	>100
10j		0.6 ± 0.1	1.0 ± 0.3
7		ND	1.5 ± 0.4

[a]IC_{50} values against cruzain were independently determined by obtaining rate measurements in triplicate for at least six inhibitor concentrations. The values represent the mean ± SD of three individual experiments. [b]IC_{50} values against T. cruzi represent the mean ± SD of two individual experiments. ND, not determined.

TABLE 3 | Cruzain inhibition and trypanocidal activity of compounds with modifications to the aromatic ring.



Compound	Structure	IC_{50} cruzain (μM)[a]	IC_{50} *T. cruzi* (μM)[b]
3k		ND	3.9 ± 0.9
3k'		75.0 ± 1.0	>100
3l		ND	>100
3l'		75.5 ± 2.5	>100
3m'		32.8 ± 8.0	>100
3d		16.7 ± 1.3	>100
3e		20.5 ± 1.3	>100
3i		2.1 ± 1.3	32.8 ± 4.9

[a]IC_{50} values against cruzain were independently determined by obtaining rate measurements in triplicate for at least six inhibitor concentrations. The values represent the mean ± SD of three individual experiments. [b]IC_{50} values against T. cruzi represent the mean ± SD of two individual experiments. ND, not determined.

Table 2 shows the data for compounds modified on the hydrophobic fragment and the amide group. Replacing the isobutyl with a benzyl group resulted in compound **10h**, which, although active against cruzain, was inactive against the parasite. Notwithstanding, the benzyl group in compound **10j** combined with the maleimide fragment generated the most active

compound against both *T. cruzi* (IC_{50} = 1.0 μM) and cruzain (IC_{50} = 0.6 μM). Compound **7**, in which the amide carbonyl was removed, displayed the opposite behavior to that of compound **10h**, i.e., **7** was active against *T. cruzi* but inactive on cruzain.

As shown in **Table 3**, except for compounds **3k** and **3i**, modifications to the aromatic ring were unfavorable for activity against both *T. cruzi* and cruzain.

Figure 8 illustrates an SAR scheme for the synthesized imide derivatives. Following our initial approach, compound **3a** was divided into five fragments, and the most relevant SARs were identified. In short, the imide function, although required for cruzain inhibition, is not essential for activity against *T. cruzi*. Benzyl is the ideal hydrophobic fragment for activity against cruzain and is tolerable regarding the activity against *T. cruzi*. Replacement of the 3-chloro-4-methoxyphenyl fragment is, in general, unfavorable for activity against *T. cruzi* and cruzain. Removal of the ester group is unfavorable for activity against cruzain, while the replacement of the secondary amide proved to be tolerable regarding the trypanocidal activity. The gathered data showed a lack of a direct correlation between the phenotypic and target-based results for some compounds; this was expected to some degree given the complexity of the intracellular environment and issues related to transport across membranes. For other compounds, however, the target-based and phenotypic data are clearly correlated. Compounds **10c**, **10e**, **10f**, and **10g**, for instance, besides being active against the enzyme, proved to be potent trypanocidal agents.

In vitro Toxicity and Selectivity

The cytotoxicity of the synthesized analogs was assessed using HFF-1 human fibroblasts and HepG2 human hepatocytes (**Table 4**). The selectivity index (SI) was calculated as the ratio between the CC_{50} values for the human cells and the IC_{50} values for *T. cruzi*. In general, the designed compounds demonstrated no significant toxicity against HFF-1 and HepG2 cells. Three compounds showed SI values for HFF-1 fibroblasts comparable or superior to that of **BZ** (SI >33): **10m** (SI >55), **7** (SI >66) and **3k** (SI >25). It is worth noting that these compounds were active against *T. cruzi*. Another aspect worth mentioning is that the most cytotoxic compounds (**10c**, **10e**, and **10j**) are Michael acceptors, which is a possible reason for their cytotoxicity. With respect to cytotoxicity to HepG2 cells, most compounds produced higher SI values than that of **BZ**, particularly those that are highly active against *T. cruzi*: **10c** (SI = 36.6), **10f** (SI >32), **10m** (SI >35), **10o** (SI = 30.6), **10j** (SI = 67), and **7** (SI = 42).

CONCLUSION

The SBDD strategy applied herein, comprising synthetic organic chemistry, molecular docking, enzyme kinetics and phenotypic assays, resulted in the discovery of potent, reversible and non-peptidic cruzain inhibitors with remarkable trypanocidal activity. The success of this experimental-computational molecular optimization approach can be illustrated by compound **10j**, the most potent cruzain inhibitor (IC_{50} = 0.6 μM), which is significantly more active than initial hit **3a** (IC_{50} = 2.2 μM). One of the most potent cruzain inhibitors reported to date, compound **10j** represents a new chemical class among the known inhibitors of this enzyme. Furthermore, compound **10j** shows trypanocidal activity (IC_{50} = 1.0 μM) that is 3-fold higher than that of the clinically used drug **BZ** (IC_{50} = 3.0 μM). Other promising compounds are **10c**, **10f**, **10m**, **10n**, **10o**, **10j**, **7** and **3k**, which showed trypanocidal activity comparable to that of **BZ** and SI values (HFF-1/*T. cruzi*) >10.

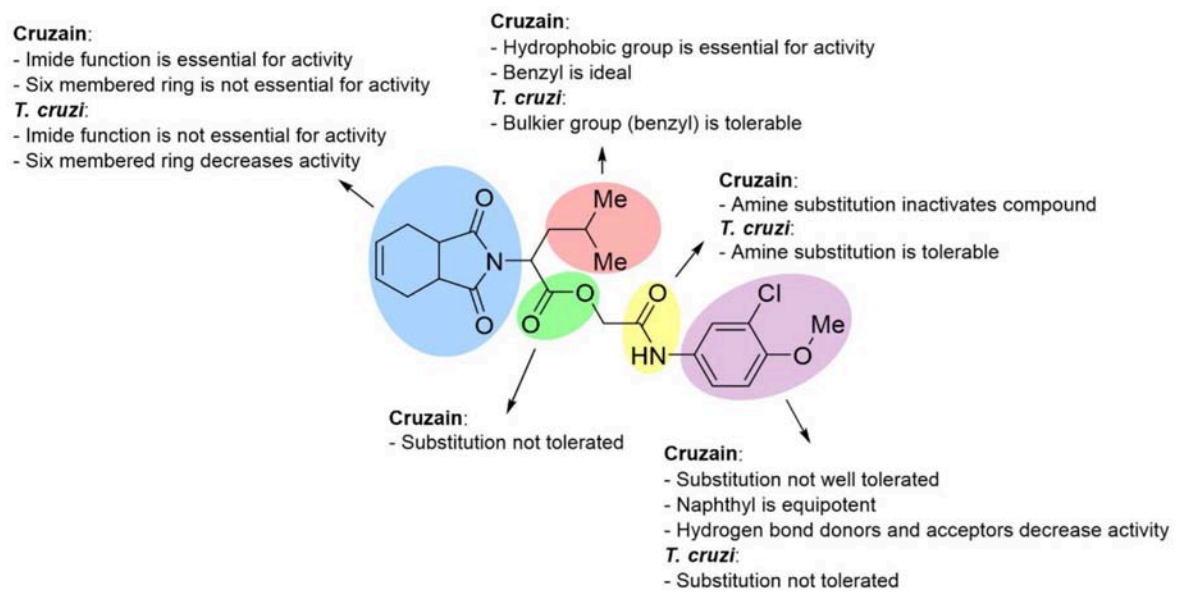


FIGURE 8 | SAR scheme for both cruzain and *T. cruzi*. Compound **3a** was used as the reference structure.

TABLE 4 | Comparison of the trypanocidal activity with cytotoxicity data obtained for HFF-1 and HepG2 cells.

Compound	IC_{50} *T. cruzi* (μM)[a]	CC_{50} HFF-1 (μM)[b]	SI HFF-1/*T. cruzi*[c]	CC_{50} HepG2 (μM)[b]	SI HepG2/*T. cruzi*[c]
BZ	3.0 ± 0.6	>100	>33	>64	>21
10c	0.9 ± 0.3	12 ± 1	13.3	33 ± 5	36.6
10d	18.5 ± 3.1	>100	>5	>64	>3
10e	1.2 ± 0.3	4.8 ± 0.5	4	5 ± 1	4.2
10f	2.0 ± 0.7	33 ± 9	16.5	>64	>32
10l	11.6 ± 2.4	>100	>8	>64	>5
10m	1.8 ± 0.4	> 100	>55	>64	>35
10n	1.6 ± 0.5	32 ± 5	20	27 ± 2	16.9
10o	1.7 ± 0.5	32 ± 3	18.8	52 ± 7	30.6
10j	1.0 ± 0.3	15.7 ± 0.8	15.7	67 ± 8	67
7	1.5 ± 0.4	>100	>66	63 ± 8	42
3k	3.9 ± 0.9	>100	>25	>64	>16
3i	32.8 ± 4.9	>100	>3	45 ± 8	1.4
Doxorubicin	–	0.21 ± 0.09	–	0.4 ± 0.1	–

[a] *T. cruzi intracellular amastigote assay. Data represent the mean ± SD of two independent assays.* [b] *Cytotoxicity assay. Data represent the mean ± SD of two independent assays.* [c] *Selectivity index (SI),* CC_{50}/IC_{50}.

The target-based results enabled the identification of relevant SARs, which allowed the uncovering of pivotal structural aspects that drive the enzyme-inhibitor molecular recognition and the activity of the investigated compounds. Moreover, these findings provide substantial insights into the design of reversible cruzain inhibitors, which can be useful to surmount the drawbacks associated with irreversible ligands. An important aspect to remark is the *in vitro* toxicity profile of some compounds, which indicates that they may be safer than the drugs currently available for the treatment of Chagas disease. In a context characterized by a lack of therapeutic innovation and serious safety and efficacy issues, the cruzain inhibitors described herein can be explored as novel chemical matter in forthcoming Chagas disease drug discovery campaigns.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

RAF: conceptualization, writing, molecular design, and organic synthesis. IP: conceptualization, writing, molecular design, and enzyme kinetics. TS: organic synthesis. CR: organic synthesis. MS: molecular design and enzyme kinetics. RSF: conceptualization, molecular design, and enzyme kinetics. LM and RK: *in vitro* experiments. LF: writing, molecular design, and *in vitro* experiments. LD: conceptualization, supervision, organic synthesis, and writing. AA: conceptualization, supervision, molecular design, and writing.

FUNDING

The National Council for Scientific and Technological Development (CNPq), Brazil, the Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil, and the São Paulo Research Foundation (FAPESP), Brazil, grants 2013/07600-3 and 2011/13789-6, for financial support and postdoctoral fellowship (grant 158926/2014-5).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article including NMR spectra (**Figures S6–S70**) can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00798/full#supplementary-material

REFERENCES

Arnal, A., Waleckx, E., Rico-Chávez, O., Herrera, C., Dumonteil, E. (2019). Estimating the current burden of Chagas disease in Mexico: A systematic review and meta-analysis of epidemiological surveys from 2006 to 2017. *PLoS Negl. Trop. Dis.* 13:e0006859. doi: 10.1371/journal.pntd.0006859

Avelar, L. A., Camilo, C. D., de Albuquerque, S., Fernandes, W. B., Gonçalez, C., Kenny, P. W., et al. (2015). Molecular design, synthesis and trypanocidal activity of dipeptidyl nitriles as cruzain inhibitors. *PLoS Negl. Trop. Dis.* 7:e0003916. doi: 10.1371/journal.pntd.0003916

Badiola, E., Fiser, B., Gómez-Bengoa, E., Mielgo, A., Olaizola, I., Urruzuno, I., et al. (2014). Enantioselective construction of tetrasubstituted stereogenic carbons through Brønsted base catalyzed michael reactions: α'-hydroxy enones as

key enoate equivalent. *J. Am. Chem. Soc.* 136, 17869–17881. doi: 10.1021/ja510603w

Barltrop, J. A., Owen, T. C., Cory, A. H., and Cory, J. G. (1991). 5-(3-carboxymethoxyphenyl)-2-(4,5-dimethylthiazolyl)-3-(4-sulfophenyl)tetrazolium, inner salt (MTS) and related analogs of 3-(4,5-dimethylthiazolyl)-2,5-diphenyltetrazolium bromide (MTT) reducing to purple water-soluble formazans as cell-viability indicators. *Bioorg. Med. Chem. Lett.* 1, 611–614. doi: 10.1016/S0960-894X(01)81162-8

Bern, C. (2015). Chagas' disease. *N. Engl. J. Med.* 373, 456–466. doi: 10.1056/NEJMra1410150

Buckner, F. S., Verlinde, C. L., La Flamme, A. C., and Van Voorhis, W. C. (1996). Efficient technique for screening drugs for activity against *Trypanosoma cruzi* using parasites expressing beta-galactosidase. *Antimicrob. Agents Chemother.* 40, 2592–2597. doi: 10.1128/AAC.40.11.2592

Clark, M., Cramer III, R. D., and Van Opdenbosch, N. (1989). Validation of the general purpose tripos 5.2 force field. *J. Comput. Chem.* 10, 982–1012. doi: 10.1002/jcc.540100804

Cucunubá, Z. M., Okuwoga, O., Basáñez, M. G., and Nouvellet, P. (2016). Increased mortality attributed to Chagas disease: a systematic review and meta-analysis. *Parasit. Vectors* 9:42. doi: 10.1186/s13071-016-1315-x

Dias, J. C., Coura, J. R., and Yasuda, M. A. (2014). The present situation, challenges, and perspectives regarding the production and utilization of effective drugs against human Chagas disease. *Rev. Soc. Bras. Med. Trop.* 47, 123–125. doi: 10.1590/0037-8682-0248-2013

Doyle, P. S., Zhou, Y. M., Hsieh, I., Greenbaum, D. C., McKerrow, J. H., and Engel, J. C. (2011). The *Trypanosoma cruzi* protease cruzain mediates immune evasion. *PLoS Pathog.* 7:e1002139. doi: 10.1371/journal.ppat.1002139

Engel, J. C., Doyle, P. S., Palmer, J., Hsieh, I., Bainton, D. F., and McKerrow, J. H. (1998). Cysteine protease inhibitors alter Golgi complex ultrastructure and function in *Trypanosoma cruzi*. *J. Cell Sci.* 111, 597–606.

Espíndola, J. W., Cardoso, M. V., Filho, G. B., Oliveira, E., Silva, D. A., Moreira, D. R., et al. (2015). Synthesis and structure-activity relationship study of a new series of antiparasitic aryloxyl thiosemicarbazones inhibiting *Trypanosoma cruzi* cruzain. *Eur. J. Med. Chem.* 101, 818–835. doi: 10.1016/j.ejmech.2015.06.048

Faghihi, K. (2008). New optically active poly(amide-imide)s based on N,N′-(pyromellitoyl)-bis-L-amino acid and 1,3-bis(4-aminophenoxy) propane: synthesis and characterization. *J. Appl. Polym. Sci.* 109, 74–81. doi: 10.1002/app.27675

Ferreira, L. G., and Andricopulo, A. D. (2017). Targeting cysteine proteases in trypanosomatid disease drug discovery. *Pharmacol. Ther.* 180, 49–46. doi: 10.1016/j.pharmthera.2017.06.004

Ferreira, L. L. G., and Andricopulo, A. D. (2019). Drugs and vaccines in the 21st century for neglected diseases. *Lancet Infect. Dis.* 19, 125–127. doi: 10.1016/S1473-3099(19)30005-2

Ferreira, R. S., Dessoy, M. A., Pauli, I., Souza, M. L., Krogh, R., Sales, A. I., et al. (2014). Synthesis, biological evaluation, and structure-activity relationships of potent noncovalent and nonpeptidic cruzain inhibitors as anti-*Trypanosoma cruzi* agents. *J. Med. Chem.* 6, 2380–2392. doi: 10.1021/jm401709b

Ferreira, R. S., Simeonov, A., Jadhav, A., Eidam, O., Mott, B. T., Keiser, M. J., et al. (2010). Complementarity between a docking and a high-throughput screen in discovering new cruzain inhibitors. *J. Med. Chem.* 53, 4891–4905. doi: 10.1021/jm100488w

Gasteiger, J., and Marsili, M. I. (1980). Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron* 36, 3219–3228. doi: 10.1016/0040-4020(80)80168-2

GBD DALYs and HALE Collaborators. (2016). Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990-2015: a systematic analysis for the Global Burden of Disease Study. *Lancet* 388, 1603–1658. doi: 10.1016/S0140-6736(16)31460-X

Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* 267, 727–748. doi: 10.1006/jmbi.1996.0897

Jose Cazzulo, J., Stoka, V., and Turk, V. (2001). The major cysteine proteinase of *Trypanosoma cruzi*: a valid target for chemotherapy of Chagas disease. *Curr Pharm Des.* 7, 1143–1156. doi: 10.2174/1381612013397528

Kinage, A. K., Upare, P. P., Shivarkar, A. B., and Gupte, S. P. (2011). Highly regioselective synthesis of β-amino alcohol by reaction with aniline and propylene carbonate in self solvent system over large pore zeolite catalyst. *GSC* 1, 76–84. doi: 10.4236/gsc.2011.13013

Latorre, A., Schirmeister, T., Kesselring, J., Jung, S., Joh,é, P., Hellmich, U. A., et al. (2016). Dipeptidyl nitroalkenes as potent reversible inhibitors of cysteine proteases rhodesain and cruzain. *ACS Med. Chem. Lett.* 7, 1073–1076. doi: 10.1021/acsmedchemlett.6b00276

Lill, M. A., and Danielson, M. L. (2011). Computer-aided drug design platform using PyMOL. *J. Comput. Aided Mol. Des.* 1, 13–19. doi: 10.1007/s10822-010-9395-8

McKerrow, J. H. (1999). Development of cysteine protease inhibitors as chemotherapy for parasitic diseases: insights on safety, target validation, and mechanism of action. *Int. J. Parasitol.* 29, 833–837. doi: 10.1016/S0020-7519(99)00044-2

Molina, I., Salvador, F., Sánchez-Montalv,á, A., Treviño, B., Serre, N., Sao Avilés, A., et al. (2015). Toxic profile of benznidazole in patients with chronic chagas disease: risk factors and comparison of the product from two different manufacturers. *Antimicrob. Agents Chemother.* 59, 6125–6123. doi: 10.1128/AAC.04660-14

Mott, B. T., Ferreira, R. S., Simeonov, A., Jadhav, A., Ang, K. K., Leister, W., et al. (2010). Identification and optimization of inhibitors of Trypanosomal cysteine proteases: cruzain, rhodesain, and TbCatB. *J. Med. Chem.* 53, 52–60. doi: 10.1021/jm901069a

Ndao, M., Beaulieu, C., Black, W. C., Isabel, E., Vasquez-Camargo, F., Nath-Chowdhury, M., et al. (2014). Reversible cysteine protease inhibitors show promise for a Chagas disease cure. *Antimicrob. Agents Chemother.* 2, 1167–1178. doi: 10.1128/AAC.01855-13

Neises, B., and Steglich, W. (1978). Simple method for the esterification of carboxylic acids. *Angew.Chem.* 17, 522–524. doi: 10.1002/anie.197805221

Neitz, R. J., Bryant, C., Chen, S., Gut, J., Caselli, E. H., Ponce, S., et al. (2015). Tetrafluorophenoxymethyl ketone cruzain inhibitors with improved pharmacokinetic properties as therapeutic leads for Chagas' disease. *Bioorg. Med. Chem. Lett.* 21, 4834–4837. doi: 10.1016/j.bmcl.2015.06.066

Olivera, M. J., Cucunubá, Z. M., Álvarez, C. A., and Nicholls, R. S. (2015). Safety profile of nifurtimox and treatment interruption for chronic chagas disease in colombian adults. *Am. J. Trop. Med. Hyg.* 93, 1224–1230. doi: 10.4269/ajtmh.15-0256

Pande, S. V., Utale, P. S., Gholse, S. B., Tekade, P. V., and Patil, S. G. (2014). Synthesis and antibacterial evaluation of carboxamide derivatives of amino acids. *Pharm. Chem. J.* 48, 29–33. doi: 10.1007/s11094-014-1040-8

Pérez-Molina, J. A., and Molina, I. (2018). Chagas disease. *Lancet* 391, 82–94. doi: 10.1016/S0140-6736(17)31612-4

Powell, M. J. D. (1977). Restart procedures for conjugate gradient method. *Math. Program.* 12, 241–254. doi: 10.1007/BF01593790

Rogers, K. E., Keränen, H., Durrant, J. D., Ratnam, J., Doak, A., Arkin, M. R., et al. (2012). Novel cruzain inhibitors for the treatment of Chagas' disease. *Chem. Biol. Drug Des.* 3, 398–405. doi: 10.1111/j.1747-0285.2012.01416.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Ferreira, Pauli, Sampaio, de Souza, Ferreira, Magalhães, Rezende, Ferreira, Krogh, Dias and Andricopulo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

REVIEW
published: 13 December 2019
doi: 10.3389/fchem.2019.00848

Could Quantum Mechanical Properties Be Reflected on Classical Molecular Dynamics? The Case of Halogenated Organic Compounds of Biological Interest

Lucas de Azevedo Santos[1], *Ingrid G. Prandi*[1] *and Teodorico C. Ramalho*[1,2]*

[1] *Department of Chemistry, Federal University of Lavras, Lavras, Brazil,* [2] *Department of Chemistry, Faculty of Science, University of Hradec Kralove, Hradec Kralove, Czechia*

Essential to understanding life, the biomolecular phenomena have been an important subject in science, therefore a necessary path to be covered to make progress in human knowledge. To fully comprehend these processes, the non-covalent interactions are the key. In this review, we discuss how specific protein-ligand interactions can be efficiently described by low computational cost methods, such as Molecular Mechanics (MM). We have taken as example the case of the halogen bonds (XB). Albeit generally weaker than the hydrogen bonds (HB), the XBs play a key role to drug design, enhancing the affinity and selectivity toward the biological target. Along with the attraction between two electronegative atoms in XBs explained by the σ-hole model, important orbital interactions, as well as relief of Pauli repulsion take place. Nonetheless, such electronic effects can be only well-described by accurate quantum chemical methods that have strong limitations dealing with supramolecular systems due to their high computational cost. To go beyond the poor description of XBs by MM methods, reparametrizing the force-fields equations can be a way to keep the balance between accuracy and computational cost. Thus, we have shown the steps to be considered when parametrizing force-fields to achieve reliable results of complex non-covalent interactions at MM level for *In Silico* drug design methods.

Keywords: halogen bonds, force-fields, molecular dynamics, non-covalent interactions, drug design

OPEN ACCESS

Edited by:
Jamie Platts,
Cardiff University, United Kingdom

Reviewed by:
Pui Shing Ho,
Colorado State University, United States
Cina Foroutan-Nejad,
Masaryk University, Czechia

***Correspondence:**
Teodorico C. Ramalho
teo@ufla.br

Specialty section:
This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

Received: *21 September 2019*
Accepted: *21 November 2019*
Published: *13 December 2019*

Citation:
Santos LdA, Prandi IG and Ramalho TC (2019) Could Quantum Mechanical Properties Be Reflected on Classical Molecular Dynamics? The Case of Halogenated Organic Compounds of Biological Interest. Front. Chem. 7:848. doi: 10.3389/fchem.2019.00848

INTRODUCTION

Biological systems are huge, they change in time and they are very sensitive to *in vivo* conditions like temperature and environment (Ramalho et al., 2009; Freitas et al., 2014; Nair and Miners, 2014; Jurinovich et al., 2015). These facts are remembered every day by drug designers, structural biologists, biophysicists and many other professionals that need to study these systems (Nair and Miners, 2014). In order to overcome these barriers, many scientists opt to model their systems using the classical atomistic Molecular Dynamics (MD) simulation method.

The classical MD is a computational method based on Molecular Mechanics (MM) physics and its first simulation was performed by Alder and Wainright (Alder and Wainwright, 1959) in the late '50s. In this pioneering work, the authors discussed the difficulties to treat the many-body

problem and proposed a numerical scheme to deal with multiple interactions of particles by solving Newton's motion equations. Although Alder and Wainright gave the first spark for the beginning of classical MD, the first realistic MD simulation was performed just in 1969 (Allen et al., 1969). In this work, by the implementation of Lennard-Jones potentials (essential to describe van der Waals interactions), Dr. Rahman and co-authors successfully modeled 864 atoms of liquid argon. Over the last decades, MD modeling was refined, and many different codes have been launched. Nowadays, drug design is one of the areas that most benefits from the enormous development that the atomistic MD has acquired over the years. However, the mindset that it is difficult, unnecessary or too time-consuming to parameterize new molecules may turn to final works that mislead the real interactions. Unphysical models or catastrophic geometries with very inaccurate interaction energies can be found along with an MD simulation, especially if the modeled molecule has too many chemical functions or π-conjugations (Davis and Patel, 2010; Prandi et al., 2016; Aytenfisu et al., 2017).

Despite the great effort being made by scientific programmers to enhance the quality of classical molecular simulation techniques, much more can be done by the user to improve inter- and intramolecular interactions outcomes. It should be kept in mind that intramolecular interactions are the driving force of most biomolecular phenomena (Martins et al., 2003; Adesokan et al., 2004; Ramalho et al., 2009; Poater et al., 2011; Ben-Naim, 2012; Hongo et al., 2013; Poznanski and Shugar, 2013). They are known to be quantum chemical phenomena that go beyond the classical description of matter and, in particular cases, they cannot be understood by simple electrostatic or dispersion schemes (Ramalho and da Cunha, 2011; Esrafili and Ahmadi, 2012; Wolters et al., 2014). This is one of the greatest challenges to force-field modeling since there is no classical analog to the quantum behavior of electrons.

That is the case of halogen bonds (XB), a real and relevant tool for rational drug design (Auffinger et al., 2004; Lu et al., 2009, 2010, 2012; Wilcken et al., 2013; Mendez et al., 2017). The XBs are non-covalent interactions between an acceptor (A), often Lewis base, and a halogenated molecule acting as a donor (D) (**Figure 1**). On one hand, some researchers address the origin of the XBs to the existence of a positive electrostatic potential region on the halogen atom (X) bond called σ-hole (Clark et al., 2007; Politzer et al., 2007, 2013; Kolár et al., 2014). On the other hand, the literature also highlights the importance of the orbital interactions, revealing the covalency part of XB (Wolters and Bickelhaupt, 2012; Wang et al., 2014; Novák et al., 2015; Wolters et al., 2015; Dominikowska et al., 2016; Bora et al., 2017; Santos et al., 2017). In contrast to molecular mechanics approaches, the XB are purely quantum chemical phenomena, whose strength grows with the size of the halogens, making chlorine, bromine, and iodine promising alternatives to promote secondary side-chain interactions inside protein cavities (Lu et al., 2010; Cavallo et al., 2016; Santos et al., 2017).

Once many compounds with biological activity have halogen atoms in their composition, the accurate description of XBs by molecular mechanics is crucial. Now, the main questions we pursue to answer are: what can we do to solve this problem?; what are the alternatives we have?; what is the best approach to build up accurate techniques to describe these interactions?

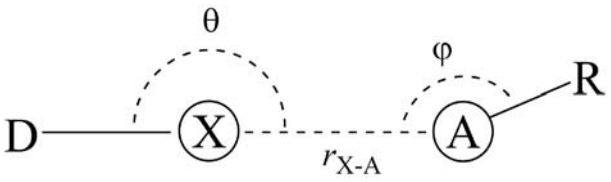


FIGURE 1 | General halogen-bond scheme. The donor (D) is bonded to a halogen atom (X) that interacts with the acceptor (A) in a distance *r*.

FORCE-FIELDS: AN OVERVIEW

Quantum Mechanics (QM) considers the electronic effects in molecules. On the other hand, Molecular Mechanics (MM) is based just on the interaction among classical charged particles, neglecting direct electronic effects.

Since the electronic environment around an atom changes accordingly to its neighborhood, we need an artifice to describe atoms with the same atomic numbers, but chemically different. For example, we need to distinguish the sp3 from sp2 carbons. To recover most of the electronic effects in MM based simulations different atom types should be employed. The atom types are atomic labels used to indicate chemically different atoms. In the example cited, different atom types should describe the carbons in ethanol.

After the atom types are set, the classical MD software associates each bonded or non-bonded molecular interaction to a set of parameters. In more detail, the MD software calculates the total potential (V_{TOT}) that acts in each particle.

$$V_{TOT} = V_S + V_A + V_D + V_{vdw} + V_C \quad (1)$$

The Equation (1) shows a generic form of a total potential: V_{S}, V_{A}, and V_{D} are the bonded terms, the stretching, angular, and dihedral potentials, respectively; and the last two terms are the non-bonded terms, in which accounts for the van der Waals interactions described by the Lennard-Jones potential (V_{vdw}) and Coulomb potential (V_{C}) that simulates the electrostatic interactions. It is important noticing that the terms in the total potential equation may vary depending on its implementation in the MD software. For instance, in Equation (1.1) we see the parameters k_μ^S, r_μ, and r_μ^0, that are the stretching force constant, the length of the bond and equilibrium distance, respectively; In Equation (1.2), k_μ^θ, θ_μ, and θ_μ^0 are the angular force constant, the angle, and equilibrium angle, respectively; The term $A_{j_\mu}^{SD}$ in Equation (1.3) is the dihedral torsional barrier, n_j^μ is the periodicity or the number of minima in the cosine function, δ_μ is the dihedral angle and γ_j^μ is a phase angle that represents the displacement of the dihedral angle (or torsional displacement); in Equation (1.4), ε_{ij}, σ_{ij}, and $\mathrm{r_{ij}}$ are the depth of the potential well, the distance at which V_{vdw} is a minimum and the distance between two particles, respectively; and finally

q and ϵ_0 in Equation (1.5) are the charge and the electrostatic constant, respectively. The description of each term in Equation (1) depends on these parameters and a complete set of equations together is named force-field (FF).

$$V_S = \sum_{\mu}^{N^{\circ}_{bonds}} k^S_{\mu}(r_{\mu} - r^0_{\mu})^2 \quad (1.1)$$

$$V_A = \sum_{\mu}^{N^{\circ}_{angles}} k^{\theta}_{\mu}(\theta_{\mu} - \theta^0_{\mu})^2 \quad (1.2)$$

$$V_D = \frac{1}{2}\sum_{\mu}^{N^{\circ}_{Sdihedrals}} \sum_{j=1}^{N^{\circ}_{\cos\mu}} A^{SD}_{j\mu}\left[1 + \cos\left(n^{\mu}_j \delta_{\mu} - \gamma^{\mu}_j\right)\right] \quad (1.3)$$

$$V_{vdw} = \sum_{i}^{N^{\circ}_{vdw\ interactions}} \sum_{i<j}^{N^{\circ}_{vdw\ interactions}} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \quad (1.4)$$

$$V_C = \sum_{i}^{N^{\circ}_{Coul.\ interac.}} \sum_{i<j}^{N^{\circ}_{Coul.\ interac.}} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad (1.5)$$

In the last decades, the use of MD has been expanded to different areas, being necessary the creation of parameters to describe a huge set of molecular interactions at the same time or, at least, those more relevant to a certain purpose. Thus, large groups of transferable parameters have been created aiming to describe chemically similar molecules. Nowadays, there are many sets of specialized parameters for the description of many different molecular groups like polymers, proteins, solvents, small organic molecules, etc. (Jorgensen et al., 1996; Schuler et al., 2001; Wang et al., 2004; Vanommeslaeghe et al., 2010; Dickson et al., 2014).

Due to the wide use of classical MD for protein modeling, here we may highlight two of the most used sets of parameters for biomodelling: AMBER (Assisted Model Building with Energy Refinement) (Case et al., 2014), created by Peter Kollman and his group at the University of California, and CHARMM (Chemistry at Harvard using Molecular Mechanics) (Vanommeslaeghe et al., 2010), initially developed by Martin Karplus and co-workers at Harvard University. Over the years, CHARMM has expanded and gained new specific parameters for the modeling of smaller molecules.

Other diffused family of parameters for biomolecular systems are OPLS and GROMOS. The OPLS (Optimized Potentials for Liquid Simulations) (Jorgensen et al., 1996) force-field was developed by Jorgensen's group to simulate proteins in solution. In 1976, GROMOS (GROningen MOlecular Simulation) (Schuler et al., 2001) started to be developed at the University of Groningen. Originally created for biomolecules modeling, until today it is constantly updated for many different classes of molecules.

Another example of a set of parameters specially designed for small and medium-sized organic molecules is the MMn (n = 1, 2, 3, 4) family of parameters developed by Allinger and coauthors (Allinger et al., 1971, 1989, 1996; Allinger, 1977; Lii and Allinger, 1989a,b; Nevins et al., 1996; Langley et al., 2001; Langley and Allinger, 2002).

With the expansion of the use of MD simulations in the pharmaceutical field, the development of a set of parameters for drug design research was urgent. Thus, in 2004, the GAFF (General AMBER Force Field) (Wang et al., 2004) family of parameters was specifically created and tested for pharmaceutical purposes. In order to guarantee a great transferability, many GAFF equilibrium parameters were extracted from the average of X-ray and *ab initio* calculations of different molecules. Besides, pure GAFF is not yet able to model the major part of metallic interactions in complexes and can poorly describe halogen bonds (Rendine et al., 2011; Li and Merz, 2017).

PARAMETERIZATION: THE KEY TO REALISTIC RESULTS

In the last decade, the US Food and Drug Administration (FDA)[1] approved more than 230 New Molecular Entity (NME) drugs. Almost 42% of the new non-biological approved drugs contain halogen atoms, and more than 3% are metallic complexes (see data in **Figure 2**). These data show the importance of a specific parameterization for new drugs since most general FFs are not able to describe with high accuracy those bonds for molecular dynamics simulations (Santos et al., 2014, 2017).

Unfortunately, specificity and transferability usually have an inversely proportional relationship. Due to their good transferability, GAFF and other general force-fields are ideal to describe molecules that are indirectly involved in the studies that we would like to do. However, for very specific cases, sets of general parameters are not enough to model physical structures or interactions and we need to remodel them.

Then, theoretical scientists have realized that molecular models need to be accurate to perform a realistic simulation. For this reason, many tools were developed aiming more straightforward paths to parameterization. The greatest part of methodologies is based on the extraction of equilibrium distances, angles and dihedrals from a QM optimized structure and the force constants are derived from the diagonalization of the Hessian matrix (extracted from a QM calculation). Some examples of tools that help computational scientists to parameterize their molecules are the following: Automated Topology Builder (ATB) (Malde et al., 2011), Paratool (Mayne et al., 2013), and Joyce (Barone et al., 2013).

More specifically, ATB is much more than an on-line tool to build biomolecular force-fields for MD or Monte Carlo simulations, it can also calculate free energies and predict hydration free enthalpies. This website is very user-friendly, does not require any installation procedure and sends an e-mail to the user when the parameters are ready.

Paratool is a plugin of the software Visual Molecular Dynamics (VMD) (Humphrey et al., 1996). It was specifically developed to build parameters in CHARMM or AMBER format.

[1]US Food and Drug Administration (FDA). Available online at: https://www.fda.gov/ (accessed November 4, 2017).

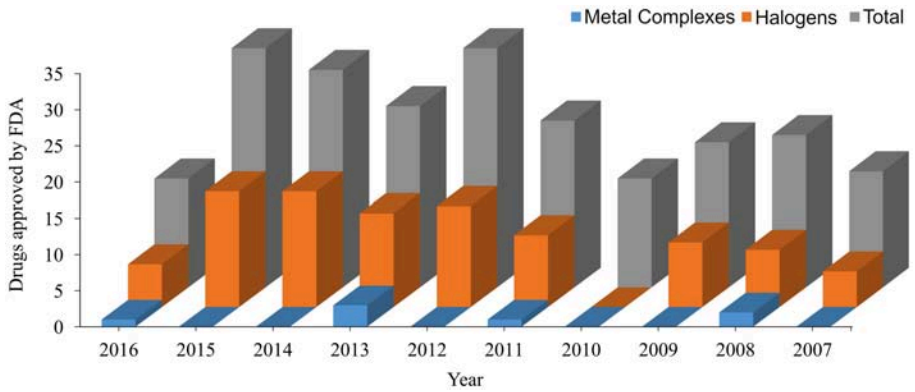


FIGURE 2 | New Molecular Entity (NME) approved by the US Food and Drug Administration (FDA) in the last 10 years. All data are taken from https://www.fda.gov.

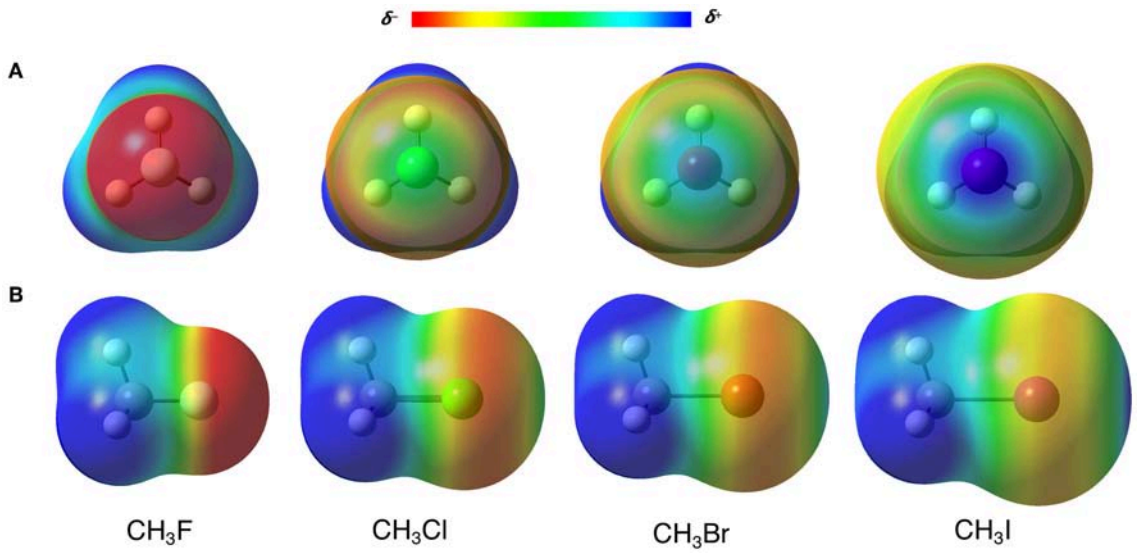


FIGURE 3 | (A) Front and **(B)** side view of the electrostatic potential surfaces (at 0.02 a.u.) from −0.3 (red) to 0.3 (blue) a.u. of CH_3X (X = F, Cl, Br, I) molecules. Computed at B3LYP-D3(BJ)/def2-TZVP, using Gaussian 09 (Frisch et al., 2009).

It is not as automated as ATB, but it is very user-friendly since it is linked to the VMD graphical interface.

Joyce is a software specially developed to assist the derivation of parameters in GROMACS (GROningen MAchine for Chemical Simulations) (van der Spoel et al., 2005) or Moscito (Paschek and Geiger, 2002) format for MD atomistic simulations. It is also a very versatile and flexible program, in which the user can symmetrize molecular groups, set dependencies between parameters and even impose specific values to the parameters.

The three aforementioned tools are just some examples of how a specific set of parameters can be derived. They were cited in ascendant order of time-consuming and effort to create a new specific FF. The choice of modeling a molecule with an FF created in a very automated way or a much more fitted one depends on the molecule, the required accuracy and how dependent the studied property is from the molecular geometry. However, another issue that cannot be neglected is the more complex intermolecular interactions, such as the halogen bonds. The difficulties of modeling intramolecular parameters are beyond the simple extraction and fitting of the parameters: they are also led by the MD software limitations.

Some MD software like AMBER do not distinguish intra from intermolecular parameters for van der Waals and Coulombic charges. Although MD simulations may give good results for many physical and macro properties of a large number of different systems, many times specific micro-interactions are not modeled in a refined way. This is the case of some vibrational modes: even if a very precise parameterization is done, coupled vibrational modes in very conjugated molecules are extremely hard to describe (Prandi et al., 2016; Andreussi et al., 2017). The difficulty of an accurate description is mirrored in the fact that most MD simulation programs do not couple molecular motions

like a stretching mode with an angular bending or the stretching modes of two adjacent atom pairs (Andreussi et al., 2017; Cerezo et al., 2018). More precisely, all mentioned terms in Equation (1) are expressed as sums of contributions, each one depending on a single internal coordinate. In this way, the off-diagonal terms (or hybrid terms) of the Hessian are not explicitly taken into account. Here, it is important to emphasize that it is mathematically and physically possible to derive parameters considering the cited couplings (Cerezo et al., 2018), but the implementation of a force-field functional form that describes the coupled terms can still not be done in many MD software.

It is evident that neither the best set of parameters can completely recover the electronic effects of a given molecule along an MD trajectory. Although some MD simulation programs are starting to be more flexible in terms of a force-field functional implementation, like GROMACS and Moscito (Cerezo et al., 2018). There is a still long path to achieve the full force-field functional form flexibility. Indeed, the maximum refinement that a normal user of mostly MD programs can do is to construct his own set of parameters. However, diving into very specific cases simple parameterizations can still be not enough.

BEYOND THE LIMITS

As defined in the introduction, the halogen bonds (XB) are non-covalent interactions between the halogen bond donor (D) and the halogen-bond acceptor (A) (**Figure 1**). Thus, the force-fields will describe these phenomena through the van der Waals term (Equation 1.4) and by the Coulombic term (Equation 1.5).

Many researchers address the origin of the XBs to the σ-hole, classifying them as σ-hole interactions (Clark et al., 2007; Politzer et al., 2008). The σ-hole is a positive region on the electrostatic potential surface (ESP), that arises from a charge anisotropy effect along with the D–X bond (Clark et al., 2007; Politzer et al., 2013). In other words, the electron density polarizes toward the D–X, generating an electron depletion in the back of the halogen (X) toward the D–X bond axis (see the blue regions over the halogens in **Figure 3**, in which D = CH_3) (Politzer et al., 2012). For the σ-hole model, the strength of the XBs, which increases along X = F < Cl < Br < I, is directly correlated to the increase of the positive electrostatic potential on the halogen. In this sense, to perform a classical FF description of the XBs, the attraction between A and X should be rigorously described by the Coulomb potential. However, here we have at least two barriers to overcome: firstly, the XB cannot be seen as the attraction of two points charges as described by Equation (1.5), but the interaction of two densities; secondly, even using the point charge scheme, halogen atoms often have negative charges that would cause an electrostatic repulsion between X and A, not allowing the XB to happen. The fact is, something totally different from the usual parameterization must be done.

The first attempt to describe the XB through molecular mechanics was suggested by Ibrahim who introduced the Explicit σ-hole (ESH) theory (Ibrahim, 2011, 2012; Kolár et al., 2013). The ESH is a way to model the σ-hole as a massless positive point charge bonded to the halogen atom at a certain distance (r_{ESH})

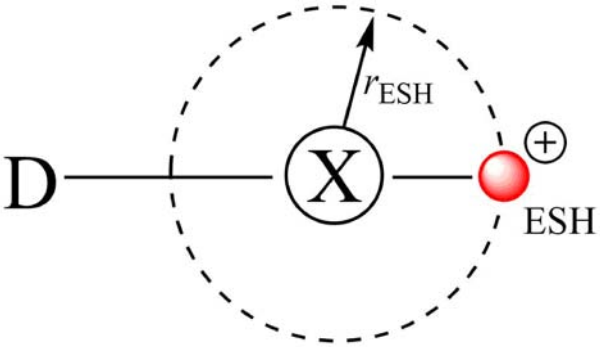


FIGURE 4 | The explicit σ-hole (ESH) scheme to model halogen bonds via the molecular mechanics approach.

(**Figure 4**). In general, there are two parameters to be fitted: the charge of the massless point and its distance to X (r_{ESH}).

The ESH strategy has promoted huge advances for the modeling of XB in a biological environment predicting energy minima points in halogen-bonded systems along the potential energy surfaces. However, it is totally based on the classical electrostatic point of view of chemical interactions, that is, the electrostatic attraction between two point charges.

In fact, the XBs are a mix of attractive dispersion, electrostatics and orbital interactions in balance with repulsive orbital interactions (Pauli repulsion) and should not be described neglecting either one of them (Huber et al., 2013; Wolters et al., 2014; Santos et al., 2017). **Figure 5** shows a simplified scheme of the halogen-bonding mechanism by Wolters and Bickelhaupt in the sight of Kohn-Sham density functional molecular orbital theory (Wolters and Bickelhaupt, 2012). An occupied molecular orbital of the acceptor, described by n*p* orbitals of the halides, interacts with an unoccupied molecular orbital of the halogenated molecule (DX) to promote an attractive orbital interaction. Here, the doubly occupied orbital can be further extrapolated to any doubly occupied MO. See that the unoccupied molecular orbital of the DX molecule will have a strong sigma anti-bonding orbital (σ^*_{D-X}) character. The Pauli repulsion originates from the interaction between the doubly occupied orbitals.

In practice, the σ-hole model often seems to work, but just by coincidence. In previous work, we have shown that the maximum ESP values on σ-hole (V_{max}) and the unoccupied orbital which contains the contribution of σ^*_{D-X} may have a similar origin (Santos et al., 2017). In other words, by increasing the value of V_{max}, the σ^*_{D-X} will be stabilized and the interaction energy will become more stable.

Once the XBs have an important contribution of non-classical interactions, they cannot be described only by the Coulomb potential to get the ideal parameterization. In the traditional FF equation (Equation 1), the other alternative is to look at the van der Waals term. The Equation (2) is the Lennard-Jones 12-6 potential (Lennard-Jones, 1931) written in a different way than in Equation (1.4). Here, the positive part is the

repulsion term and the negative part is the attractive term. The parameters are reduced to ε, the potential energy depth, and r_e, the equilibrium distance.

$$V_{vdw} = V_{LJ\ 12-6} = \varepsilon \left\{ \left(\frac{r_e}{r}\right)^{12} - 2\left(\frac{r_e}{r}\right)^{6} \right\} \quad (2)$$

In theory, the repulsive part of Equation (2) would account for the Pauli repulsion, which is decently described by the traditional FF being the result of the steric hindrance between two atoms. The attractive term would account for the dispersive and orbital interactions. The first is also quite well-parametrized but the same cannot be said for the orbital interactions (Wu et al., 2012; Santos et al., 2014). The main problem of neglecting the orbital interactions in molecular mechanics is to get overestimated destabilizing energies at low range distances (Santos et al., 2014, 2017). At this interaction bond length, the Pauli repulsion and charge transfer are exponentially intensified.

One way to minimize this problem is to use the LJ 10-6 potential (Equation 3). With a lower exponential factor in the repulsion term, the interaction energies at low range distances are less destabilized (Du et al., 2013; Santos et al., 2014, 2017). However, the LJ 10-6 does not bring geometric improvements in comparison with the LJ 12-6, mainly in the cases that the non-covalent interactions are extremely directional, as the halogen bonds. The Lennard-Jones potential can model if a Lewis base will approximate toward the σ_{D-X} bond axis or not and how it would affect the interaction energy (Soteras Gutiérrez et al., 2016; Bernardes and Canongia Lopes, 2017).

$$V_{vdw} = V_{LJ\ 10-6} = \varepsilon \left\{ \left(\frac{r_e}{r}\right)^{10} - 2\left(\frac{r_e}{r}\right)^{6} \right\} \quad (3)$$

CHANGING THE POTENTIAL EQUATIONS

If the actual model does not work for a specific system, we must reformulate it. So, why not do the same for FF equations? Nevertheless, it is not necessary to build an equation from scratch. Wiser is to modify a well-known model. In the case of reformulating new non-bonded force-field terms, subtly modifications into the V_C or V_{LJ} have been done to get more accurate functions.

The directional dependence of halogen bonds can be understood by looking at the σ-hole and MO theories together. The interaction angle θ must be close to 180° to lead the interaction toward the D–X bond axis (**Figure 1**). This is the geometry configuration that would maximize the electrostatic and orbital interactions for both σ-hole and MO theories (Riley et al., 2009, 2013; Esrafili and Ahmadi, 2012; Santos et al., 2017). The angle φ depends on the electronic structure of the acceptor (A) in order to maximize the attractive donor-acceptor orbital overlaps (see **Figure 5**). For instance, having an sp^2 oxygen as acceptor, φ would be around 120° to provide the frontal overlap between the lone pair of the oxygen (LP_O) and the σ^*_{D-X} orbital. By the same perspective, for an *sp* nitrogen as acceptor, φ would be around 180° and, for π acceptors, φ would be around 90° (Cavallo et al., 2016; Nziko and Scheiner, 2016; Santos et al., 2017).

In a very clever way, Carter and co-workers (Carter et al., 2012; Scholfield et al., 2015; Koebel et al., 2016) have introduced the angular dependence into the LJ 12-6 and Coulombic potentials to describe bromine bonds, which was later extended to chlorine and iodine. The Equations (4) and (5) are the *ff*BXB functions to describe the non-bonding terms of XBs. The effective halogen charge (Z_X) is defined by the amplitude (A) and the baseline (B) of the cosine function, which has the period ν and $\alpha = 180 - \theta$. In V_{LJ}, $\langle r_{vdw(X)} \rangle$ is now the average radius of the bromine at the energy minimum.

$$V_C = \frac{Z_X Z_A e^2}{D r^n} \quad ; \quad Z_X = A\cos(\nu\alpha) + B \quad (4)$$

$$V_{LJ} = \sqrt{\varepsilon_X \varepsilon_A} \left\{ \left(\frac{r_{vdw(A)} + \langle r_{vdw(X)} \rangle - \Delta r_X \cos(\nu\alpha)}{r} \right)^{12} - 2\left(\frac{r_{vdw(A)} + \langle r_{vdw(X)} \rangle}{r} \right)^{6} \right\} \quad (5)$$

Parameterized to predict the halogen bonds in DNA junctions, the variation in the interaction energies were ∼0.06 to ∼0.7 kcal.mol^{-1} in comparison to the experimental data. The *ff*BXB functions also give good values of θ, from ∼146 to ∼122°.

Du et al. have introduced new polarizable non-bonded functions to the force-field equations in order to reproduce the XBs, the PEff model (Du et al., 2013). The electrostatic potential is defined by (6), in which Q is a constant, α, β, and ζ are parameters from *ab initio* electrostatic potential, r_1 is a coordinate in the equatorial area, R is the distance from the halogen atom toward the D–X axis and r is the halogen-bond length.

$$V_{elst}(r_1, R, r) = Q \cdot \left[\exp(-\alpha r_1 - \beta R) - exp(\zeta r)\right] / r \quad (6)$$

The Lennard-Jones potential was used to simulate the repulsion and dispersion interactions (7), in which r_e is a function of θ, $r_{e,T}$ is the transverse distance, $r_{e,L}$ is the longitudinal distance and λ is a steepness parameter manually set to 1.26.

$$V_{rd} = 4\varepsilon \left\{ \left(\frac{r_e(\theta)}{r}\right)^{10} - \left(\frac{r_e(\theta)}{r}\right)^{6} \right\};$$
$$r_e(\theta) = r_{e,T} sin^2(\lambda\theta) + r_{e,L} cos^2(\lambda\theta) \quad (7)$$

The third and last term is the polarization energy (8), in which E is the electronic field, E_{tot} incorporates the induce dipole effects and α is the isotropic polarizability of the halogen.

$$V_{pol} = -\frac{1}{2}\alpha E \cdot E_{tot} \exp\left(1.0 - \left(\frac{r}{r_{min}}\right)^2\right) \quad (8)$$

The PEff functions have demonstrated a good performance in comparison with MP2 methods to predict the binding energies for chlorine, bromine, and iodine. Applied to well-known crystal structures, the deviation of the halogen-bond lengths was <0.1 Å and giving bond angles close to 180°.

Both *ff*BXB and PEff are complete force-field and already functional, albeit some tests with molecular dynamics must be

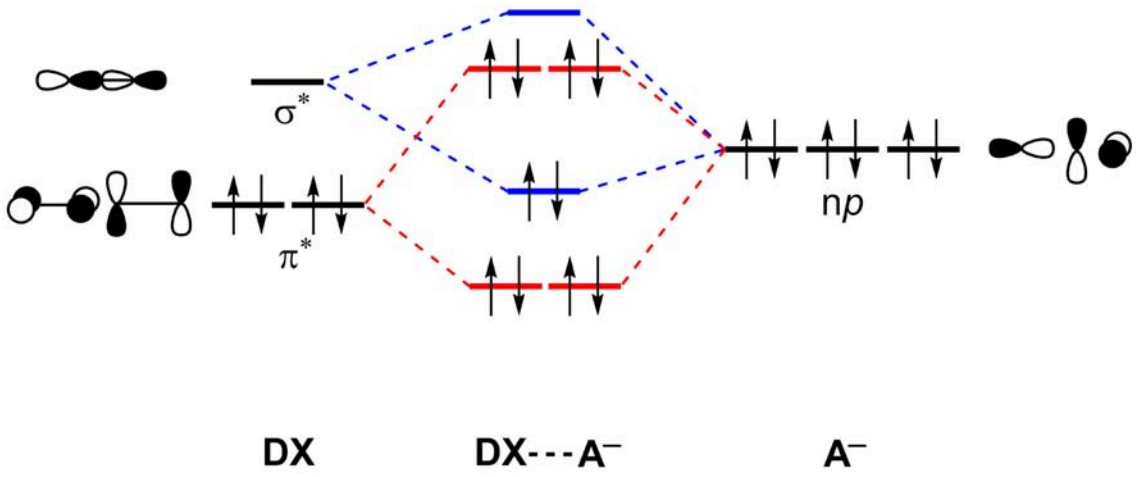


FIGURE 5 | Simplified molecular orbital perspective of halogen-halide bonds. Main attractive interactions (blue) and repulsive interactions (red) are highlighted. X, D, A = F, Cl, Br, I.

carried out. Moreover, they only consider Cl, Br, and I as donors and Lewis basis with available electron lone pairs (i.e., A = S, O, N) as acceptors. Nonetheless, molecules with π orbitals can also act as halogen bond acceptors and must be considered since there are several aromatic structures in the biological environment. In this sight, a new LJ 10-6 function has been proposed that takes into account the halogen-bond acceptor nature and also includes the fluorine in the XB donor, the E_{mod} (Equation 10) (Santos et al., 2017). Indeed, at certain conditions regarding the electronic structure of the whole donor molecule, the fluorine atom can form strong halogen-bonded complexes, sometimes as strong as the chlorine bonds (Wolters and Bickelhaupt, 2012; Santos et al., 2017).

$$V_{LJ} = E_{mod} = \varepsilon \left\{ \left(\frac{r_{e+\frac{\delta cos\Theta}{6}}}{r} \right)^{10} - 2 \left(\frac{r_e}{r+\gamma} \right)^{6} \right\} \qquad (9)$$

The E_{mod} empirical potential (10) has two new parameters: δ and γ (11). The parameter δ accounts for the attractive orbital interactions regarding the angular dependence to minimize the repulsion term, based on the synergy between V_{max} and the $\sigma^*_{\mathrm{D-X}}$ energies, as aforementioned. The V_{max} is calculated by QM methods, α is the van der Waals radius of the halogen atom and β is a constant, in which β = 2.5 for lone pair acceptors and β = 0.432 for aromatic acceptors. The parameter γ is a function of δ to rebalance the potential.

$$\delta = \frac{\beta V_{\mathrm{max}}}{4\pi\alpha^3} \;\; ; \;\; \gamma = \left[\frac{2^{2-\delta}\,(1-\delta)}{25r} \right] \qquad (10)$$

The E_{mod} was designed to use the r_{e} already parameterized by any force-field without halogen-bond corrections. In practice, V_{max} should be obtained by a QM calculation and used to fully define the parameters in Equation (11). Also, the E_{mod} could be used the general V_{LJ} function of the FF, since when V_{max} is not given (i.e., equal to zero), the parameters δ and γ will tend to zero, and this function will behave like a traditional LJ 10-6 potential. However, E_{mod} has not been tested with a complete force-field equation and there are no parameters for iodine.

Performing subtle modifications in the traditional empirical potentials is a good strategy to improve force-field equations. There are many other examples of modified potentials fitted to obtain reliable results of complex properties at the molecular mechanics level (Bernardes and Canongia Lopes, 2017; Franchini et al., 2018; Lin and MacKerell, 2018; Nunes et al., 2018). This approach eases the implementation of these functions by not requiring a huge effort to build a code from the beginning but using an already existing open-access and well-working code.

The use parameters obtained from previous quantum mechanical calculations can surely improve the results of a molecular dynamics simulation, but the next step is to rework the potentials in Equation (1) to further decrease the level of empiricism. That is where we find the *ab initio* force-fields (McDaniel et al., 2016; Xu et al., 2018; Pérez-Conesa et al., 2019). In principle, it would be possible to properly describe any non-covalent interactions with *ab initio* derived potentials, considering their specific properties, with a manageable computational cost.

SUMMARY AND OUTLOOK

Through the last decades, computational methods have been employed in the investigations around chemical properties of the matter. The evolution of technology has allowed us to go deeper into the atomic level to retrieve information about chemical bonds and non-covalent interactions. However, the computational cost has been the border of how further our knowledge could go. To overcome these borders, cheap computational approaches based on classical mechanics have emerged.

In this review, we have discussed how cheap approaches, like molecular dynamics (MD) and molecular mechanics

(MM) calculations, can be improved. Toward this goal, the parameterization of their force field (FF) equations is the key. Most of the parameters can be obtained by quantum mechanical (QM) calculations together with specific tools to modify and generate more accurate FF. Nevertheless, we have further explored one case that only setting up better parameters is not enough to retrieve the real information from a non-covalent interaction.

Being purely quantum chemical phenomena, the halogen bonds (XBs) have required the replacement of some FF potentials, since simple classical equations could not describe the properties of systems they are involved in. This replacement has been wisely done by modifying and introducing new parameters to well-known potentials. The new potentials to describe XBs were fit to high-level QM calculations, showing good agreement with crystal structure data. Thus, we strongly believe that the classical mechanical approaches will evolve by introducing new potentials based on *ab initio* calculations. The scope of this review is to highlight the relevance of ab initio parameterizations if the recovering of quantum chemical effects, lost by MM simulations, is wanted.

AUTHOR CONTRIBUTIONS

LS and IP held the literature research and prepared the initial draft of this review. LS, IP, and TR were responsible for proofing and preparing the final copy of this review.

FUNDING

We thank the Brazilian agencies FAPEMIG, CNPq, and CAPES for the financial support. This work was also supported by University of Hradec Kralove (Faculty of Science, VT2019-2021).

REFERENCES

Adesokan, A. A., Roberts, V. A., Lee, K. W., Lins, R. D., and Briggs, J. M. (2004). Prediction of HIV-1 integrase/viral DNA interactions in the catalytic domain by fast molecular docking. *J. Med. Chem.* 47, 821–828. doi: 10.1021/jm0301890

Alder, B. J., and Wainwright, T. E. (1959). Studies in molecular dynamics. General method. *J. Chem. Phys.* 31:459. doi: 10.1063/1.1730376

Allen, R. E., de Wette, F. W., and Rahman, A. (1969). Calculation of dynamical surface properties of noble-gas crystals. II. *Phys. Rev.* 179:887. doi: 10.1103/PhysRev.179.887

Allinger, N. L. (1977). Conformational analysis. 130. MM2. a hydrocarbon force field utilizing V1 and V2 torsional terms. *J. Am. Chem. Soc.* 99, 8127–8134. doi: 10.1021/ja00467a001

Allinger, N. L., Chen, K., and Lii, J. H. (1996). An improved force field (MM4) for saturated hydrocarbons. *J. Comput. Chem.* 17, 642–668. doi: 10.1002/(SICI)1096-987X(199604)17:5/6<;642::AID-JCC6>;3.0.CO;2-U

Allinger, N. L., Thomas Tribble, M., Miller, M. A., and Wertz, D. H. (1971). Conformational analysis. LXIX. an improved force field for the calculation of the structures and energies of hydrocarbons. *J. Am. Chem. Soc.* 93, 1637–1648. doi: 10.1021/ja00736a012

Allinger, N. L., Yuh, Y. H., and Lii, J. H. (1989). Molecular mechanics. the MM3 force field for hydrocarbons. 1. *J. Am. Chem. Soc.* 111, 8551–8566. doi: 10.1021/ja00205a001

Andreussi, O., Prandi, I. G., Campetella, M., Prampolini, G., and Mennucci, B. (2017). Classical force fields tailored for QM applications: is it really a feasible strategy? *J. Chem. Theory Comput.* 13, 4636–4648. doi: 10.1021/acs.jctc.7b00777

Auffinger, P., Hays, F. A., Westhof, E., and Ho, P. S. (2004). Halogen bonds in biological molecules. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8742–8747. doi: 10.1073/pnas.0407607101

Aytenfisu, A. H., Spasic, A., Grossfield, A., Stern, H. A., and Mathews, D. H. (2017). Revised RNA dihedral parameters for the Amber force field improve RNA molecular dynamics. *J. Chem. Theory Comput.* 13, 900–915. doi: 10.1021/acs.jctc.6b00870

Barone, V., Cacelli, I., De Mitri, N., Licari, D., Monti, S., and Prampolini, G. (2013). Joyce and Ulysses: integrated and user-friendly tools for the parameterization of intramolecular force fields from quantum mechanical data. *Phys. Chem. Chem. Phys.* 15, 3736–3751. doi: 10.1039/c3cp44179b

Ben-Naim, A. (2012). Levinthal's question revisited, and answered. *J. Biomol. Struct. Dyn.* 30, 113–124. doi: 10.1080/07391102.2012.674286

Bernardes, C. E. S., and Canongia Lopes, J. N. (2017). Modeling halogen bonds in ionic liquids: a force field for imidazolium and halo-imidazolium derivatives. *J. Chem. Theory Comput.* 13, 6167–6176. doi: 10.1021/acs.jctc.7b00645

Bora, P. L., Novák, M., Novotný, J., Foroutan-Nejad, C., and Marek, R. (2017). Supramolecular covalence in bifurcated chalcogen bonding. *Chem. Eur. J.* 23, 7315–7323. doi: 10.1002/chem.201700179

Carter, M., Rappé, A. K., and Ho, P. S. (2012). Scalable anisotropic shape and electrostatic models for biological bromine halogen bonds. *J. Chem. Theory Comput.* 8, 2461–2473. doi: 10.1021/ct3001969

Case, D. A., Babin, V., Berryman, J. T., Betz, R. M., Cai, Q., Cerutti, D. S., et al. (2014), *AMBER 14.* San Francisco, CA: University of California.

Cavallo, G., Metrangolo, P., Milani, R., Pilati, T., Priimagi, A., Resnati, G., et al. (2016). The halogen bond. *Chem. Rev.* 116, 2478–2601. doi: 10.1021/acs.chemrev.5b00484

Cerezo, J., Prampolini, G., and Cacelli, I. (2018). Developing accurate intramolecular force fields for conjugated systems through explicit coupling terms. *Theor. Chem. Acc.* 137:80. doi: 10.1007/s00214-018-2254-8

Clark, T., Hennemann, M., Murray, J. S., and Politzer, P. (2007). Halogen bonding: the σ-Hole. *J. Mol. Model.* 13, 291–296. doi: 10.1007/s00894-006-0130-2

Davis, J. E., and Patel, S. (2010). Revised charge equilibration parameters for more accurate hydration free energies of alkanes. *Chem. Phys. Lett.* 484, 173–176. doi: 10.1016/j.cplett.2009.09.061

Dickson, C. J., Madej, B. D., Skjevik, A. A., Betz, R. M., Teigen, K., Gould, I. R., et al. (2014). Lipid14: the amber lipid force field. *J. Chem. Theory Comput.* 10, 865–879. doi: 10.1021/ct4010307

Dominikowska, J., Bickelhaupt, F. M., Palusiak, M., and Fonseca Guerra, C. (2016). Source of cooperativity in halogen-bonded haloamine tetramers. *Chem. Phys. Chem.* 17, 474–480. doi: 10.1002/cphc.201501130

Du, L., Gao, J., Bi, F., Wang, L., and Liu, C. (2013). A polarizable ellipsoidal force field for halogen bonds. *J. Comput. Chem.* 34, 2032–2040. doi: 10.1002/jcc.23362

Esrafili, M. D., and Ahmadi, B. (2012). A theoretical investigation on the nature of Cl. . . N and Br. . . N halogen bonds in FArX. . . NCY complexes (X=Cl, Br and Y=H, F, Cl, Br, OH, NH2, CH3 and CN). *Comput. Theor. Chem.* 997, 77–82. doi: 10.1016/j.comptc.2012.07.038

Franchini, D., Dapiaggi, F., Pieraccini, S., Forni, A., and Sironi, M. (2018). Halogen bonding in the framework of classical force fields: the case of chlorine. *Chem. Phys. Lett.* 712, 89–94. doi: 10.1016/j.cplett.2018.09.052

Freitas, L. B., Borgati, T. F., de Freitas, R. P., Ruiz, A. L., Marchetti, G. M., de Carvalho, J. E., et al. (2014). Synthesis and antiproliferative activity of 8-hydroxyquinoline derivatives containing a 1,2,3-triazole moiety. *Eur. J. Med. Chem.* 84, 595–604. doi: 10.1016/j.ejmech.2014.07.061

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., et al. (2009). *Gaussian 09, Revision D.01*. Wallingford, CT: Gaussian Inc.

Hongo, K., Cuong, N. T., and Maezono, R. (2013). The importance of electron correlation on stacking interaction of adenine-thymine base-pair step in B-DNA: a quantum Monte Carlo study. *J. Chem. Theory Comput.* 9, 1081–1086. doi: 10.1021/ct301065f

Huber, S. M., Scanlon, J. D., Jimenez-Izal, E., Ugalde, J. M., and Infante, I. (2013). On the directionality of halogen bonding. *Phys. Chem. Chem. Phys.* 15:10350. doi: 10.1039/c3cp50892g

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graph.* 14, 33–38. doi: 10.1016/0263-7855(96)00018-5

Ibrahim, M. A. (2011). Molecular mechanical study of halogen bonding in drug discovery. *J. Comput. Chem.* 32, 2564–2574. doi: 10.1002/jcc.21836

Ibrahim, M. A. (2012). Molecular mechanical perspective on halogen bonding. *J. Mol. Model.* 18, 4625–4638. doi: 10.1007/s00894-012-1454-8

Jorgensen, W. L., Maxwell, D. S., and Tirado-Rives, J. (1996). Development and testing of the OLPS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* 118, 11225–11236. doi: 10.1021/ja9621760

Jurinovich, S., Viani, L., Prandi, I. G., and Mennucci, B. (2015). Towards an *ab initio* description of the optical spectra of light-harvesting antennae: application to the CP29 complex of photosystem II. *Phys. Chem. Chem. Phys.* 17, 14405–14416. doi: 10.1039/C4CP05647G

Koebel, M. R., Schmadeke, G., Posner, R. G., and Sirimulla, S. (2016). AutoDock VinaXB: implementation of XBSF, new empirical halogen bond scoring function, into AutoDock Vina. *J. Chem. Inform.* 8:27. doi: 10.1186/s13321-016-0139-1

Kolár, M., Hobza, P., and Bronowska, A. K. (2013). Plugging the explicit σ-holes in molecular docking. *Chem. Commun.* 49, 981–983. doi: 10.1039/C2CC37584B

Kolár, M., Hostaš, J., and Hobza, P. (2014). The strength and directionality of a halogen bond are co-determined by the magnitude and size of the σ-Hole. *Phys. Chem. Chem. Phys.* 16, 9987–9996. doi: 10.1039/C3CP55188A

Langley, C. H., and Allinger, N. L. (2002). Molecular mechanics (MM4) calculations on amides. *J. Phys. Chem. A* 106, 5638–5652. doi: 10.1021/jp014426r

Langley, C. H., Lii, J. H., and Allinger, N. L. (2001). Molecular mechanics (MM4) calculations on carbonyl compounds part I: aldehydes. *J. Comput. Chem.* 24, 1283–1286. doi: 10.1002/jcc.1177

Lennard-Jones, J. E. (1931). Cohesion. *Proc. Phys. Soc.* 43, 461–482. doi: 10.1088/0959-5309/43/5/301

Li, P., and Merz, K. M. Jr. (2017). Metal ion modeling using classical mechanics. *Chem. Rev.* 117, 1564–1686. doi: 10.1021/acs.chemrev.6b00440

Lii, J. H., and Allinger, N. L. (1989a). Molecular mechanics. the MM3 force field for hydrocarbons. 2. Vibrational frequencies and thermodynamics. *J. Am. Chem. Soc.* 111, 8566–8575. doi: 10.1021/ja00205a002

Lii, J. H., and Allinger, N. L. (1989b). Molecular mechanics. the MM3 force field for hydrocarbons. 3. The van Der Waals' potentials and crystal data for aliphatic and aromatic hydrocarbons. *J. Am. Chem. Soc.* 111, 8576–8582. doi: 10.1021/ja00205a003

Lin, F. Y., and MacKerell, A. D. (2018). Polarizable empirical force field for halogen-containing compounds based on the classical drude oscillator. *J. Chem. Theory Comput.* 14, 1083–1098. doi: 10.1021/acs.jctc.7b01086

Lu, Y., Liu, Y., Xu, Z., Li, H., Liu, H., and Zhu, W. (2012). Halogen bonding for rational drug design and new drug discovery. *Expert Opin. Drug Discov.* 7, 375–383. doi: 10.1517/17460441.2012.678829

Lu, Y., Shi, T., Wang, Y., Yang, H., Yan, X., Luo, X., et al. (2009). Halogen bonding-a novel interaction for rational drug design? *J. Med. Chem.* 52, 2854–2862. doi: 10.1021/jm9000133

Lu, Y., Wang, Y., and Zhu, W. (2010). Nonbonding interactions of organic halogens in biological systems: implications for drug discovery and biomolecular design. *Phys. Chem. Chem. Phys.* 12, 4543–4551. doi: 10.1039/b926326h

Malde, A. K., Zuo, L., Breeze, M., Stroet, M., Poger, D., Nair, P. C., et al. (2011). An automated force field topology builder (ATB) and repository: version 1.0. *J. Chem. Theory Comput.* 7, 4026–4037. doi: 10.1021/ct200196m

Martins, T. L. C., Ramalho, T. C., Figueroa-Villar, J. D., Flores, A. F. C., and Pereira, C. M. P. (2003). Theoretical and experimental ^{13}C and ^{15}N NMR investigation of guanylhydrazones in solution. *Magn. Reson. Chem.* 41, 983–988. doi: 10.1002/mrc.1299

Mayne, C. G., Saam, J., Schulten, K., Tajkhorshid, E., and Gumbart, J. C. (2013). Rapid parameterization of small molecules using the force field toolkit. *J. Comput. Chem.* 34, 2757–2770. doi: 10.1002/jcc.23422

McDaniel, J. G., Choi, E., Son, C. Y., Schmidt, J. R., and Yethiraj, A. (2016). *Ab Initio* force fields for imidazolium-based ionic liquids. *J. Phys. Chem. B* 120, 7024–7036. doi: 10.1021/acs.jpcb.6b05328

Mendez, L., Henriquez, G., Sirimulla, S., and Narayan, M. (2017). Looking back, looking forward at halogen bonding in drug discovery. *Molecules* 22, 22–25. doi: 10.3390/molecules22091397

Nair, P. C., and Miners, J. O. (2014). Molecular dynamics simulations: from structure function relationships to drug discovery. *In Silico Pharmacol.* 2:4. doi: 10.1186/s40203-014-0004-8

Nevins, N., Chen, K., and Allinger, N. L. (1996). Molecular mechanics (MM4) calculations on alkenes. *J. Comput. Chem.* 17, 669–694. doi: 10.1002/(SICI)1096-987X(199604)17:5/6<;669::AID-JCC7>;3.0.CO;2-S

Novák, M., Foroutan-Nejad, C., and Marek, R. (2015). Asymmetric bifurcated halogen bonds. *Phys. Chem. Chem. Phys.* 17, 6440–6450. doi: 10.1039/C4CP05532B

Nunes, R., Vila-Viçosa, D., Machuqueiro, M., and Costa, P. J. (2018). Biomolecular simulations of halogen bonds with a GROMOS force field. *J. Chem. Theory Comput.* 14, 5383–5392. doi: 10.1021/acs.jctc.8b00278

Nziko, V. P. N., and Scheiner, N. (2016). Comparison of π-hole tetrel bonding with σ-hole halogen bonds in complexes of XCN (X = F, Cl, Br, I) and NH_3. *Phys. Chem. Chem. Phys.* 18, 3581–3590. doi: 10.1039/C5CP07545A

Paschek, D., and Geiger, A. (2002). *MOSCITO 4.* Dortmund: Department of Physical Chemistry; University of Dortmund.

Pérez-Conesa, S., Torrico, F., Martínez, J. M., Pappalardo, R. R., and Marcos, E. S. (2019). A general study of actinyl hydration by molecular dynamics simulations using *ab initio* force fields. *J. Chem. Phys.* 150:104504. doi: 10.1063/1.5083216

Poater, J., Swart, M., Fonseca Guerra, C., and Bickelhaupt, F. M. (2011). Selectivity in DNA replication. interplay of steric shape, hydrogen bonds, π-stacking and solvent effects. *Chem. Commun.* 47, 7326–7328. doi: 10.1039/c0cc04707d

Politzer, P., Lane, P., Concha, M. C., Ma, Y., and Murray, J. S. (2007). An overview of halogen bonding. *J. Mol. Model.* 13, 305–311. doi: 10.1007/s00894-006-0154-7

Politzer, P., Murray, J. S., and Clark, T. (2013). Halogen bonding and other σ-hole interactions: a perspective. *Phys. Chem. Chem. Phys.* 15, 11178–11189. doi: 10.1039/c3cp00054k

Politzer, P., Murray, J. S., and Concha, M. C. (2008). σ-hole bonding between like atoms; a fallacy of atomic charges. *J. Mol. Model.* 14, 659–665. doi: 10.1007/s00894-008-0280-5

Politzer, P., Riley, K. E., Bulat, F. A., and Murray, J. S. (2012). Perspectives on halogen bonding and other σ-hole interactions: lex parsimoniae (Occam's Razor). *Comput. Theor. Chem.* 998, 2–8. doi: 10.1016/j.comptc.2012.06.007

Poznanski, J., and Shugar, D. (2013). Halogen bonding at the ATP binding site of protein kinases: preferred geometry and topology of ligand binding. *Biochim. Biophys. Acta* 1834, 1381–1396. doi: 10.1016/j.bbapap.2013.01.026

Prandi, I. G., Viani, L., Andreussi, O., and Mennucci, B. (2016). Combining classical molecular dynamics and quantum mechanical methods for the description of electronic excitations: the case of carotenoids. *J. Comput. Chem.* 37, 981–991. doi: 10.1002/jcc.24286

Ramalho, T. C., and da Cunha, E. F. F. (2011). Thermodynamic framework of the interaction between protein and solvent drives protein folding. *J. Biomol. Struct. Dyn.* 28, 645–646. doi: 10.1080/073911011010524975

Ramalho, T. C., Rocha, M. V. J., da Cunha, E. F. F., and Freitas, M. P. (2009). The search for new COX-2 inhibitors: a review of 2002-2008 patents. *Expert Opin. Ther. Pat.* 19, 1193–1228. doi: 10.1517/13543770903059125

Rendine, S., Pieraccini, S., Forni, A., and Sironi, M. (2011). Halogen bonding in ligand–receptor systems in the framework of classical force fields. *Phys. Chem. Chem. Phys.* 13, 19508–19516. doi: 10.1039/c1cp22436k

Riley, K. E., Murray, J. S., Politzer, P., Concha, M. C., and Hobza, P. (2009). Br. . . O complexes as probes of factors affecting halogen bonding: interactions of bromobenzenes and bromopyrimidines with acetone. *J. Chem. Theory Comput.* 5, 155–163. doi: 10.1021/ct8004134

Riley, K. E., Rezáč, J., and Hobza, P. (2013). Competition between halogen dihalogen and hydrogen bonds in bromo- and iodomethanol dimers. *J. Mol. Model.* 19, 2879–2883. doi: 10.1007/s00894-012-1727-2

Santos, L. A., da Cunha, E. F. F., Freitas, M. P., and Ramalho, T. C. (2014). Hydrophobic noncovalent interactions of inosine-phenylalanine: a theoretical model for investigating the molecular recognition of nucleobases. *J. Phys. Chem. A* 118, 5808–5817. doi: 10.1021/jp411230w

Santos, L. A., da Cunha, E. F. F., and Ramalho, T. C. (2017). Toward the classical description of halogen bonds: a quantum based generalized empirical potential for fluorine, chlorine, and bromine. *J. Phys. Chem. A* 121, 2442–2451. doi: 10.1021/acs.jpca.6b13112

Scholfield, M. R., Ford, M. C., Vander Zanden, C. M., Billman, M. M., Ho, P. S., and Rappé, A. K. (2015). Force field model of periodic trends in biomolecular halogen bonds. *J. Phys. Chem. B* 119, 9140–9149. doi: 10.1021/jp509003r

Schuler, L. D., Daura, X., and van Gunsteren, W. F. (2001). An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.* 22, 1205–1218. doi: 10.1002/jcc.1078

Soteras Gutiérrez, I., Lin, F. Y., Vanommeslaeghe, K., Lemkul, J. A., Armacost, K. A., Brooks, C. L., et al. (2016). Parametrization of halogen bonds in the CHARMM general force field: improved treatment of ligand–protein interactions. *Bioorg. Med. Chem.* 24, 4812–4825. doi: 10.1016/j.bmc.2016.06.034

van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005). GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26, 1701–1718. doi: 10.1002/jcc.20291

Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., et al. (2010). CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* 31, 671–690. doi: 10.1002/jcc.21367

Wang, C., Danovich, D., Mo, Y., and Shaik, S. (2014). On the nature of the halogen bond. *J. Chem. Theory Comput.* 10, 3726–3737. doi: 10.1021/ct500422t

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *J. Comput. Chem.* 25, 1157–1174. doi: 10.1002/jcc.20035

Wilcken, R., Zimmermann, M. O., Lange, A., Joerger, A. C., and Boeckler, F. M. (2013). Principles and applicationsof halogen bonding in medicinal chemistry and chemical biology. *J. Med. Chem.* 56, 1363–1388. doi: 10.1021/jm3012068

Wolters, L. P., and Bickelhaupt, F. M. (2012). Halogen bonding versus hydrogen bonding: a molecular orbital perspective. *Chem. Open* 1, 96–105. doi: 10.1002/open.201100015

Wolters, L. P., Schyman, P., Pavan, M. J., Jorgensen, W. L., Bickelhaupt, F. M., and Kozuch, S. (2014). The many faces of halogen bonding: a review of theoretical models and methods. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 4, 523–540. doi: 10.1002/wcms.1189

Wolters, L. P., Smits, N. W. G., and Fonseca Guerra, C. (2015). Covalency in resonance-assisted halogen bonds demonstrated with cooperativity in *n*-halo-guanine quartets. *Phys. Chem. Chem. Phys.* 17, 1585–1592. doi: 10.1039/C4CP03740E

Wu, X., Sun, Y., Li, C., and Yang, W. (2012). Parametric effects of the potential energy function on the geometrical features of ternary Lennard-Jones clusters. *J. Phys. Chem. A* 116, 8218–8225. doi: 10.1021/jp3037395

Xu, P., Guidez, E. B., Bertoni, C., and Gordon, M. S. (2018). Perspective: *ab initio* force field methods derived from quantum mechanics. *J. Chem. Phys.* 148:090901. doi: 10.1063/1.5009551

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Santos, Prandi and Ramalho. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 19 December 2019
doi: 10.3389/fchem.2019.00820

Novel Quinoline Compound Derivatives of NSC23925 as Potent Reversal Agents Against P-Glycoprotein-Mediated Multidrug Resistance

Xingping Quan[1†], Hongzhi Du[2†], Jingjing Xu[3†], Xiaoying Hou[1], Xiaofeng Gong[1], Yao Wu[1], Yuqi Zhou[1], Jingwei Jiang[1], Ligong Lu[4], Shengtao Yuan[1], Xiangyu Yang[4], Lei Shi[3]* and Li Sun[1]**

[1] Jiangsu Key Laboratory of Drug Screening, China Pharmaceutical University, Nanjing, China, [2] School of Pharmacy, Hubei University of Chinese Medicine, Wuhan, China, [3] Henan Key Laboratory of Organic Functional Molecules and Drug Innovation, School of Chemistry and Chemical Engineering, Henan Normal University, Xinxiang, China, [4] Zhuhai Interventional Medical Center, Zhuhai Precision Medical Center, Zhuhai People's Hospital, Zhuhai Hospital Affiliated with Jinan University, Zhuhai, China

OPEN ACCESS

Edited by:
Kamil Kuca,
University of Hradec Králové, Czechia

Reviewed by:
Bin Yu,
Zhengzhou University, China
José Carlos Menéndez,
Complutense University of Madrid, Spain

***Correspondence:**
Xiangyu Yang
yangxygdzhh@163.com
Lei Shi
shirlei@sina.com
Li Sun
sunli@cpu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: 21 July 2019
Accepted: 12 November 2019
Published: 19 December 2019

Citation:
Quan X, Du H, Xu J, Hou X, Gong X, Wu Y, Zhou Y, Jiang J, Lu L, Yuan S, Yang X, Shi L and Sun L (2019) Novel Quinoline Compound Derivatives of NSC23925 as Potent Reversal Agents Against P-Glycoprotein-Mediated Multidrug Resistance. Front. Chem. 7:820. doi: 10.3389/fchem.2019.00820

Multidrug resistance is a serious problem and a common cause of cancer treatment failure, leading to patient death. Although numerous reversal resistance inhibitors have been evaluated in preclinical or clinical trials, efficient and low-toxicity reversal agents have not been identified. In this study, a series of novel quinoline compound derivatives from NSC23925 were designed to inhibit P-glycoprotein (P-gp). Among them, YS-7a showed a stronger inhibitory effect against P-gp than verapamil, as a positive control, when co-incubated with chemotherapy drugs at minimally cytotoxic concentrations. YS-7a suppressed the P-gp transport function without affecting the expression of P-gp but stimulated the ATPase activity of P-gp in a dose-dependent manner. Next, molecular docking was used to predict the six most probable binding sites, namely, SER270, VAL273, VAL274, ILE354, VAL357, and PHE390. Moreover, YS-7a had no effect on cytochrome P450 3A4 activity and showed little toxicity to normal cells. In addition, combined treatment of YS-7a with vincristine showed a better inhibitory effect than the positive control verapamil *in vivo* without a negative effect on mouse weight. Overall, our results showed that YS-7a could reverse cancer multidrug resistance through the inhibition of P-gp transport function *in vitro* and *in vivo*, suggesting that YS-7a may be a novel therapeutic agent.

Keywords: molecular docking, multidrug resistance (MDR), P-glycoprotein (P-gp), quinoline, reversal cancer resistance

INTRODUCTION

Cancer multidrug resistance (MDR) is a major cause of chemotherapy failure leading to patient death. MDR cancer cells often show pleiotropic cross-resistance to a wide range of chemotherapy drugs. Mechanisms of MDR can be classified into non-cellular-based and cellular-based resistance mechanisms (Krishna and Mayer, 2000). Numerous potential mechanisms of MDR have been reported involving the ABC transporter family, DNA damage and repair, cancer stem cell

regulation, microRNA regulation, and epigenetic regulation (Wu et al., 2014). Among these, the ABC transporter families play an important role in cellular-based resistance mechanisms by facilitating exocytosis of chemotherapy drugs (Choi and Yu, 2014; Wu et al., 2014). These transporters are universally expressed across MDR cancer cells, especially P-glycoprotein (P-gp; encoded by MDR1), which is functionally equivalent to an efflux pump that translocate substrates or chemotherapy drugs from the intracellular to the extracellular environment (Fojo et al., 1987; Konstantinos, 2015). Studies have confirmed that P-gp is highly expressed or overactivated in a large number of patients with failed chemotherapy (Alfarouk et al., 2015). Therefore, P-gp is a potential target for reversing drug resistance.

P-gp inhibitors, also known as MDR modulators, have been used to reverse MDR and block P-gp function in combination with chemotherapy drugs (Coley, 2010; Kumar and Jaitak, 2019). Several pharmacological P-gp inhibitors have been developed, including verapamil (VP), PSC-833, and tariquidar. VP was the first to be identified, and is commonly used as a P-gp inhibitor for its low affinity and other pharmacological activities, but has many side effects (Bellamy et al., 1988; Yusa and Tsuruo, 1989). Dexverapamil (Pirker et al., 1995; Thürlimann et al., 1995) and PSC-833 (Boesch et al., 1991; Kusunoki et al., 2010) lack the pharmacological activities of VP and cyclosporin A, but inhibit cytochrome P450 (CYP) 3A4 activity. This leads to complicated drug–drug interactions and limits their application (Chico et al., 2001; Labrie et al., 2006). Meanwhile, tariquidar (XR9576) (Federica et al., 2004; Fox and Bates, 2007), LY335979 (Dantzig et al., 1999; Shepard et al., 2003), and HM30181 (Cha et al., 2013; Köhler and Wiese, 2015) show more specific affinities to P-gp with fewer side effects. Several clinical trials are underway that are expected to address clinical drug resistance. Oraxol, the oral preparation consisting of paclitaxel and HM30181A, showed a strong trend in progression-free survival ($p = 0.077$), favoring oral paclitaxel over *intravenous* paclitaxel and a strong trend in overall survival ($p = 0.11$) (https://ir.athenex.com/, ClinicalTrials.gov Identifier: NCT02594371). However, improved novel inhibitors are required to make P-gp a reliable target for reversing resistance.

NSC23925 was identified from 2,000 small molecule compounds using a high-throughput cell-based screening assay, and can specifically inhibit P-gp and reverse MDR with no effect on P-gp expression (Duan et al., 2009). NSC23925 was shown to prevent the emergence of MDR in ovarian cancer both *in vitro* and *in vivo* (Yang et al., 2015) and in osteosarcoma (Yang et al., 2014) without affecting P-gp expression. NSC23925 also reversed MDR in cancer cells (Duan et al., 2012).

In this study, we designed and synthesized a series of NSC23925 analogs with improved potency through two mechanisms (**Figures 1B,C**). Owing to their new structure and superior activity, these compounds were granted a patent (CN 108017615A, CN 107973781A). Among the synthesized compounds, YS-7a and YS-7b showed better P-gp inhibition than the positive control VP and the parent compound NSC23925. Our findings demonstrated that substituting -OH with -OMe increased the intracellular accumulation of Rhodamine 123 (Rho123); therefore, YS-7a was selected for further evaluation of its potent reversal effect. Next, the target of YS-7a was confirmed using small interfering (si)RNA. YS-7a had no effect on mRNA and protein expression of P-gp but inhibited its efflux pumping effect and stimulating P-gp ATPase instead, supporting its direct effect on P-gp. The binding sites were predicted through molecular docking. There was no significant effect on CYP3A4 activity and almost no toxicity toward normal cells. Finally, YS-7a improved the anti-tumor effect of chemotherapy drugs, showing better reversal of drug resistance than VP when combined with vincristine (VCR) *in vivo*. Overall, our study showed that YS-7a inhibited P-gp with high efficiency and low toxicity both *in vitro* and *in vivo*. Therefore, YS-7a is a novel P-gp inhibitor that may be used for the treatment of MDR cancers.

MATERIALS AND METHODS

General Chemistry

All the reagents were obtained from commercial sources and used without further purification unless otherwise indicated. All organic solvents were dried and freshly distilled before use by standard methods. The reactions were monitored by thin layer chromatography (TLC) on GF254 silica gel coated plates and visualized by UV light (254 and 365 nm). Purification by column and flash chromatography was carried out using silica gel (200–300 mesh). Melting points were taken on a X-4B melting-point apparatus and were uncorrected. ^{1}H and ^{13}C NMR spectra were recorded in DMSO-d6 or $CDCl_3$ on a Bruker Avance/600 (^{1}H: 600 MHz, ^{13}C: 150 MHz at 25°C) or Bruker Avance/400 (^{1}H: 400 MHz, ^{13}C: 100 MHz at 25°C, Bruker Instruments, Inc., Billerica, MA, USA) Chemical shifts are expressed in values (ppm) relative to tetramethylsilane as an internal standard, and coupling constants (J values) were given in hertz (Hz). Abbreviations are represented as follows: br, broad; s, singlet; d, doublet; dd, double doublet; t, triplet; q, quartet; m, multiplet. HRMS analysis was performed on a mass spectrometer using electrospray ionization (ESI-oa-TOF), and the purity of all samples used for HRMS (>95%) was confirmed by ^{1}H and ^{13}C NMR spectroscopic analyses. HPLC was performed on Agilent Technologies 1200 LC Column 250 × 4.6 nm and using H_2O (95–5%)/MeOH (5–95%) during 22 min as the mobile phase. Flow rate was1.0 mL/min (all solvents were HPLC grade). The HPLC system was monitored at 254 nm.

Biology

Cell Lines and Cell Culture

Human leukemia cell line K562, human oral squamous carcinoma KB cells, human hepatocellular carcinoma cell line HepG2 and Human umbilical vein endothelial cells (HUVEC) were obtained from the Cell Bank of the Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences (Shanghai, China).The MDR1-overexpressed cell lines, 3 μM adriamycin (ADR)-selected (Dalian Meilun Biotech Co., Ltd., China) multidrug resistance cell K562/ADR and 0.1 μM vincristine (VCR)-selected (Lingnan Pharmaceutical Co., China) multidrug resistance cell KB/VCR were obtained from Nanjing Shenghe Pharmaceutical Ltd (Nanjing, China). All the cells were cultured in RPMI-1640 medium (Gibco) supplemented with 10% (v/v)

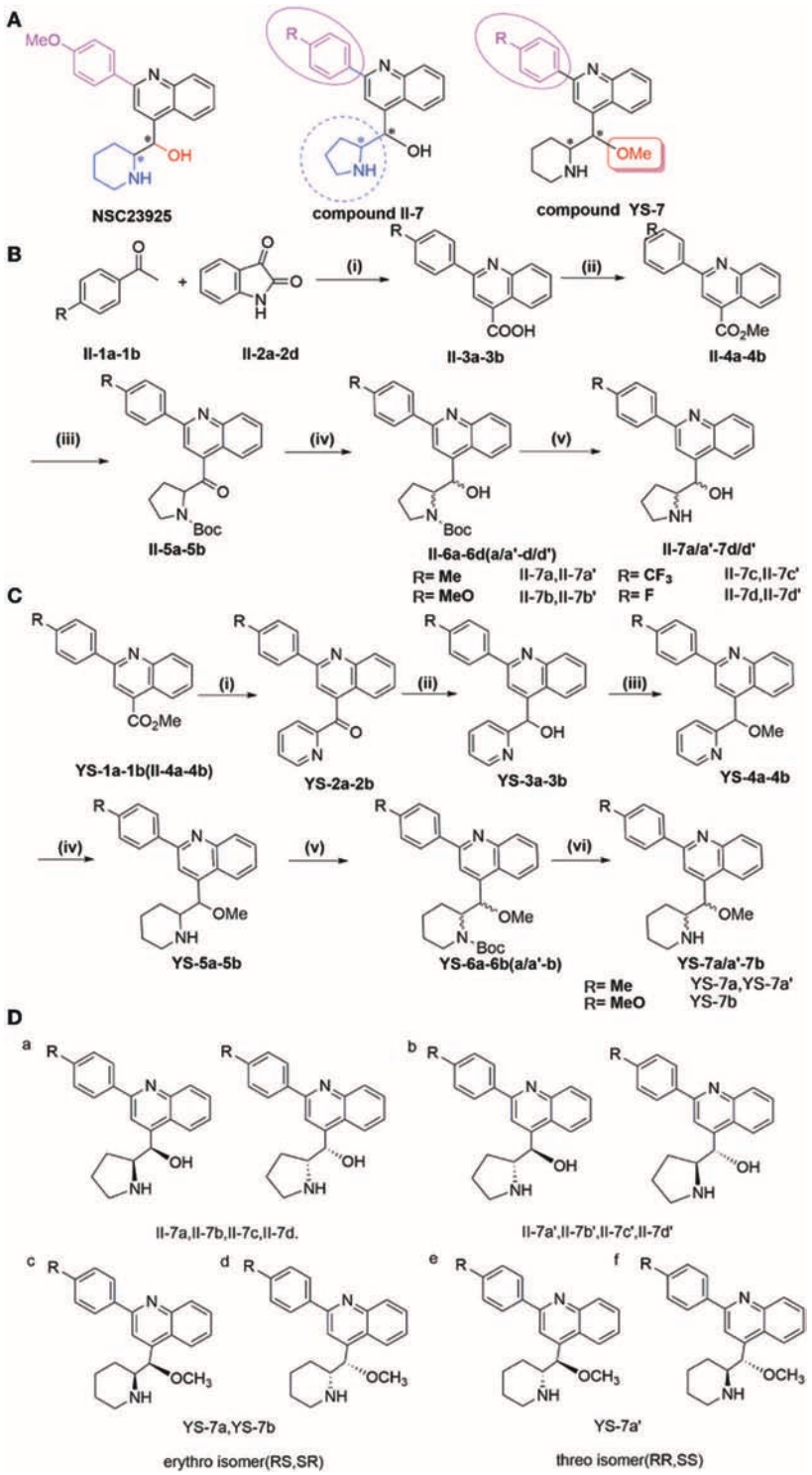


FIGURE 1 | (A) Synthesis of compounds II-7a, II-7a′, II-7b, II-7b′, II-7c, II-7c′, II-7d, and II-7d′; **(B)** synthesis of compounds YS-7a, YS-7a′, and YS-7b. Reagents and conditions **(A)**: (i) 1-(4-3-methoxyphenyl)ethanone, KOH, EtOH, 85°C, 24 h; (ii) concentrated H_2SO_4, MeOH, 65°C reflux overnight; (iii) 1-Boc-pyrrole (1 equiv), sec-BuLi (1.3 equiv) tetramethylethylenediamine 1.7 mL, −78°C, 2 h, dropping compound, stirring 2 h at room temperature, dry tetrahydrofuran; (iv) $NaBH_4$ (9 equiv), EtOH, 0°C, 0.75–1 h. (v) 2 M HCl, 30°C, 48 h. **(B)**: (i) 2-bromopyridine, n-BuLi, Et_2O/tetrahydrofuran, −78°C, 2 h; (ii) 7 $NaBH_4$, EtOH, 0°C, 1 h; (iii) dry N,N-dimethylformamide, NaCl, N_2, 0°C, 10 min, methyl iodide, stirring 1 h at 25°C; (iv) platinum dioxide, HCl, MeOH, H_2; (v) triethylamine, di-tert-butyl decarbonate dissolved in tetrahydrofuran at 0°C, stirring at room temperature overnight; (vi) 2 M HCl, 30°C, 48 h. **(C)** Design of target compounds. **(D)** Structure of compounds II-7a, II-7a′, II-7b, II-7b′, II-7c, II-7c′, II-7d, II-7d′, YS-7a, YS-7a′, and YS-7b (a–d).

heat-inactivated fetal bovine serum (FBS, HyClone) at 37°C and 5% CO_2.

Antibodies and Reagents

The antibodies and reagents included β-actin (AC026, ABclonal), MDR1 (22336-1-AP, proteintech), Anti-rabbit IgG (H+L) (Biotinylated Antibody #14708, Cell Signaling Technology), Anti-mouse IgG (H+L) (Biotinylated Antibody #14709, Cell Signaling Technology).

Rhodamine-123 (Rho-123) Accumulation

Intracellular fluorescence intensity was measured by Flow Cytometry (Wang et al., 2000). Cells in the exponential growth phase were seeded in 12-well plates, with about 10^4 cells per well. After a 24 h incubation at 37°C in a 5% CO_2, cells were treated with various concentrations of candidate compounds and verapamil (VP) for 4 h. Then 1 μg/mL Rhodamine-123 (Rho-123) was added directly to the cells. Additional incubation for 1 h at 37°C protected from light, the cells were harvested and immediately detected by flow cytometric (BD) at an excitation wavelength of 488 nm and emission wavelength of 530 nm. The results were calculated by GraphPad Prism 6.0 software.

Cell Proliferation Assays *in vitro*

The inhibition of the compounds on the growth of cancer cells KB/VCR and KB or human leukemia cell line K562/ADR and K562 cells were estimated by the 3-(4,5-dimethyl-2-thiazolyl)-2,5-diphenyl-2-H-tetrazolium bromide (MTT). Cells were plated in 96-well-plates, with about 1–2 × 10^3 cells per well. After 24 h, cells were incubated with various concentrations of compounds and verapamil (VP) for 72 h. MTT was added directly to the cells, and incubated for a further 4 h at 37°C protected from light. Finally, the absorbance at 490 nm was read on a microplate reader (Thermo Fisher Scientific). Experiments were conducted in triplicate at least and repeated three times independently. The inhibition rate was calculated as follows (Wei et al., 2019): inhibition rate (%) = (1– absorbance of the treated group/absorbance of the control group) × 100%. The reversing tumor resistance fold (RF) in resistance cancer cells = IC_{50} (concentration at half-maximum inhibition) of single chemotherapeutic drug/IC_{50} of chemotherapeutic drug combined with YS-7a.

siRNA Treatment

All siRNA fragments were synthesized by GenePharma (Shanghai, China). Cells in the logarithmic growth phase were seeded in 6-well-plates, with about 5–6 × 10^5 cells per well. Once the cells attached (often 8 h later), the medium with 10% FBS was replaced by fresh serum-free medium containing different siRNA for 6–8 h. The sequence of siRNA are as follows:

Negative Control	5′-UUCUCCGAACGUGUCACGUTT-3′
	5′-ACGUGACACGUUCGGAGAATT-3′
MDR1-homo-824	5′-GACCAGGUAUGCCUAUUAUTT-3′
	5′-AUAAUAGGCAUACCUGGUCTT-3′
MDR1-homo-2187	5′-GCGAAGCAGUGGUUCAGGUTT-3′
	5′-ACCUGAACCACUGCUUCGCTT-3′
MDR1-homo-3323	5′-CACCCAGGCAAUGAUGUAUTT-3′
	5′-AUACAUCAUUGCCUGGGUGTT-3′.

The medium should then be changed to fresh medium with 10% FBS for a further 24 h. Cells were used to evaluate the effectiveness of knockdown and the cytotoxicity effects; methods were the same as cell proliferation assays *in vitro*. The reversing tumor resistance fold (RF) in P-gp knockdown cancer cells = IC_{50} of single chemotherapeutic drug in P-gp knockdown cancer cells/IC_{50} of chemotherapeutic drug combined with YS-7a in P-gp knockdown cancer cells.

RT-qPCR Assays

The expression of the relative genes of cells was detected by RT-qPCR as reported (Hou et al., 2018b). The total RNA was isolated with the TRIzol® Reagent (Vazyme) and then reverse transcribed with the HiScript™ QRT SuperMix for qPCR (Vazyme). The mRNA level was measured using the SYBR Green master mix (Vazyme). The β-actin mRNA served as the control. Primer sequences used for qRT-PCR were as follows:

β-actin	5′-GGACTTCGAGCAAGAGATGG-3′ (forward)
	5′- AGCACTGTGTTGGCGTACAG-3′ (reverse)
MDR1	5′-GGAGCCTACTTGGTGGCACATAA-3′(forward)
	5′-TGGCATAGTCAGGAGCAAATGAAC-3′ (reverse).

Western Blot

The expression of the MDR1 protein was analyzed by Western Blot assays. After being treated with compounds, cells were harvested and lysed, the total protein concentrations were consistent according to the BCA kit (Beyotime Biotechnology, China). Protein lysates (20–30 μg protein per lane) were separated by 8% SDS-PAGE, and then the PVDF membranes were incubated with primary antibodies and second antibodies. The results were quantified by image analyzer (Bio-Rad, USA). The expression of β-actin was used as the control.

P-gp ATPase Activity Assays

The P-gp ATPase activity was tested by Pgp-Glo™ Assay Systems (Promega). The impact of candidate compounds on P-gp ATPase activity were examined by comparing untreated samples and samples treated with Na_3VO_4 (sodium orthovanadate). The compounds could be assessed as stimulating, inhibiting, or having no effect on basal P-gp ATPase activity.

Molecular Docking

Schrödinger was used for the molecular modeling studies as we have used before (Hou et al., 2018a). The crystal structure of the P-gp (PDB ID: 3WME) was prepared using Protein Preparation Wizard. The structure of compounds YS-7a were prepared using ChemBioDraw Ultra 13.0. The calculation was performed based on the force field OPLS (optimized potentials for liquid simulations) 2005 selecting water as the solvent. The following structure was obtained from the result of 1,000 calculation cycles.

CYP3A4 Activity Assays

The effect of the candidate compounds against CYP3A4 activity was performed by P450-Glo™ CYP3A4 Assay (Luciferin-IPA) Cell-Based/Biochemical Assay (Promega). The expression

of the CYP3A4 gene of HepG2 cells could be induced by 25 μM rifampicin. Cells were typically exposed for 24–72 h. The cells were then treated with various concentrations of test compounds for 72 h. Ketoconazole was used as the positive group. Finally, the P450-Glo™ CYP3A4 Assay (Luciferin-IPA) Cell-Based/Biochemical Assay (Promega) was added to pyrolysis, the treated HepG2 cells, and the fluorescence was detected by SoftMax Pro 6.1 (Beckman counter). The relative CYP3A4 activity was estimated as the formula: (the fluorescence of YS-7a treated/YS-7a treated HepG2 cells)/(the fluorescence of control/control HepG2 cells).

In vivo Experiments

Female BALB/c mice aged 4–6 weeks and weighing 18–20 g were purchased by the Model Animal Research Center of Nanjing University (Nanjing, China). KB/VCR cells (2×10^6) were transplanted to establish the subcutaneous xenograft model. Intraperitoneal administration once the tumor volume (TV) reached 100 mm^3, all the mice could be divided into six groups: the normal group (saline, 0.2 ml/kg), single chemotherapy drug group (VCR 0.5 mg/kg), positive drug single group (VP 10 mg/kg), single YS-7a group (YS-7a 10 mg/kg), positive drug group (VP 10 mg/kg+ VCR 0.5 mg/kg), and the combination YS-7a group (YS-7a 10 mg/kg + VCR 0.5 mg/kg). All the drugs were given by intraperitoneal injection every 3 days. After 24 days, all the animals were sacrificed and the final tumor obtained to calculate the relative tumor volume (RTV). TV and RTV were calculated as follows: TV (mm^3) = ½ × A × B^2, where A represents the longest diameter of tumor, B represents the shortest diameter. RTV = V_{24}/V_0, V_{24} represents the TV of day 24, and V_0 represents the TV of day 0. Additionally, all of the final tumors were weighted, and the weight loss ratio counted. The weight loss ratio (%) = $M_{24(\text{drugtreated})}/M24^*_{(\text{control})}100\%$. In this study, animal administration was guided by the Animal Care and Control Committee of the China Pharmaceutical University.

Statistical Analysis

All results are shown as the mean ± S.D of triplicate experiments. One-way ANOVA or the student's *t*-test was performed for the statistical analysis using the GraphPad Prism 5.0 software as previously reported (Du et al., 2018). All comparisons are made relative to the untreated controls. $^*P < 0.05$ was considered statistically significant; $^{**}P < 0.01$ and $^{***}P < 0.001$ was considered very statistically significant.

RESULTS

Design Novel Derivatives With Potent Reversal Activity

The natural compound NSC23925 contains a quinoline structure with one benzene ring and two chiral carbons in the side chain (**Figure 1A**). Based on the structural mode, we designed and obtained some derivatives **II-7a-a′, II-7b-b′**, **II-7c-c′**, **II-7d-d′**, **YS-7a-a′,** and **YS-7b** by changing the substituent on benzene ring in the side chain and replacing piperidine with tetrahydropyrrole (**compound II-7**) or replacing -OH with -OMe (**compound YS-7**) (**Figure 1A**). Each compound was purified by column chromatography and the purity of compounds was evaluated using high-performance liquid chromatography (HPLC) (**Table S1**, **Figures S1**–**S11**).

The general procedure for the synthesis of the compounds shown in **Figures 1B,C** is as follows: 2-(4-R-Phenyl)quinoline-4-carboxylic acid (**II-3**) was prepared from 4-R-1-phenyl ethenone (**II-1**) and indoline-2,3-dione (**II-2**) (yield: 80–82%), and was then esterified to afford methyl-2-(4-R-phenyl)quinoline-4-carboxylate (**II-4** or **YS-1**) (yield: 80–88%). Using n-BuLi as a base, 2-bromopyrrole or 2-bromopyridine was reacted with (**II-4** or **YS-1**) to give (2-(4-R-phenyl)quinolin-4-yl)(pyrrole-2-yl)-methanone (**II-5**) (yield: 20–25%) and (2-(4-R-phenyl)quinolin-4-yl)(pyridin-2-yl)-methanone (**YS-2**) (yield: 80–85%). Subsequent reduction of **II-5** or **YS-2** by $NaBH_4$ afforded 2-(4-R-phenyl) (quinolin-4-yl)(pyridin-2-yl)methanol (**II-6**) (yield: 87–96%) and 2-(4-R-phenyl) (quinolin-4-yl)(pyridin-2-yl)methanol (**YS-3**) (yield: 90–98%). **II-6** was separated by column chromatography to obtain a pair of diastereomers **II-6** (**a, a′-d, d′**) that could be separated based on polarity; the less polar compound was named **II-6** (**a-d**) and the more polar was named **II-6** (**a′-d′**). Methyl iodide was used as a methyl donor to produce methyl-2-(4-R-phenyl) (quinolin-4-yl)(pyridin-2-yl)methyl ether (**YS-4**) (yield: 85%). We then performed a reduction reaction of **YS-4** by hydrogen to produce 2-(hydroxy(2-(4-R-phenyl) quinolin-4-yl) methyl) piperidine-1-carboxylic acid tert-butyl ester (**YS-5**) (yield: 95%). A triethylamine and di-tert-butyl decarbonate reaction with **YS-5** was used to produce **YS-6**, after which **YS-6** was separated by column chromatography to obtain a pair of diastereomers that could be separated based on polarity; the less polar compound was named **YS-6** (**a** or **b**) and the more polar was named **YS-6** (**a′** or **b′**) (yield: 75%). The deprotection reaction of **II-6** or **YS-6** was performed and finally solid washed with dichloromethane to yield the corresponding diastereomer **II-7** (**a, a′-d, d′**) and **YS-7** (**a, a′** or **b**).

Based on a previous study of the Erythro/Threo configuration of phenyl 2-piperidylcarbinols by NMR, after knowing its absolute configuration (Lapidus and Fauley, 1971; Solladié-Cavallo et al., 2003), the ^{3}J-erythro and ^{3}J-threo in the CH(OH)-CH(NH) systems were different, always showing ^{3}J-erythro(RS, SR) $<^3$J-threo(RR, SS). Therefore, ^{1}H NMR was applied to determine the structure of these compounds. By analyzing these data, we found that the low polarity compounds (II-7a,7b,7c) could belong to etythro (RS, SR) isomer, and the highly polar compounds (II-7a',7b',7c') belong to threo (RR, SS) isomer. Therefore, we inferred that YS-7a and YS-7b (the lower polarity) belongs to etythro (RS, SR) isomer compounds, and YS-7a' belongs to the threo (RR, SS) isomer compound (**Figure 1D**). All NMR spectra of these compounds are provided in **Figures S12**–**S55**. After obtaining these 11 compounds, the reversal activities were measured.

Screening of Novel P-gp Inhibitors in MDR Cells

We evaluated the cancer MDR reversal activities of the candidate compounds. First, quantitative reverse transcription PCR and

Western blotting were performed to characterize the resistance of the KB/VCR and K562/ADR cell lines (**Figures 2A–F**). P-gp expression was increased at both the mRNA and protein levels. MDR of KB/VCR and K562/ADR cells was detected based on the half-maximal inhibitory concentration (IC_{50}) using the MTT assay. KB/VCR cells ($IC_{50} = 0.8294 \pm 0.241\,\mu M$) showed a 46.8-fold greater resistance to VCR than KB cells ($IC_{50} = 0.01770 \pm 0.094\,\mu M$), whereas K562/ADR cells ($IC_{50} = 6.919 \pm 0.01992\,\mu M$) showed a 347.3-fold greater resistance to adriamycin (ADR) than K562 cells ($IC_{50} = 0.01992 \pm 0.008\,\mu M$). Because the resistance of cancer cells decreased in the absence of chemotherapy drugs, compounds at a dose below the 20% inhibition concentration (0.1 μM VCR in KB/VCR cells; 3 μM ADR in K562/ADR cells) were co-incubated during the entire process. These findings supported the use of KB/VCR and K562/ADR cells for subsequent experimentation.

After confirming the overexpression of P-gp in MDR cancer cells, we detected the intracellular accumulation of Rho123 to screen for potential novel P-gp inhibitors and to assess P-gp inhibition. Rho123 was added to KB and KB/VCR cells at a final concentration of 5 μM for 1 h; 10 μM VP in KB/VCR cells were used as the positive control. The intracellular concentration of Rho123 decreased significantly in KB/VCR cells, which was reversed by VP (**Figure 2G**). Subsequently, II-7a, II-7a′, II-7b, II-7b′, II-7c, II-7c′, II-7d, II-7d′, YS-7a, YS-7a′, and YS-7b were evaluated by flow cytometry for their capacity to inhibit Rho123 efflux. Among the candidate compounds, the low-polarity compounds YS-7a (fold change = 82.19 ± 17.79) and YS-7b (65.85 ± 10.04) showed the greatest degree of Rho123 efflux inhibition, compared to VP (9.79 ± 0.70) and NSC23925 (40.53 ± 0.49) (**Figure 2H**), suggesting that they can potently reverse MDR. Further characterization was performed to measure their reversal efficiency *in vitro*.

Structure–Activity Relationship

Next, we investigated the structure–activity relationship (SAR) of the candidate compounds (**Table 1**). The low-polarity compound **II-7b**, which substituted piperidine for tetrahydropyrrole, showed significantly decreased activity compared to the parent compound. Changing the substituent on the benzene ring in the side chain with an electron-donating or electron-withdrawing group resulted in compounds **II-7a-a′** (R_1: -Me), **II-7c-c′** (R_1: -CF_3), and **II-7d-d′** (R_1: -F); these compounds did not improve the efflux of intracellular Rho123. Among these eight compounds (i.e., **II-7a-a′ to d-d′**), higher polarity damaged the activity.

Based on these results, we focused on the replacement of -OH with -OMe in the chiral carbons. The addition of a methoxy group in chiral carbons (**YS-7b**) promoted the activity, suggesting that reducing the polarity of the molecule and hydrogen bonding are beneficial to its activity. Interestingly, **YS-7a**, with an electron-donating methyl substitution, was more potent (fold change = 82.19 ± 17.79) than NSC23925, whereas the high-polarity compound **YS-7a′** (38.42 ± 25.09) showed no obvious improvement in activity. Considering the bioactivity and polarity of the compounds, we selected compounds **YS-7a** and **YS-7b** for further characterization.

In vitro Drug Resistance Reversal Effects of YS-7a and YS-7b

As potential P-gp inhibitors, YS-7a and YS-7b may reverse the resistance to chemotherapy drugs. Before testing YS-7a and YS-7b in combination with chemotherapy drugs, the antiproliferation effects of YS-7a and YS-7b were measured using the MTT assay. A concentration with low antiproliferation effects (<20% inhibition) was used to evaluate their MDR reversal effects (**Figures 3A,B**). In both KB/VCR and K562/ADR cells, 10 μM YS-7a, 2.5 μM YS-7b, and 2.5 μM VP were selected as the low-toxicity dose.

The drug resistance reversal effects of YS-7a and YS-7b were measured in combination with chemotherapy drugs based on cytotoxicity. Series concentrations of the chemotherapy drugs VCR and doxorubicin (ADR) were applied individually to explore the reversal effects of YS-7a and YS-7b. YS-7a exerted powerful reversal activity compared to the classical P-gp inhibitor VP in both KB/VCR and K562/ADR cells (**Figures 3C,D**). The IC_{50} values of the combined YS-7a + chemotherapy drug in drug-resistant KB/VCR and K562/ADR cells were 0.0376 ± 0.0116 μM and 0.268 ± 0.053 μM, respectively, whereas the IC_{50} values of the chemotherapy drugs in drug-sensitive KB and K562 cells were 0.435 ± 0.286 μM and 6.571 ± 1.758 μM, respectively; therefore, YS-7a showed a 10.11 ± 3.51 and 30.59 ± 5.83 reversal of drug resistance in KB/VCR and K562/ADR cells, respectively. Meanwhile, the IC_{50} values of the combined YS-7b + chemotherapy drug in drug-resistant KB/VCR and K562/ADR cells were 0.0561 ± 0.0390 μM and 0.616 ± 0.185 μM, respectively, whereas the IC_{50} values of the chemotherapy drugs in drug-sensitive KB and K562 cells were 0.375 ± 0.246 μM and 6.571 ± 1.758 μM, respectively; therefore, YS-7b showed a 6.92 ± 0.55 and 10.79 ± 0.87 reversal of drug resistance in KB/VCR and K562/ADR cells, respectively. Overall, YS-7a showed a significantly better drug resistance reversal effect in MDR cells and was selected for further experimentation of its reversal capabilities.

P-gp May Be a Target of YS-7a

Although YS-7a was screened for its potent reversal effect in MDR cancer cells, it remained unclear whether it targeted P-gp. To clarify this, knockdown of P-gp was implemented by three siMDR1 fragments. siMDR1-2 and siMDR1-3 were chosen for further experiments, both of which downregulated P-gp mRNA and protein levels in KB/VCR and K562/ADR cells (**Figures 4A–D**). Knockdown by siMDR1-2 and siMDR1-3 decreased P-gp transporter function (**Figures 4E,F**). After knockdown of P-gp, the drug resistance reversal potency of YS-7a decreased from 16.12- to 1.57-fold (siMDR1-2) and 1.63-fold (siMDR1-3) in KB/VCR cells, whereas that of VP decreased from 6.26- to 1.56-fold (siMDR1-2) and 1.63-fold (siMDR1-3) (**Figure 4G**). In other words, the reversal effect of YS-7a almost entirely disappeared after P-gp knockdown. Altogether, these results show that the drug resistance reversal effect of YS-7a relies on P-gp.

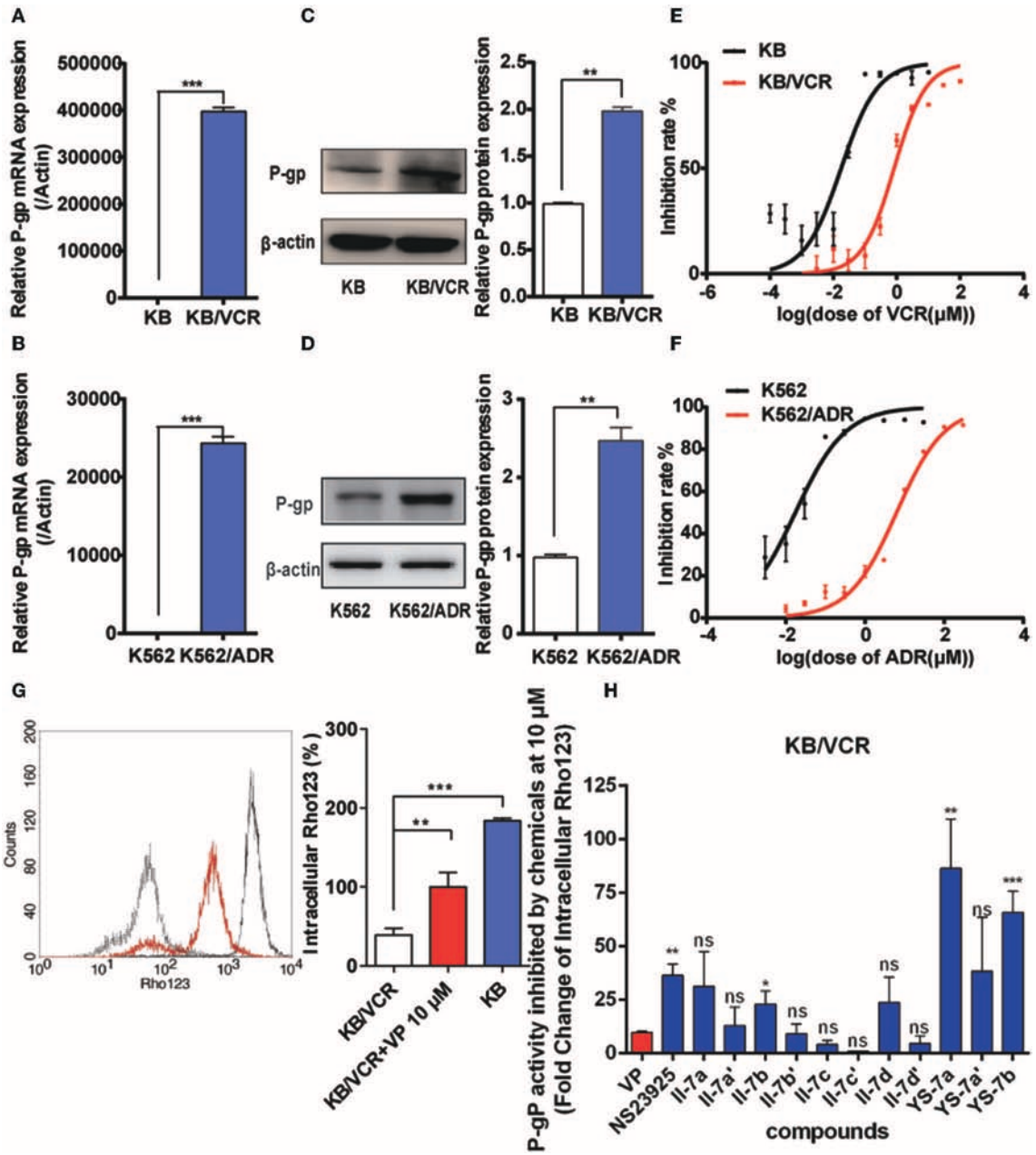
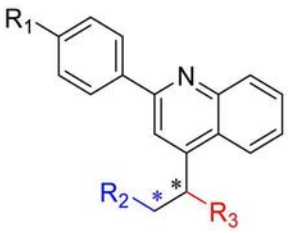


FIGURE 2 | Drug resistance of the KB/VCR and K562/ADR cell lines and screening for novel P-gp inhibitors. **(A,B)** Expression of P-gp mRNA in KB, KB/VCR, K562, and K562/ADR cells. **(C,D)** Expression of P-gp protein in KB, KB/VCR, K562, and K562/ADR cells. **(E,F)** The IC_{50} was determined after exposure to a series concentration of chemotherapy drugs (VCR or ADR) in KB, KB/VCR, K562, and K562/ADR cells for 72 h, and were measured using the MTT assay. **(G)** Intracellular content of Rho123 in KB cells, KB/VCR cells, and KB/VCR cells exposed to 10 μM VP for 4 h. **(H)** Flow cytometry was used to screen compounds that exhibited strong P-gp inhibition after incubation for 4 h; YS-7a and YS-7b resulted in more than 50 times accumulation of Rho123, and were chosen for further study. All experiments were repeated at least three times. $^{**}P < 0.01$ and $^{***}P < 0.001$.

YS-7a Does Not Affect P-gp Expression but Directly Inhibits Transport Function

MDR cells often exhibit overactivated or overexpressed P-gp and abnormal P-gp ATPase activity. To explore the specific mechanisms of YS-7a, flow cytometry was performed to monitor the function of P-gp transporters *via* Rho123 efflux. Treatment of KB/VCR cells with 10 μM YS-7a resulted in a significant increase compared to the group treated with 2.5 μM VP (positive control) (**Figure 5A**). Consistent results were obtained in K562/ADR cells (**Figure 5B**). However, YS-7a did not affect P-gp mRNA and protein levels in KB/VCR and K562/ADR cells (**Figures 5C–E**). These results showed that YS-7a inhibited P-gp function, but

TABLE 1 | SAR study of novel quinoline compounds.



Compound	-R_1	-R_2	-R_3	Structure	tR(min) /HPLC	Fold change of intracellular Rho123 (FC)
NSC23925	-OMe		-OH			36.50 ± 5.20
II-7a	-Me		-OH	Etythro (RS, SR) isomer	tR = 19.23 min 96.61%	31.35 ± 16.05
II-7a′				Threo (RR, SS) isomer	tR = 19.718 min 95.66%	12.94 ± 8.72
II-7b	-OMe			Etythro (RS, SR) isomer	tR = 19.068 min 96.17%	23 ± 6.13
II-7b′				Threo (RR, SS) isomer	tR = 19.488 min 95.10%	9.16 ± 4.66
II-7c	-CF_3			Etythro (RS, SR) isomer	tR = 10.917 min 97.35%	4.15 ± 1.96
II-7c′				Threo (RR, SS) isomer	tR = 13.772 min 95.03%	1.04 ± 0.01
II-7d	-F			Etythro (RS, SR) isomer	tR = 9.931 min 95.70%	23.79 ± 11.86
II-7d′				Threo (RR, SS) isomer	tR = 12.429 min 95.00%	4.63 ± 3.57
YS-7a	-Me		-OMe	Etythro (RS, SR) isomer	tR = 18.222 min 96.33%	86.43 ± 22.92
YS-7-a′				Threo (RR, SS) isomer	tR = 19.923 min 96.93%	38.42 ± 25.09
YS-7b	-OMe			Etythro (RS, SR) isomer	tR = 19.409 min 97.22%	65.85 ± 10.04

not expression. We further measured P-gp ATPase activity in the presence of YS-7a using the Pgp-Glo™ Assay Systems Kit (Promega; Madison, WI, USA), and found that YS-7a stimulated P-gp ATPase activity in a dose-dependent manner (**Figure 5F**). These results demonstrate that YS-7a may inhibit the P-gp substrate binding site in a similar manner as VP.

To explore the involvement of P-gp as a therapeutic target of YS-7a, we performed molecular docking experiments of YS-7a with P-gp (PDB: 3WME). The binding energies for all YS-7a poses were −11.3 kcal/mol, indicating that YS-7a could strongly bind to P-gp. Our results also showed that YS-7a formed six hydrophobic interactions with residues SER270, VAL273, VAL274, ILE354, VAL357, and PHE390 (**Figure 5G**). Based on these observations, YS-7a can directly bind to the functional domains of P-gp. Altogether, YS-7a may suppress the P-gp transport function without affecting its expression, by stimulating the ATPase activity of P-gp by directly binding to the six probable sites instead.

YS-7a Has No Effect on CYP3A4 Activity and Little Toxicity Toward Normal Cells

P-gp inhibitors, such as PSC-833 (Boesch et al., 1991; Kusunoki et al., 2010), can inhibit CYP3A4 activity, resulting in complicated drug–drug interactions and unexpected side effects; the ensuing toxicity can lead to failure of the final inhibitor clinical trial. To explore the potential for drug–drug interactions, we measured CYP3A4 activity after treatment with YS-7a. The results showed that YS-7a had no effect on CYP3A4 activity, even

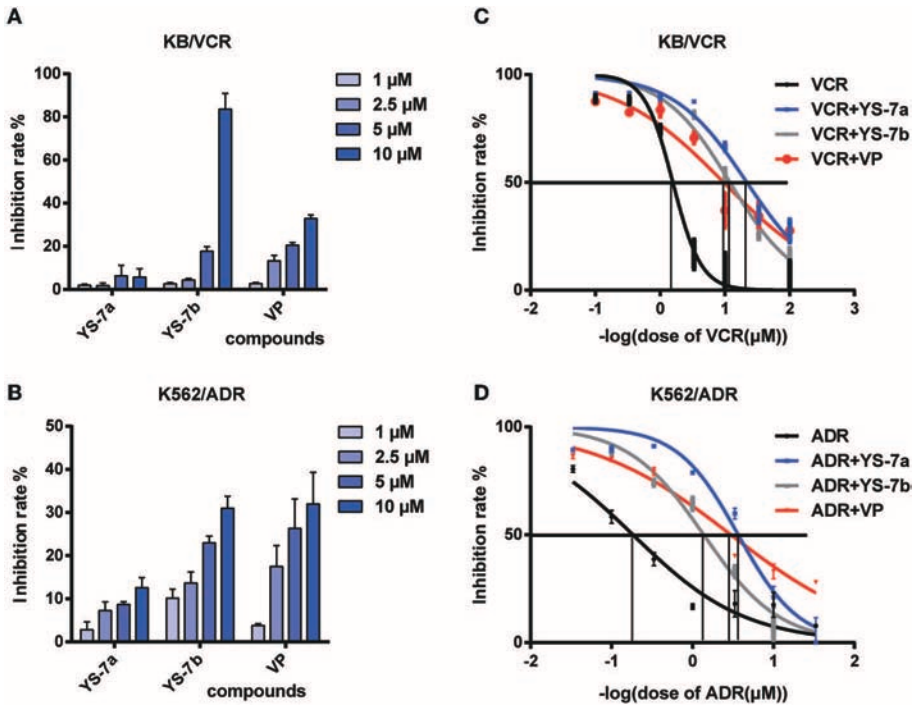


FIGURE 3 | Cytotoxicity and reversal effect of the candidate compounds YS-7a and YS-7b in KB/VCR and K562/ADR cells. The compounds were diluted to four concentrations, 1, 2.5, 5, and 10 μM, and cytotoxicity was evaluated after 72 h using the MTT assay in **(A)** KB/VCR and **(B)** K562/ADR cells. **(C)** Combined treatment of 10 μM YS-7a and 2.5 μM YS-7b with a series of VCR concentrations in KB/VCR cells for 72 h. **(D)** Combined treatment of 10 μM YS-7a and 2.5 μM YS-7b with a series of ADR concentrations in K562/ADR cells for 72 h. The data shown represent the mean ± SD of three independent experiments.

at high doses (100 μM) (**Figure 6A**). Thus, YS-7a may not have drug metabolism interactions.

Drugs can also show toxicity to vascular endothelial cells after entering the blood circulation (Cao et al., 2017). Thus, we evaluated the potential toxicity of YS-7a in human umbilical vein endothelial cells (HUVECs). The IC_{50} of YS-7a in HUVECs was 45.33 ± 2.55 μM (**Figure 6B**), suggesting that YS-7a (10 μM) has low potential for toxicity at concentrations used to reverse drug resistance. These results support the application of YS-7a as a novel potent P-gp inhibitor that can inhibit its transporter functions without undesirable side effects on CYP3A4 activity or endothelial cell toxicity.

In vivo Drug Resistance Reversal Effect of YS-7a

The mechanism and potential toxicity of YS-7a was confirmed *in vitro*. However, whether YS-7a reverses MDR *in vivo* remained unclear. Thus, we performed an *in vivo* xenograft experiment to evaluate the reversal effect of YS-7a (**Figure 7**). When combined with 0.5 mg/kg VCR, YS-7a at a dose of 10 mg/kg showed tumor growth inhibition of approximately 50.11%, while the single 10 mg/kg YS-7a group and single VCR group showed poor inhibition rates. Simultaneously, no YS-7a groups showed a decrease in mouse weight, indicating that YS-7a may have minimal toxicity *in vivo*. These findings indicate that YS-7a can reverse MDR *in vivo* with minimal potential toxicity.

DISCUSSION

To combat MDR cancer, the development of novel P-gp inhibitors is important, and most P-gp inhibitors are in preclinical or clinical trials. In this study, we synthesized 11 novel quinoline compounds, which could be divided into low-polarity (II-7a, II-7b, II-7c, II-7d, YS-7a, and YS-7b) and high-polarity (II-7a′, II-7b′, II-7c′, II-7d′, and YS-7a′) groups based on HPLC. Among these, YS-7a had the best MDR reversal effect *in vitro*, showing a reversal effect of over 10-fold in KB/VCR cells and over 30-fold in K562/ADR cells at low-toxicity concentrations.

We confirmed that YS-7a directly inhibited the transporter function of P-gp without affecting its expression, stimulating P-gp ATPase activity in a dose-dependent manner instead. Furthermore, YS-7a did not inhibit CYP3A4 activity and showed little cytotoxicity toward HUVECs at a concentration of 10 μM. In the KB/VCR xenograft model, 10 mg/kg YS-7a combined with 0.5 mg/kg VCR showed significant differences in tumor volumes compared to the control, with an average tumor growth inhibition exceeding 50%. These findings support YS-7a as a novel P-gp inhibitor, and can provide a reference for the design and development of additional P-gp inhibitors.

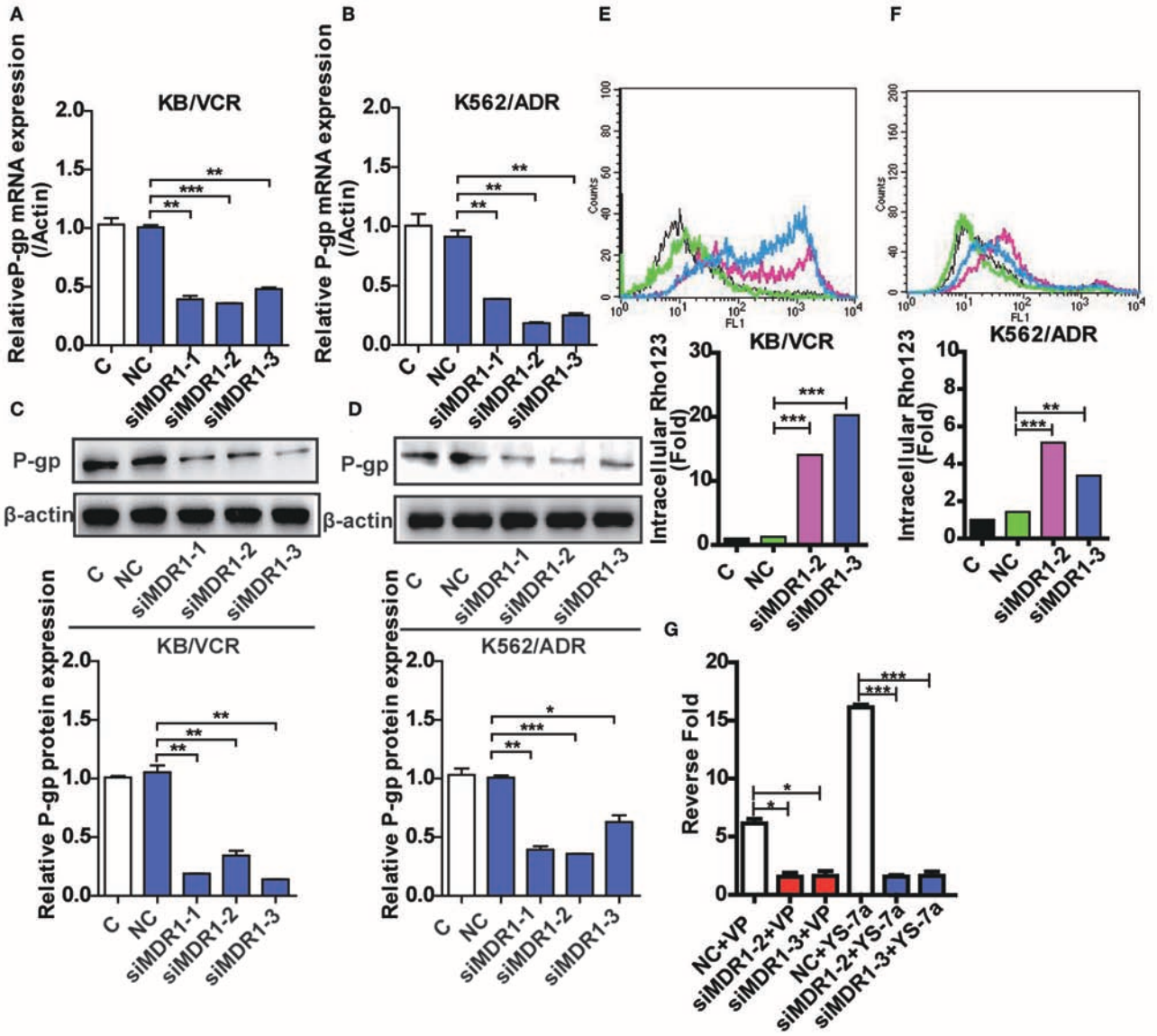


FIGURE 4 | Confirmation of the target of YS-7a in MDR cells. The expression of P-45 gp mRNA and protein in *MDR1* knockdown **(A,C)** KB/VCR cells and **(B,D)** K562/ADR cells. Changes in P-gp transporter function in *MDR1* knockdown **(E)** KB/VCR cells and **(F)** K562/ADR cells. **(G)** The target of YS-7a was verified through the fold change in drug resistance reversal in KB/VCR cells using the MTT assay after *MDR1* knockdown. All results were repeated at least three times. $^{*}P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$.

P-gp inhibitors are generally classified based on their inhibition mechanism: inhibiting the substrate binding site, interfering with ATP hydrolysis, or altering the integrity or fluidity of cell membrane lipids, which inhibits P-gp structural transformation (Shapiro and Ling, 1997; Varma et al., 2003; Drori et al., 2010). Most P-gp inhibitors inhibit the substrate binding site. Based on these reports, we screened P-gp inhibitors by performing Rho123 efflux experiments (**Figure 3**). Furthermore, MDR cancer cells show overexpression or excessive activation of P-gp, and P-gp inhibitors may inhibit the expression or function of P-gp (Silva et al., 2015). Therefore, the effects of YS-7a on P-gp function and expression were measured (**Figure 5**). Our results showed that YS-7a suppressed P-gp transport function without affecting its expression. In addition, P-gp ATPase activity is affected by various drug resistance regulators, such as VP (Sharom et al., 1995). The YS-7a inhibition effect of P-gp ATPase was reflected by ATP consumption measurement using the P-gp-Glo™ Assay Systems Kit (**Figure 5F**). Our results suggest that YS-7a inhibits the P-gp substrate binding site, similar to VP; however, further studies are required to verify its mechanism of action.

P-gp and CYP3A4 play important roles in reducing intracellular concentrations of xenobiotics and drug absorption through their respective roles in xenobiotic excretion and

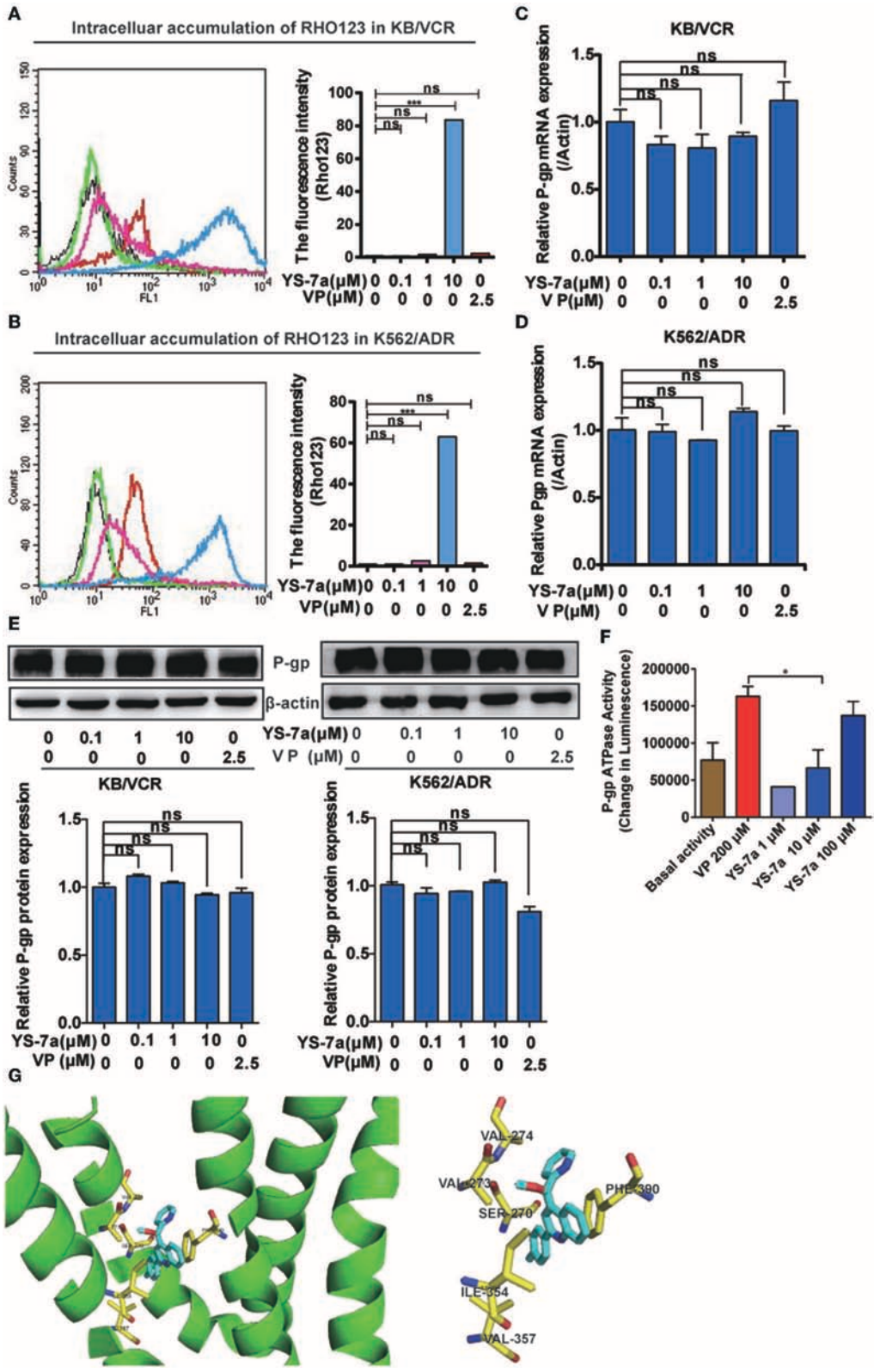


FIGURE 5 | Mechanism of P-gp inhibition by YS-7a. Intracellular accumulation of Rho123 in **(A)** KB/VCR and **(B)** K562/ADR cells after exposure to YS-7a. Effect of YS-7a on the expression of P-gp **(C,D)** mRNA and **(E)** protein in KB/VCR and K562/ADR cells. **(F)** Effect of YS-7a on P-gp ATPase activity after different drug concentrations were incubated with recombinant P-gp protein; VP was used as the positive control and basal activity as the negative control. **(G)** Molecular docking of YS-7a with P-gp (3WME); the yellow amino acid residues represent a distance of 1 angstrom or less, $^{***}P < 0.001$.

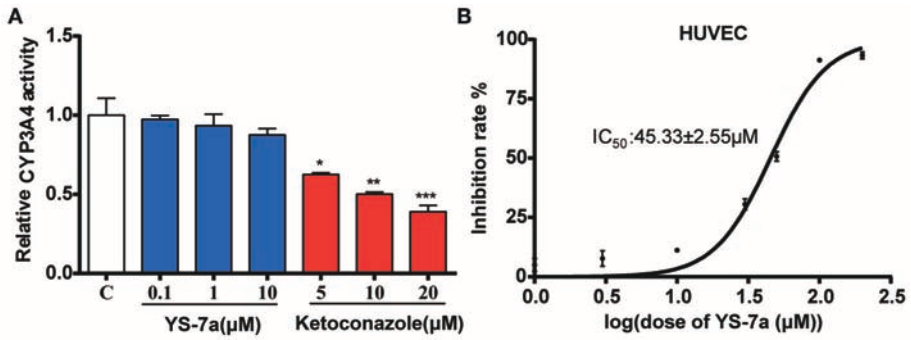


FIGURE 6 | Potential toxicity of YS-7a *in vitro*. **(A)** Effect of different concentrations of YS-7a on CYP3A4 activity. **(B)** The 72-h cytotoxicity of YS-7a in HUVECs measured using the MTT assay. All experiments were repeated at least three times. $^{*}P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$.

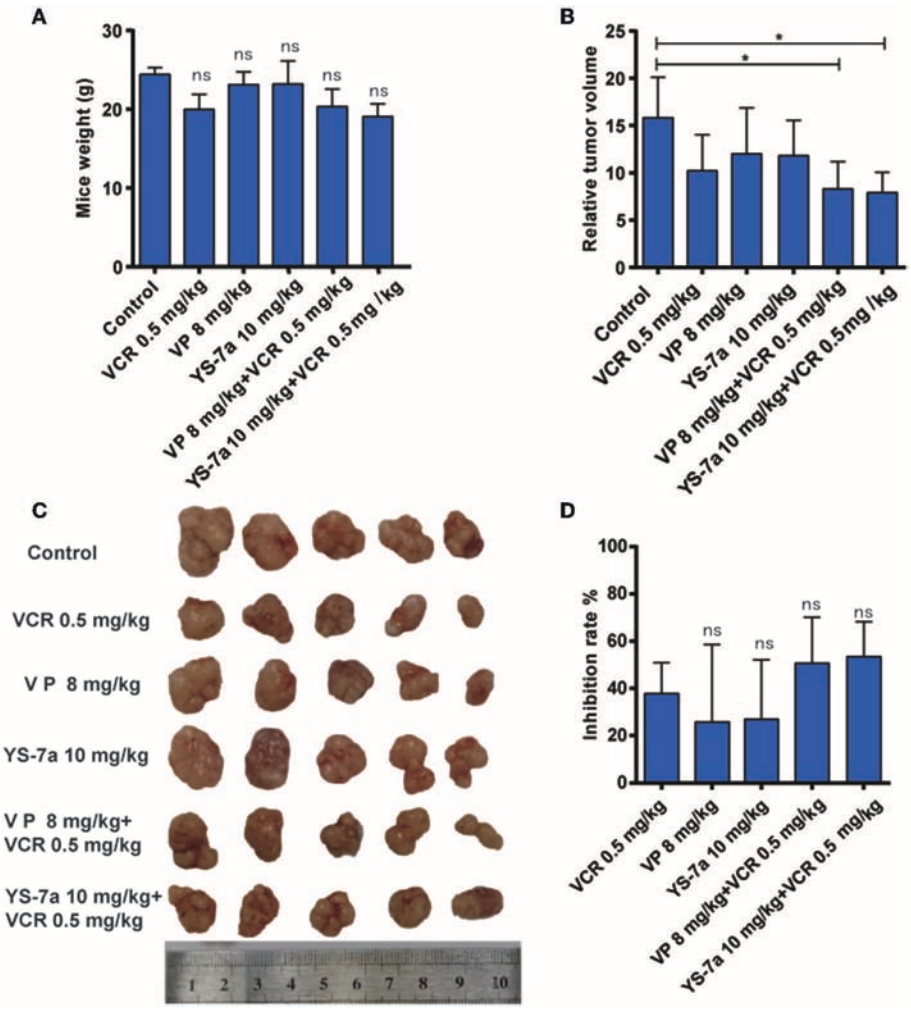


FIGURE 7 | *In vivo* drug resistance reversal effect of YS-7a in KB/VCR xenograft nude mice. **(A)** The weight of KB/VCR xenograft nude mice in all groups after treatment for 24 days. **(B)** The relative tumor volume of KB/VCR xenograft nude mice in all treatment groups after 24 days. **(C)** Images of tumors from the KB/VCR xenograft nude mice in every treatment group. **(D)** Tumor inhibition rate in every treatment group after 24 days, $^{*}P < 0.05$.

metabolism. P-gp and CYP3A4 work in coordination, given their co-localization in intestinal epithelial tissue and similarly overlapping substrates (Watkins, 1997; Katoh et al., 2001; Oliver et al., 2004). Several studies have reported that P-gp inhibitors inhibit CYP3A4, leading to unexpected toxicity (Wacher et al., 1998; Mathias et al., 2015). To predict potential drug–drug interactions, we measured the inhibitory effects of YS-7a on CYP3A4 activity (**Figure 6**). Our results suggested that YS-7a did not inhibit CYP3A4 activity. Therefore, YS-7a may have fewer side effects than PSC-833 and dexverapamil.

As shown *in vitro* (**Figures 3–6**), the novel P-gp inhibitor YS-7a showed superior reversal effects compared to VP by binding directly to P-gp. However, confirming whether YS-7a binds at the sites predicted by molecular docking, requires further study. Various methods, such as mutating the corresponding sites or radioisotope tracing (Hrycyna et al., 1999; Tsujimura et al., 2008), could be applied. *In vivo*, the drug resistance reversal effect of YS-7a was relatively low (about 50%). However, YS-7a at a dose of 10 mg/kg did not significantly decrease mouse body weight, suggestive of little-to-no toxicity or side effects. Thus, YS-7a may exhibit better reversal effects at higher doses. Many reports (Krepler et al., 2016; Vaidhyanathan et al., 2016) have shown that the patient-derived xenograft (PDX) model is ideal to evaluate the efficiency and toxicity of small-molecule inhibitors *in vivo*. Therefore, future studies should apply the PDX model to confirm the drug resistance reversal effect of YS-7a. Moreover, the pharmacokinetics of YS-7a should be explored to investigate its potential therapeutic mechanism in future studies.

CONCLUSIONS

We obtained a novel potent quinoline P-gp inhibitor derived from NSC23925, which showed a cancer MDR reversal effect both *in vitro* and *in vivo*. First, 11 novel quinoline compounds were synthesized, and potential P-gp inhibitors were screened using the classic screening model. YS-7a showed a significant inhibition effect against cellular Rho123 efflux. The MDR reversal effect and potential mechanisms of YS-7a were verified *in vitro*. YS-7a suppressed the P-gp transport function without affecting its expression, by stimulating the ATPase activity of P-gp in a dose-dependent manner instead. In addition, potential binding sites were predicted based on molecular docking. Finally, *in vitro* experiments support the low toxicity of YS-7a and the MDR reversal effect of YS-7a was verified in a KB/VCR cancer xenograft model with minimal toxicity. Overall, these results suggest that YS-7a may be a potential candidate compound for the development for new agents to reverse cancer MDR.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

This animal study was reviewed and approved by China Pharmaceutical University.

AUTHOR CONTRIBUTIONS

XQ, HD, JX, XY, LSh, and LSu defined the research subject and its aims, conceived, and designed the experiments. XQ, HD, JX, XH, XG, YZ, LSh, and YW conducted the experiments. XQ, HD, JJ, LL, and SY analyzed the data and wrote the paper.

FUNDING

This work was supported by National Key Research and Development Program of China (No. 2017YFA0205200), the National Natural Science Foundation of China (No. 21702051, 81773766, and 81903845), the Double First-Class University Project (No. CPU2018GY38), the Natural Science Foundation of Jiangsu Province (BK20181330) and the Key Scientific Research Project of Henan Province (No. 18A150009 and 17A350006).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00820/full#supplementary-material

REFERENCES

Alfarouk, K. O., Stock, C. M., Taylor, S., Walsh, M., Muddathir, A. K., Verduzco, D., et al. (2015). Resistance to cancer chemotherapy: failure in drug response from ADME to P-gp. *Cancer Cell Int.* 15:71. doi: 10.1186/s12935-015-0221-1

Bellamy, W. T., Dalton, W. S., Kailey, J. M., Gleason, M. C., Mccloskey, T. M., Dorr, R. T., et al. (1988). Verapamil reversal of doxorubicin resistance in multidrug-resistant human myeloma cells and association with drug accumulation and DNA damage. *Cancer Res.* 48, 6365–6370.

Boesch, D., Muller, K., Pourtiermanzanedo, A., and Loor, F. (1991). Restoration of daunomycin retention in multidrug-resistant P388 cells by submicromolar concentrations of SDZ PSC 833, a nonimmunosuppressive cyclosporin derivative. *Exp. Cell Res.* 196, 26–32. doi: 10.1016/0014-4827(91)90452-Z

Cao, Y., Gong, Y., Liu, L., Zhou, Y., Fang, X., Zhang, C., et al. (2017). The use of human umbilical vein endothelial cells (HUVECs) as an *in vitro* model to assess the toxicity of nanoparticles to endothelium: a review. *J. Appl. Toxicol.* 37, 1359–1369. doi: 10.1002/jat.3470

Cha, Y. J., Lee, H., Gu, N., Kim, T. E., Lim, K. S., Yoon, S. H., et al. (2013). Sustained increase in the oral bioavailability of loperamide after a single oral dose of HM30181, a P-glycoprotein Inhibitor, in healthy male participants. *Basic Clin. Pharmacol. Toxicol.* 113, 419–424. doi: 10.1111/bcpt.12108

Chico, I., Kang, M. H., Bergan, R., Abraham, J., Bakke, S., Meadows, B., et al. (2001). Phase I study of infusional paclitaxel in combination with the P-glycoprotein antagonist PSC 833. *J. Clin. Oncol.* 19, 832–842. doi: 10.1200/JCO.2001.19.3.832

Choi, Y. H., and Yu, A. M. (2014). ABC transporters in multidrug resistance and pharmacokinetics, and strategies for drug development. *Curr. Pharm. Des.* 20, 793–807. doi: 10.2174/13816128200514021652 12

Coley, H. M. (2010). Overcoming multidrug resistance in cancer: clinical studies of p-glycoprotein inhibitors. *Methods Mol. Biol.* 596, 341–358. doi: 10.1007/978-1-60761-416-6_15

Dantzig, A. H., Shepard, R. L., Law, K. L., Tabas, L., Pratt, S., Gillespie, J. S., et al. (1999). Selectivity of the multidrug resistance modulator, LY335979, for P-glycoprotein and effect on cytochrome P-450 activities. *J. Pharmacol. Exp. Ther.* 290, 854–862.

Drori, S., Eytan, G. D., and Assaraf, Y. G. (2010). Potentiation of anticancer-drug cytotoxicity by multidrug-resistance chemosensitizers involves alterations in membrane fluidity leading to increased membrane permeability. *FEBS J.* 228, 1020–1029. doi: 10.1111/j.1432-1033.1995.tb20352.x

Du, H., Huang, Y., Hou, X., Quan, X., Jiang, J., Wei, X., et al. (2018). Two novel camptothecin derivatives inhibit colorectal cancer proliferation via induction of cell cycle arrest and apoptosis *in vitro* and *in vivo*. *Eur. J. Pharm. Sci.* 123, 546–559. doi: 10.1016/j.ejps.2018.08.018

Duan, Z., Choy, E., and Hornicek, F. J. (2009). NSC23925, identified in a high-throughput cell-based screen, reverses multidrug resistance. *PLoS ONE* 4:e7415. doi: 10.1371/journal.pone.0007415

Duan, Z., Li, X., Huang, H., Yuan, W., Zheng, S.-L., Liu, X., et al. (2012). Synthesis and evaluation of (2-(4-Methoxyphenyl)-4-quinolinyl)(2-piperidinyl)methanol (NSC23925) isomers to reverse multidrug resistance in cancer. *J. Med. Chem.* 55, 3113–3121. doi: 10.1021/jm300117u

Federica, D. N., Knight, L. A., Sharon, G., Whitehouse, P. A., Mercer, S. J., Sanjay, S., et al. (2004). *Ex vivo* reversal of chemoresistance by tariquidar (XR9576). *Anticancer. Drugs* 15, 861–869. doi: 10.1097/00001813-200410000-00006

Fojo, A. T., Ueda, K., Slamon, D. J., Poplack, D. G., Gottesman, M. M., and Pastan, I. (1987). Expression of a multidrug-resistance gene in human tumors and tissues. *Proc. Natl. Acad. Sci. U.S.A.* 84, 265–269. doi: 10.1073/pnas.84.1.265

Fox, E., and Bates, S. E. (2007). Tariquidar (XR9576): a P-glycoprotein drug efflux pump inhibitor. *Expert Rev. Anticancer Ther.* 7, 447–459. doi: 10.1586/14737140.7.4.447

Hou, X., Du, H., Quan, X., Shi, L., Zhang, Q., Wu, Y., et al. (2018a). Silibinin inhibits NSCLC metastasis by targeting the EGFR/LOX pathway. *Front. Pharmacol.* 9:21. doi: 10.3389/fphar.2018.00021

Hou, X., Du, H., Yang, R., Qi, J., Huang, Y., Feng, S., et al. (2018b). The antitumor activity screening of chemical constituents from Camellia nitidissima Chi. *Int. J. Mol. Med.* 41, 2793–2801. doi: 10.3892/ijmm.2018.3502

Hrycyna, C. A., Ramachandra, M., Germann, U. A., Cheng, P. W., Pastan, I., and Gottesman, M. M. (1999). Both ATP sites of human P-glycoprotein are essential but not symmetric. *Biochemistry* 38, 13887–13899. doi: 10.1021/bi991115m

Katoh, M., Nakajima, M., Yamazaki, H., and Yokoi, T. (2001). Inhibitory effects of CYP3A4 substrates and their metabolites on P-glycoprotein-mediated transport. *Eur. J. Pharm. Sci.* 12, 505–513. doi: 10.1016/S0928-0987(00)00215-3

Köhler, S. C., and Wiese, M. (2015). HM30181 derivatives as novel potent and selective inhibitors of the breast cancer resistance protein (BCRP/ABCG2). *J. Med. Chem.* 58, 3910–3921. doi: 10.1021/acs.jmedchem.5b00188

Konstantinos, B. (2015). Structural basis for the mechanism of ABC transporters. *Biochem. Soc. Trans.* 43, 889–893. doi: 10.1042/BST20150047

Krepler, C., Xiao, M., Katrin, S., Brafford, P., Shannan, B., Beqiri, M., et al. and Herlyn, M. (2016). Personalized pre-clinical trials in BRAF inhibitor resistant patient derived xenograft models of melanoma identify c-Met as an effective second line combination therapy target. *Clin. Cancer Res.* 75, 1592–1602. doi: 10.1158/1078-0432.CCR-15-1762

Krishna, R., and Mayer, L. D. (2000). Multidrug resistance (MDR) in cancer-mechanisms, reversal using modulators of MDR and the role of MDR modulators in influencing the pharmacokinetics of anticancer drugs. *Eur. J. Pharm. Sci.* 11, 265–283. doi: 10.1016/S0928-0987(00)00114-7

Kumar, A., and Jaitak, V. (2019). Natural products as multidrug resistance modulators in cancer. *Eur. J. Med. Chem.* 176, 268–291. doi: 10.1016/j.ejmech.2019.05.027

Kusunoki, N., Takara, K., Tanigawara, Y., Yamauchi, A., Ueda, K., Komada, F., et al. (2010). Inhibitory effects of a cyclosporin derivative, SDZ PSC 833, on transport of doxorubicin and vinblastine via human P-glycoprotein. *Cancer Sci.* 89, 1220–1228. doi: 10.1111/j.1349-7006.1998.tb00518.x

Labrie, P., Maddaford, S. P., Lacroix, J., Catalano, C., Lee, D. K. H., Rakhit, S., et al. (2006). *In vitro* activity of novel dual action MDR anthranilamide modulators with inhibitory activity at CYP-450. *Bioorg. Med. Chem.* 14, 7972–7987. doi: 10.1016/j.bmc.2006.07.055

Lapidus, J. B., and Fauley, J. J. (1971). Absolute configuration of the phenyl 2-piperidylcarbinols. *J. Organic Chem.* 36, 3065–3067. doi: 10.1021/jo00819a045

Mathias, T. J., Natarajan, K., Shukla, S., Doshi, K. A., Singh, Z. N., Ambudkar, S. V., et al. (2015). The FLT3 and PDGFR inhibitor crenolanib is a substrate of the multidrug resistance protein ABCB1 but does not inhibit transport function at pharmacologically relevant concentrations. *Invest. New Drugs* 33, 300–309. doi: 10.1007/s10637-015-0205-y

Oliver, V. R., Oliver, B., Fromm, M. F., Thon, K. P., Michel, E., and Kivistö, K. T. (2004). Cytochrome P450 3A4 and P-glycoprotein expression in human small intestinal enterocytes and hepatocytes: a comparative analysis in paired tissue specimens. *Clin. Pharm. Ther.* 75, 172–183. doi: 10.1016/j.clpt.2003.10.008

Pirker, R., Zöchbauer, S., Kupper, H., Laßmann, A., Gsur, A., Frass, M., et al. (1995). Dexverapamil as resistance modifier in acute myeloid leukaemia. *J. Cancer Res. Clin. Oncol.* 3, 21–24. doi: 10.1007/BF02351067

Shapiro, A. B., and Ling, V. (1997). Effect of quercetin on Hoechst 33342 transport by purified and reconstituted P-glycoprotein. *Biochem. Pharmacol.* 53, 587–596. doi: 10.1016/S0006-2952(96)00826-X

Sharom, F. J., Yu, X., Chu, J. W., and Doige, C. A. (1995). Characterization of the ATPase activity of P-glycoprotein from multidrug-resistant Chinese hamster ovary cells. *Biochem. J.* 308(Pt 2), 381–390. doi: 10.1042/bj3080381

Shepard, R. L., Cao, J., Starling, J. J., and Dantzig, A. H. (2003). Modulation of P-glycoprotein but not MRP1- or BCRP-mediated drug resistance by LY335979. *Int. J. Cancer* 103, 121–125. doi: 10.1002/ijc.10792

Silva, R., Vilas-Boas, V., Carmo, H., Dinis-Oliveira, R. J., Carvalho, F., De Lourdes Bastos, M., et al. (2015). Modulation of P-glycoprotein efflux pump: induction and activation as a therapeutic strategy. *Pharmacol. Ther.* 149, 1–123. doi: 10.1016/j.pharmthera.2014.11.013

Solladié-Cavallo, A., Marsol, C., Yaakoub, M., Azyat, K., Klein, A., Roje, M., et al. (2003). erythro-1-Naphthyl-1-(2-piperidyl)methanol: synthesis, resolution, NMR relative configuration, and VCD absolute configuration. *J. Org. Chem.* 68, 7308–7315. doi: 10.1021/jo0345502

Thürlimann, B., Kröger, N., Greiner, J., Mross, K., Schüller, J., Schernhammer, E., et al. (1995). Dexverapamil to overcome epirubicin resistance in advanced breast cancer. *J. Cancer Res. Clin. Oncol.* 121, R3–R6. doi: 10.1007/BF02351063

Tsujimura, S., Saito, K., Nawata, M., Nakayamada, S., and Tanaka, Y. (2008). Overcoming drug resistance induced by P-glycoprotein on lymphocytes in patients with refractory rheumatoid arthritis. *Ann. Rheum. Dis.* 67, 380–388. doi: 10.1136/ard.2007.070821

Vaidhyanathan, S., Wilken-Resman, B., Ma, D. J., Parrish, K. E., Mittapalli, R. K., Carlson, B. L., et al. (2016). Factors influencing the central nervous system distribution of a novel phosphoinositide 3-kinase/mammalian target of rapamycin inhibitor GSK2126458: implications for overcoming resistance with combination therapy for melanoma brain metastases. *J. Pharmacol. Exp. Ther.* 356, 251–259. doi: 10.1124/jpet.115.229393

Varma, M. V., Ashokraj, Y., Dey, C. S., and Panchagnula, R. (2003). P-glycoprotein inhibitors and their screening: a perspective from bioavailability enhancement. *Pharmacol. Res.* 48, 347–359. doi: 10.1016/S1043-6618(03)00158-0

Wacher, V. J., Silverman, J. A., Zhang, Y., and Benet, L. Z. (1998). Role of P-glycoprotein and cytochrome P450 3A in limiting oral absorption of peptides and peptidomimetics. *J. Pharm. Sci.* 87, 1322–1330. doi: 10.1021/js980082d

Wang, E., Casciano, C. N., Clement, R. P., and Johnson, W. W. (2000). *In vitro* flow cytometry method to quantitatively assess inhibitors of P-glycoprotein. *Drug Metab. Dispos.* 28, 522–528.

Watkins, P. B. (1997). The barrier function of CYP3A4 and P-glycoprotein in the small bowel. *Adv. Drug Deliv. Rev.* 27, 161. doi: 10.1016/S0169-409X(97)00041-0

Wei, X., Mao, T., Li, S., He, J., Hou, X., Li, H., et al. (2019). DT-13 inhibited the proliferation of colorectal cancer via glycolytic metabolism and AMPK/mTOR signaling pathway. *Phytomedicine* 54, 120–131. doi: 10.1016/j.phymed.2018.09.003

Wu, Q., Yang, Z., Nie, Y., Shi, Y., and Fan, D. (2014). Multi-drug resistance in cancer chemotherapeutics: Mechanisms and lab approaches. *Cancer Lett.* 347, 159–166. doi: 10.1016/j.canlet.2014.03.013

Yang, X., Feng, Y., Gao, Y., Shen, J., Choy, E., Cote, G., et al. (2015). NSC23925 prevents the emergence of multidrug resistance in ovarian cancer *in vitro* and *in vivo*. *Gynecol. Oncol.* 137, 134–142. doi: 10.1016/j.ygyno.2015.02.002

Yang, X., Yang, P., Shen, J., Osaka, E., Choy, E., Cote, G., et al. (2014). Prevention of multidrug resistance (MDR) in osteosarcoma by NSC23925. *Br. J. Cancer* 110, 2896–2904. doi: 10.1038/bjc.2014.254

Yusa, K., and Tsuruo, T. (1989). Reversal mechanism of multidrug resistance by verapamil: direct binding of verapamil to P-glycoprotein on specific sites and transport of verapamil outward across the plasma membrane of K562/ADM cells. *Cancer Res.* 49, 5002–5006.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Quan, Du, Xu, Hou, Gong, Wu, Zhou, Jiang, Lu, Yuan, Yang, Shi and Sun. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

REVIEW
published: 08 January 2020
doi: 10.3389/fchem.2019.00873

In silico Methods for Design of Kinase Inhibitors as Anticancer Drugs

Zarko Gagic[1], *Dusan Ruzic*[2], *Nemanja Djokovic*[2], *Teodora Djikic*[2] *and Katarina Nikolic*[2]*

[1] Department of Pharmaceutical Chemistry, Faculty of Medicine, University of Banja Luka, Banja Luka, Bosnia and Herzegovina, [2] Department of Pharmaceutical Chemistry, Faculty of Pharmacy, University of Belgrade, Belgrade, Serbia

Rational drug design implies usage of molecular modeling techniques such as pharmacophore modeling, molecular dynamics, virtual screening, and molecular docking to explain the activity of biomolecules, define molecular determinants for interaction with the drug target, and design more efficient drug candidates. Kinases play an essential role in cell function and therefore are extensively studied targets in drug design and discovery. Kinase inhibitors are clinically very important and widely used antineoplastic drugs. In this review, computational methods used in rational drug design of kinase inhibitors are discussed and compared, considering some representative case studies.

Keywords: kinase inhibitors, rational drug design, molecular modeling, drug discovery, pharmacophore

OPEN ACCESS

Edited by:
Simone Brogi,
University of Pisa, Italy

Reviewed by:
Elif Ozkirimli,
Bogaziçi University, Turkey
Anna Maria Almerico,
University of Palermo, Italy
Mahmoud A. Al-Sha'er,
Zarqa Private University, Jordan
Chandrabose Selvaraj,
Central European Institute of Technology, Brno University of Technology, Czechia

***Correspondence:**
Katarina Nikolic
knikolic@pharmacy.bg.ac.rs

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *28 September 2019*
Accepted: *04 December 2019*
Published: *08 January 2020*

Citation:
Gagic Z, Ruzic D, Djokovic N, Djikic T and Nikolic K (2020) In silico Methods for Design of Kinase Inhibitors as Anticancer Drugs. Front. Chem. 7:873. doi: 10.3389/fchem.2019.00873

KINASES AS TARGETS FOR DEVELOPING ANTICANCER DRUGS

Kinases belong to a large family of enzymes that catalyze transfer of high energy phosphate group from adenosine triphosphate (ATP) to substrates, such as proteins (the protein-tyrosine kinases, the serine-threonine specific kinases), lipids (phosphatidylinositol kinases, sphingosine kinases), carbohydrates, and nucleic acids (Duong-Ly and Peterson, 2013). Phosphorylation of the substrate modulates its activity and/or interaction with other molecules leading to different physiological responses. It is estimated that 50% of all proteins are constantly undergoing reversible phosphorylation and dephosphorylation, which emphasizes the role of protein kinases in almost all aspects of cell function, including proliferation, cell growth, apoptosis, and signal transduction (Graves and Krebs, 1999; Manning et al., 2002).

Dysregulated, overexpressed, or mutated protein kinases are found in many diseases, including cancer, and over the past two decades they became extensively examined targets for the development of new antineoplastic drugs (Blume-Jensen and Hunter, 2001; Cohen, 2002). There are 53 kinase inhibitors (KIs) currently approved by the FDA (FDA, 2019), while over 200 potential inhibitors are in different phases of clinical trials worldwide (Carles et al., 2018). Majority of the approved drugs are orally active and effective against various malignancies (**Table 1**; Roskoski, 2019a,b).

Structures of the selected KIs commonly used for treatment of cancer are shown in **Figure 1**. These drugs target different protein kinases that are frequently upregulated in cancer cells. The epidermal growth factor receptor (EGFR) is a member of the ErbB family of tyrosine kinase receptors that is overexpressed or mutated in non-small cell lung cancer and represents the primary target for drugs such erlotinib and gefitinib (Bethune et al., 2010). Lapatinib and neratinib bind to intracellular domain of the human epidermal growth factor receptor 2 (HER2/neu), another member of the ErbB tyrosine kinases, which elevated levels are found in approximately 20–30% of breast cancers (Collins et al., 2019). Imatinib possesses activity against non-receptor breakpoint

cluster region (Bcr)-Abelson leukemia virus (Abl) tyrosine kinase that is formed as a result of a chromosome rearrangement and has been implicated in the pathogenesis of nearly all cases of chronic myeloid leukemia (CML) and acute lymphoblastic leukemia with the Philadelphia chromosome (Iqbal and Iqbal, 2014). Although imatinib is a relatively specific Bcr-Abl inhibitor, it also inhibits the CD117 tyrosine kinase associated with gastrointestinal stromal tumors and has consequently been approved for this indication (Buchdunger et al., 2000). The vascular endothelial growth factor family of receptors (VEGFR) contains a tyrosine kinase domain which activation can lead to induction of signaling pathways that regulate cell proliferation, survival, and promotion of tumor angiogenesis (Morabito et al., 2006). Agents that target VEGFR, including lenvatinib, sorafenib and vandetanib, are frequently used for treatment of thyroid cancers. Vemurafenib, dabrafenib, and encorafenib target BRAF, a serine/threonine protein kinase which mutation is expressed at about 50–60% of cutaneous melanomas where it leads to continuous activation of mitogen-activated protein kinase (MAPK) pathway and uncontrolled proliferation of cancer cells (Yu et al., 2019).

TABLE 1 | Therapeutic indications of selected FDA-approved protein kinase inhibitors.

Therapeutic indication	Drug
Breast cancer	Everolimus, lapatinib, neratinib, palbociclib, ribociclib
Non-small cell lung cancer	Afatinib, alectinib, brigatinib, ceritinib, crizotinib, dabrafenib, dacomitinib, erlotinib, gefitinib, lorlatinib, osimertinib
Leukemia	Bosutinib, dasatinib, gilteritinib, ibrutinib, imatinib, midostaurin, nilotinib, ponatinib
Melanoma	Binimetinib, cobimetinib, dabrafenib, encorafenib, trametinib, vemurafenib
Thyroid cancer	Cabozantinib, lenvatinib, vandetanib
Renal cancer	Axitinib, pazopanib, sorafenib, temsirolimus
Gastrointestinal cancer	Regorafenib, sunitinib

Structure of Protein Kinases

The human genome encodes at least 518 protein kinases (Manning et al., 2002). Out of them, 478 share highly conserved catalytic domains. The remaining 40 do not share the sequence similarity, but their folding is similar to the folding of "typical" PKs (Caballero and Alzate-Morales, 2012). In 1991, Knighton solved the X-ray structure of cyclic AMP-dependent PK and described its structure for the first time. This description can apply to all currently known protein kinases. The characteristic architecture of the catalytic domain of PK consists of a small, amino-terminal N-lobe and a large α-helical carboxy-terminal C-lobe which are connected with a small hinge region (**Figure 2**; Knighton et al., 1991). The N-terminal lobe is dominated by five β-strands (β1–β5) and one conserved α-helix (helix C) that occurs in active (αC-in) or inactive (αC-out) orientations. The C-lobe consists of eight α-helices and four short conserved β-strands (β6–β9) which include residues that participate in the phosphorylation of protein substrates. The small and large lobes form a catalytic cleft where ATP binds (Knighton et al., 1991; Roskoski, 2019a). The hydrophobic residues of the cleft form a binding pocket for ATP. The charged residues in the active site bind and position the γ-phosphate of ATP and divalent cation and take part in the catalysis (Knight et al., 2007).

Conserved residues play crucial roles in positioning ATP, stabilizing the active-conformation and in the catalytic mechanism, and they are mostly found in and around the active site but also in other parts of the protein kinase domain (Knight et al., 2007). Almost all protein kinases possess a conserved K/E/D/D (Lys/Glu/Asp/Asp) signature that is important for the

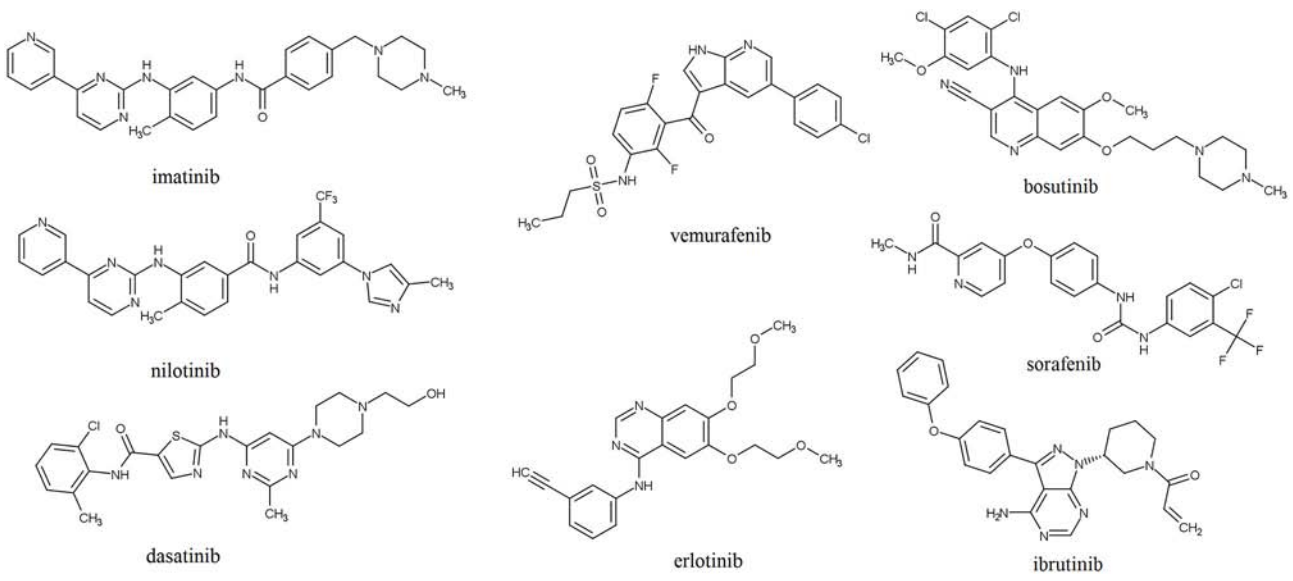


FIGURE 1 | Structures of selected protein kinase inhibitors that have been approved for clinical use.

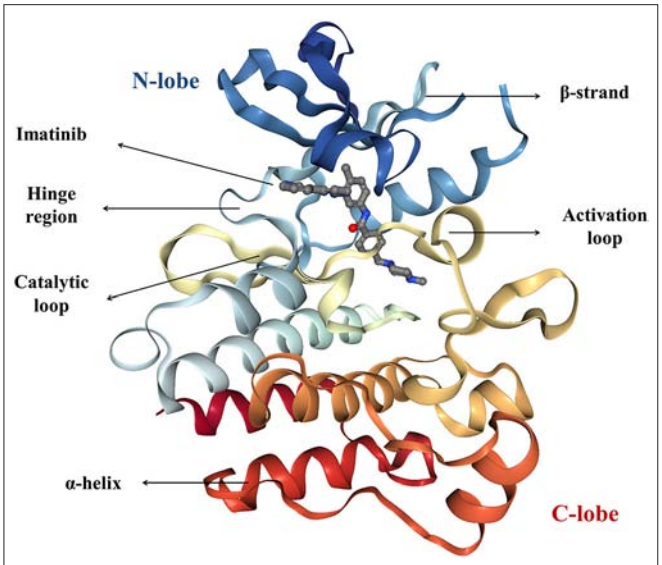


FIGURE 2 | The crystal structure of imatinib-bound form of the Abl kinase (PDB accession code: 2HYY), colored as rainbow from N-lobe (blue) to C-lobe (red). Imatinib is represented as ball and stick.

catalysis. Lysine and glutamic acid residues belong to the N-lobe, and the two aspartic acid residues are found in the C-lobe. Lysine residue binds to the α- and β-phosphates of ATP. Formation of the salt bridge between the carboxylate group of aspartic acid and the amino group of lysine stabilizes its interactions with the α- and β-phosphates, and it is required for kinase activation (Roskoski, 2015, 2019a,b).

The N-lobe contains a conserved flexible glycine-rich GxGxxG motif (also called P-loop) between β1 and β2 that folds over the nucleotide and places the γ-phosphate of ATP during the catalysis (Taylor and Kornev, 2011). As mentioned above, lysine from the β3-strand forms a salt bridge with the conserved glutamate near the center of the protein-kinase αC-helix which is necessary for the formation of the active enzyme, and this structure corresponds to the "αC-in" conformation (Roskoski, 2015, 2019a,b).

The C-lobe is important for both the protein-substrate binding as well as nucleotide binding (Roskoski, 2015). The C-lobe contains a mobile activation loop of 20–30 residues which can take open or closed conformation. The activation loop begins with the DFG motif (Asp-Phe-Gly) and extends up to an APE motif (Ala-Pro-Glu) (Modi and Dunbrack, 2019). In the active conformation a divalent metal ion, Mg^{2+} (or sometimes Mn^{2+}), interacts with a highly conserved aspartic acid residue from the DFG motif. It coordinates with the α and γ phosphates of ATP and facilitates the phosphorylation and coordinates the ATP binding (Adams, 2001). At the other end, glutamic acid from APE motif is fixed by the formation of a salt bridge with arginine from the C-lobe (Roskoski, 2015, 2019b). In addition to these, another motif on the C-lobe is highly conserved suggesting it plays an important role in the catalysis—HRD (rarely YRD) motif. The aspartate residue of this motif is required for the orientation of the hydroxyl group of the substrate peptide at the P-site and the transfer of the phosphoryl group. Arginine residue interacts with the phosphorylated activation segment thereby contributing to its correct orientation. Histidine (or in rare cases tyrosine) is considered to be involved in the maintenance of the conserved rigid organization of the catalytic core (La Sala et al., 2016).

The main differences between tyrosine kinases and serine/threonine kinases are found in the protein-substrate binding site. In serine/threonine kinases, the phosphorylatable serine or threonine of the protein substrate interacts with backbone residues near the end of the activation segment. Basic residues of the protein-substrate N-terminal interact with surface acidic residues of the C-lobe. Additionally, peptide substrate is fixed by serine in the glycine rich loop and lysine in the catalytic loop and also to threonine in the P+1 loop. These three residues are highly conserved in the majority of protein-serine/threonine kinases, and they are positioning the target hydroxyl group of a substrate in the catalytic cleft (near the γ-phosphate of ATP) where the phosphotransfer reaction happens (P-site). Since both serine and threonine hydroxyls are linked to the β-carbon, they have similar mechanisms of the catalysis. On the other hand, in tyrosine kinases, after DFG motif, there is a very stable region that contains the three tyrosine phosphorylation sites. The protein substrate chain positions in a manner that one of the tyrosines is oriented with its hydroxyl group lying in phosphorylation site P-site. The following tyrosine residue lies in the P+1 site. Proline residue interacts with the tyrosyl residue of the protein-substrate and is responsible for positioning the P-site tyrosine in the phosphotransfer site. The tyrosine ring is also positioned by Arg (Hubbard et al., 1994; Taylor et al., 1995; Roskoski, 2015).

Moreover, many protein kinases are regulated by different mechanisms such as dimerization, binding of allosteric effectors, or other modifications important for subcellular localization that can modulate their activity. Binding of an allosteric modulator leads to conformational changes that mostly involve structural reorganization of the activation loop, making it a primary end point of allosteric regulation. Effectors or regulatory subunits bind outside the catalytic site, causing the changes in loop conformation through conformational changes of other substructural elements. In most of the cases, regulators bind the αC helix at different locations, allowing control of catalysis from distal regions (Shi et al., 2006). Nevertheless, the αC helix is not the only allosteric binding site, in fact, they are very diverse (**Figure 3**; Ohren et al., 2004; Vanderpool et al., 2009; Jahnke et al., 2010; Martin et al., 2012; Park et al., 2015; Rettenmaier et al., 2015; Ung et al., 2018). Therefore, understanding the diversity of allosteric regulatory sites among the kinase superfamily gives a unique opportunity for the creation of novel selective allosteric kinase antagonists (Lamba and Ghosh, 2012).

Active/Inactive States

Basically, protein kinases reside in one active state and multiple inactive states (**Figure 4**). In active kinase conformation, activation loop forms a cleft that binds the substrate. When the substrate peptide binds, it interacts with the HRD motif (His-Arg-Asp). Asp from the DFG motif binds a magnesium ion that interacts directly with an oxygen atom of the β phosphate of

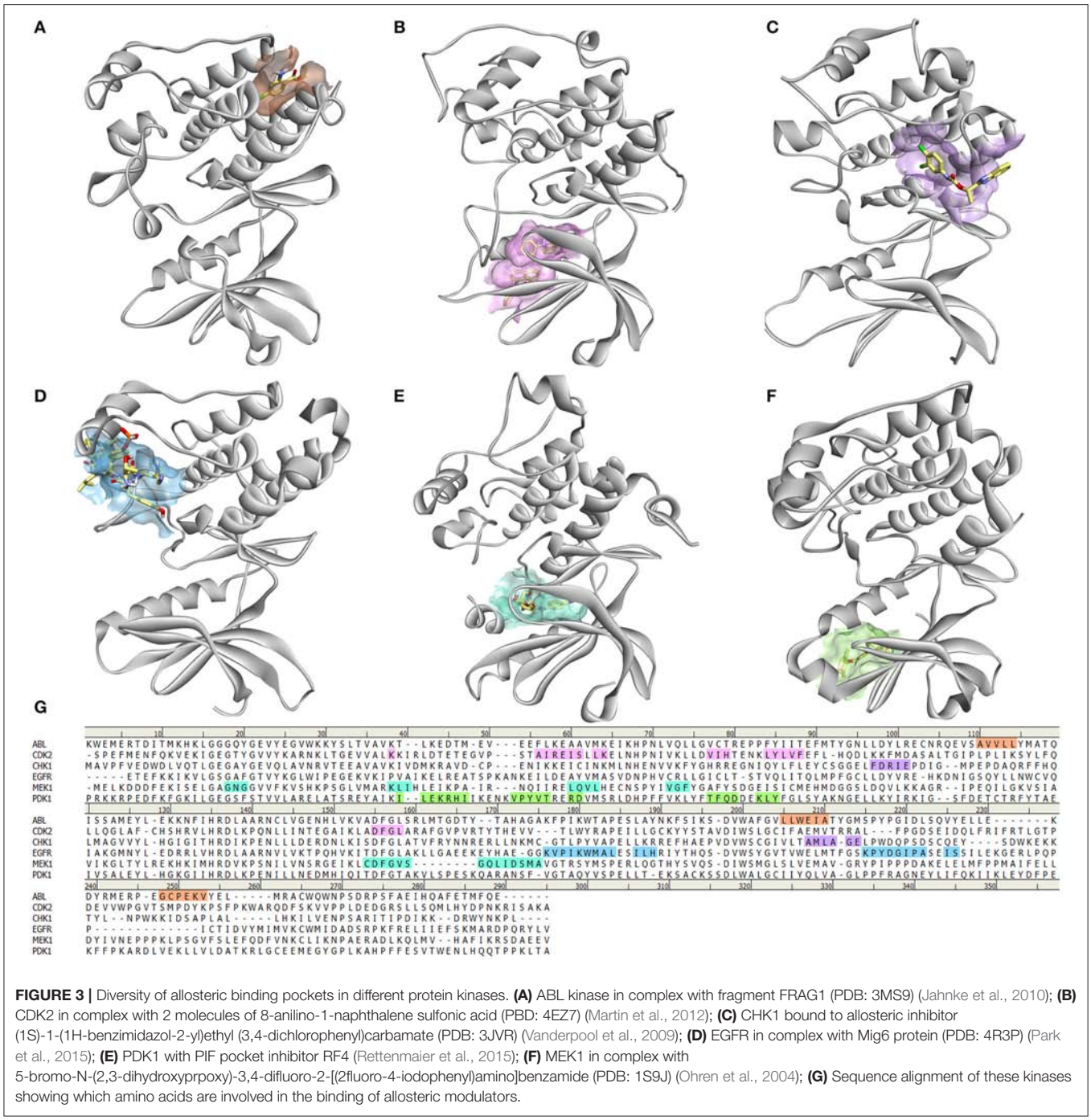


FIGURE 3 | Diversity of allosteric binding pockets in different protein kinases. **(A)** ABL kinase in complex with fragment FRAG1 (PDB: 3MS9) (Jahnke et al., 2010); **(B)** CDK2 in complex with 2 molecules of 8-anilino-1-naphthalene sulfonic acid (PBD: 4EZ7) (Martin et al., 2012); **(C)** CHK1 bound to allosteric inhibitor (1S)-1-(1H-benzimidazol-2-yl)ethyl (3,4-dichlorophenyl)carbamate (PDB: 3JVR) (Vanderpool et al., 2009); **(D)** EGFR in complex with Mig6 protein (PDB: 4R3P) (Park et al., 2015); **(E)** PDK1 with PIF pocket inhibitor RF4 (Rettenmaier et al., 2015); **(F)** MEK1 in complex with 5-bromo-N-(2,3-dihydroxyprpoxy)-3,4-difluoro-2-[(2fluoro-4-iodophenyl)amino]benzamide (PDB: 1S9J) (Ohren et al., 2004); **(G)** Sequence alignment of these kinases showing which amino acids are involved in the binding of allosteric modulators.

ATP. This is followed by formation of a salt bridge between the Glu from the C-helix with a Lys residue in the β3 strand. When the salt bridge is formed, the lysine side chain forms hydrogen bonds with oxygen atoms of α and β phosphates of ATP. The Glycine-rich Loop of the N-lobe stabilizes the phosphates of the bound ATP molecule during catalysis (Taylor and Kornev, 2011; Modi and Dunbrack, 2019). In an inactive conformation, usually the activation loop is blocking the substrate binding, and DFG motif is incompatible with the binding ATP and magnesium ion required for catalysis. Many attempts have been made in order to achieve classification for these conformations and to study interaction of inhibitors in different states (Mobitz, 2015; Ung et al., 2018; Modi and Dunbrack, 2019), and they are all based on the position of highly conserved DFG motif.

The most recent classification was published by Modi and Dunbrack. They have divided kinase structures into three clusters based on the spatial position of the DFG-Phe side chain into DFG-in, DFG-out, and DFG-inter (intermediate) conformations.

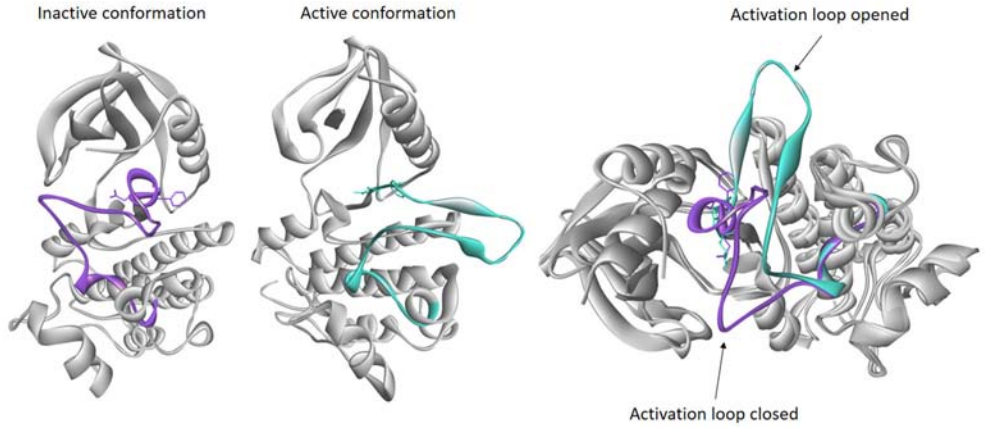


FIGURE 4 | Examples of inactive and active conformations of epidermal growth factor receptor (EGFR) kinase (PDB accession codes: 4HJO and 2GS6, respectively), and their superimposition. The activation loop of inactive conformation (purple) is closed therefore blocking the substrate to enter catalytic loop, while activation loop in active conformation (turquoise) is opened allowing the ATP to bind.

Each of these three clusters was further divided based on the dihedral angles required to place the Phe side chain, resulting in total of eight clusters: six for DFG-in and one cluster each for the DFG-out and DFG-inter groups. DFG-in represents the DFG motif orientations where DFG-Phe is packed against or under the C-helix. It contains many conformations, among all the typical DFG-in active conformation belongs to this group. DFG-out represents the structures where DFG-Phe is moved into the ATP binding pocket. DFG-inter represents the conformations in which the DFG-Phe side chain is out of the C-helix pocket but has not moved completely to a DFG-out conformation. Usually in this conformation DFG-Phe is pointing upward toward the β-sheets while dividing the active site into two halves (Modi and Dunbrack, 2019). This classification offers insight into active and inactive kinase conformations which are of great importance in structure-based design of kinase inhibitors.

Types of Kinase Inhibitors

Many reviewers have categorized KIs based on their binding modes into three classes, labeled as types I, II, and III kinase inhibitors (Roskoski, 2016; Bhullar et al., 2018). Type I inhibitors, such as gefitinib, bind to the active DFG-in conformation of a kinase in the phosphorylated ATP catalytic site, and they usually contain a heterocycle that mimic the purine ring of ATP. Considering that the ATP active site is highly conserved among different protein kinases, these ATP-competitive inhibitors display low selectivity profile which may lead to off-target side effects. While the physiological relevance of many off-target effects is still unclear, it was demonstrated that the lack of selectivity is connected with preclinical and clinical cardiotoxicity of kinase inhibitors (Force and Kolaja, 2011; Yang and Papoian, 2012). Possible mechanism behind the KI induced cardiotoxicity lies in binding of these drugs to colony-stimulating factor 1 receptor (CSF1R) (Hasinoff, 2010). Type II inhibitors, such as imatinib, bind to the inactive (DFG-out) conformation of a kinase in the unphosphorylated ATP catalytic site (Dar and Shokat, 2011). Zuccotto introduced type I½ inhibitors as compounds that bind to active ATP catalytic site as type I inhibitors but elongate into the back cavity of the ATP site giving rise to interactions specific for type II pharmacophore. These inhibitors represented by dasatinib, lapatinib, and vemurafenib, display higher selectivity profile as compared to Type I KIs (Zuccotto et al., 2010). Types III and IV are allosteric inhibitors that bind outside the ATP-binding site. Type III includes trametinib that binds to the allosteric site close to the ATP pocket, whereas Type IV inhibitors bind to a pocket distant from the ATP-binding site. Bivalent inhibitor spanning two regions of the protein kinase is termed as Type V (Wong et al., 2017), while KIs that form an irreversible covalent bond with the catalytic site represent Type VI inhibitors (afatinib and ibrutinib) (Roskoski, 2016).

Developing and Overcoming Resistance to Kinase Inhibitors

Despite the significant advances achieved by the use of protein kinase inhibitors, drug resistance remains one of the greatest challenges toward successful cancer treatment. Various mechanisms can underpin the development of resistance to KIs, including alterations in protein kinases, aberration of downstream pathways, or bypass mechanism that activates parallel signaling pathways (Holohan et al., 2013). Mutations of Bcr-Abl kinase domain were found in over 90% of patients with CML who relapsed after an initial response to imatinib. These mutations include different amino acid substitutions at the active site residues or changes in the kinase flexibility that impair its ability to adopt the inactive conformation required for optimal imatinib binding (Shah et al., 2002). Dasatinib is a novel Abl kinase inhibitor that can bind to both the active (mutated) and inactive (normal) conformations of Bcr-Abl, and its activity has been demonstrated in all imatinib-resistant CML patients, with the exception of those with the

T315I mutation that prevents formation of critical hydrogen bond (Shah et al., 2004; Burgess et al., 2005). Docking of dasatinib to three Bcr-Abl conformations (active, inactive, and intermediate inactive) showed that dasatinib binds preferentially to an active conformation, and that binding affinity significantly decreases when the kinase adopts inactive conformation (Laurini et al., 2013). Drug combinations targeting different upstream and downstream components within a single pathway, or targeting parallel kinase pathways, have been proved in clinical trials as an efficient method to overcame or delay therapeutic resistance. For instance, treatment with dabrafenib, a selective BRAF inhibitor, and trametinib, a selective MAPK kinase inhibitor, significantly improved progression-free survival of melanoma patients (Flaherty et al., 2012).

IN SILICO METHODS USED IN DRUG DESIGN

Since the approval of imatinib in 2001, protein kinases have received significant attention from academic and pharmaceutical companies, reflected in a large number of publications, solved crystal structures, and identified small molecule inhibitors for about one-fifth of the human kinome (Wu et al., 2015b). Considerable progress in this field is much owed to the use of computational methods that were able to provide valuable information on structural characteristic of both the kinase and the ligand that are important for favorable interaction and desired inhibitory activity (Agafonov et al., 2015). To design inhibitors for protein kinases it is necessary to understand the structure and dynamics of these enzymes, substrate recognition, and reaction of phosphorylation, product release as well as differences between active and inactive conformations.

There are two main approaches within the framework of computer-aided drug design (CADD): structure-based drug design (SBDD), and ligand-based drug design (LBDD). SBDD is based on structural information gathered from biological targets and includes *in silico* methods such as molecular docking, structure-based virtual screening (SBVS), and molecular dynamics (MD). In contrast, in the absence of information on targets, LBDD relies on the knowledge of ligands that interact with a specific target, and these methods include ligand-based virtual screening (LBVS), similarity searching, quantitative structure-activity relationship (QSAR) modeling, and pharmacophore generation (Ferreira et al., 2015). Over the last years, a large number of studies have reported successful use of CADD in design and discovery of new drugs (Lu et al., 2018b). In this study we provide the comprehensive review of computational tools that led to discovery, design and optimization of KIs as anticancer drugs.

Ligand-Based Methods in Drug Design

QSAR modeling involves the formation of a mathematical relationship between experimentally determined biological activity and quantitatively defined chemical characteristics that describe the analyzed molecule (descriptors) within a set of structurally similar compounds. The QSAR concept originated in the 1860s, when Crum-Brown and Fraser proposed the idea that the physiological action of a compound in a particular biological system is a function of its chemical constituent, while the modern era of QSAR modeling is associated with the work of Hansch et al. in the early 1960s (Hansch et al., 1962). The aim of the QSAR modeling is to utilize the information on structure and activity obtained from a relatively small series of data to ensure that the best lead compounds enter further studies, minimizing the time and the expense of drug development process (Cherkasov et al., 2014).

Classical 2D-QSAR models correlate physicochemical parameters, such as electronic, hydrophobic or steric characteristics of compounds, to biological activity, while the more advanced 3D-QSAR modeling adds quantum chemical parameters. One of the first approaches used in deriving 3D-QSAR models was CoMFA (comparative molecular field analysis). With this analysis, molecules were described with electrostatic and steric fields, which were correlated to biological activity by means of partial least squares regression (PLS) (Cramer et al., 1988). In addition to the steric and electrostatic descriptors, another approach used in deriving 3D-QSAR models was Comparative Molecular Similarity Index Analysis (CoMSIA). CoMSIA approach additionally uses three novel fields comparing to CoMFA, describing the ligand's hydrophobic properties, the presence of the hydrogen bond donors (HBD), and the presence of hydrogen bond acceptors (HBA) (Klebe et al., 1994). The main limitation of the CoMFA/CoMSIA methods is that they are largely dependent on the alignment of 3D-molecular structures which is often a slow process prone to subjectivity. Recently, modern QSAR programs that use new generation of 3D-descriptors, so-called grid-independent (GRIND) descriptors, have been developed and used for multivariate analyses and 3D-QSAR modeling (Pastor et al., 2000; Duran et al., 2009; Smajić et al., 2015; Gagic et al., 2016b).

Recent cases of reported QSAR studies aimed at providing useful information to guide the discovery of new potent KIs are listed in **Table 2**. Some of them will be discussed in this chapter.

Koneru et al. have used QSAR combined with molecular dynamics to redesign second-generation Src kinase inhibitor RL-45 in order to withstand the gatekeeper residue mutation

TABLE 2 | Selected studies that have used QSAR in the design of kinase inhibitors.

Target kinase	QSAR method	Software package	References
Mer	3D	Pentacle	Shiri et al., 2016
Lyn	2D	JMP	Naboulsi et al., 2018
HER2, EGFR	3D	SYBYL	de Angelo et al., 2018
EGFR	2D and 3D	SYBYL	Simeon et al., 2019
IKK-β	2D and 3D	Discovery studio; Schrödinger suite	Wang et al., 2019a
EGFR	3D	SYBYL	Zhao et al., 2019a
Src	3D	Vlife MDS	Koneru et al., 2019
VEGFR-2	3D	MOE	Mohamed et al., 2019
PKMYT1	2D	MOE	Najjar et al., 2019

and enhance binding affinity. They integrated fragment-based drug discovery (FBDD) technique with QSAR and molecular dynamics to assess novel Src kinase inhibitors. Newly designed compounds were assumed to be able to mitigate mutation-related Src kinase resistance and to bind more efficiently to the kinase active site and were proposed for further synthesis (Koneru et al., 2019). Wang et al. applied QSAR studies on a series of 2-acylamino-3-aminothienopyridine analogs in order to design new IKK-β inhibitors (Wang et al., 2019a). Obtained information on physicochemical, structural, electrostatic, and steric properties revealed that bulky aryl substituents at position C3 on the piperidine ring have favorable effect on activity, which led to the design of an in-house library. Compounds with best predicted activities were further subjected to docking studies. Based on these results two new compounds B01 and B02 were identified as potential IKK-β inhibitors, with predicted pIC_{50} activities of 7.18 and 7.17, and binding affinities of 41.6 and 40.1 kcal/mol, respectively.

Comparative 2D- and 3D-QSAR studies, followed by molecular docking were conducted on a series of quinazoline derivatives acting as EGFR inhibitors (Noolvi and Patel, 2013). According to the 2D-QSAR multiple linear regression (MLR) model, anticancer activity of quinazoline derivatives was influenced by lipophilicity and number of hydrogen bond donors. Presence of short chain ethers such as methoxy-, ethoxy- at C-6 and C-7 positions of quinazoline was found favorable for the activity, while N-containing groups should not be directly attached to the quinazoline ring. 3D-QSAR kNN-MFA (k-nearest neighbor molecular field analysis) revealed that the presence of electronegative groups on the anilino moiety site, electropositive groups at position C7, and a bulky aromatic substituent at C4 increases the EGFR kinase inhibitory activity.

Virtual screening (VS) refers to a group of *in silico* methods widely used in drug discovery to search large-scale compound databases in order to select a more manageable number of candidates with the highest probability of displaying the desired biological activity (Gagic et al., 2016a; Oluic et al., 2017; Vucicevic et al., 2017; Banegas-Luna et al., 2018). This method has been very popular among pharmaceutical companies since it enables developing drugs in time and cost-effective manner and increases the chance of selected candidates to reach clinical studies. Considering the constant improvement of computational power, it is expected that in the near future VS will be a reasonable alternative to high throughput screening (HTS) (Kumar et al., 2015). There are generally two approaches to screen molecular libraries: LBVS that will be discussed in this section and SBVS.

LBVS is often applied when there are known active compounds, but the target of action is not known, or the crystallographic structure of the protein is not available. These active compounds are then used as ligands to screen molecular libraries based on the similar property principle, which states that structurally similar compounds should possess similar biochemical properties (Nikolic et al., 2015; Bajorath, 2017). For each compound from the virtual library, the similarity with the known active is calculated. Many different strategies for measuring similarity have been developed, including Cosine coefficient, Euclidean distance, Soergel distance, and Tanimoto coefficient (Bajusz et al., 2015). Compounds are ranked based on the similarity score and those at the top are selected as virtual hit molecules for further optimization and synthesis. Modern VS protocols include additional filtering steps in order to exclude compounds that e.g., have low similarity score, do not fall within the Lipinski's rule of five, are not feasible for synthesis or are not available for purchase (Neves et al., 2018).

Besides similarity searches, pharmacophore search is one of the most commonly used LBVS techniques. Given a list of known actives, pharmacophore model can be derived to define the minimum structural requirements that molecule must possess in order to exhibit good activity profile (Vittorio et al., 2019). It is then possible to search large databases, such as PubChem (Kim et al., 2019), ChEMBL (Mendez et al., 2019), and DrugBank (Wishart et al., 2018), for identification of lead compounds that fit to the pharmacophore structure (Bacilieri and Moro, 2006). Several studies that describe the use of LBVS methodology in discovery of potential kinase inhibitors have been listed in **Table 3**. Pharmacophore-based VS model was employed to search for new tumor progression locus-2 (Tpl2) inhibitors (Teli and Rajanikant, 2012). Tpl2 is a serine/threonine kinase in the MAPK signaling pathway that regulates cell proliferation, survival, and death and participates in many processes of tumor development (Lee et al., 2015). For this purpose, Asinex database was screened using PHASE 3.0 module of the Schrodinger molecular modeling software which resulted in six potential Tpl2 kinase inhibitors. A 3D QSAR pharmacophore model was developed from the structures of known inhibitors of MAPK1 (ERK2) and used for virtual screening of ZINC database (Irwin et al., 2012) that contains over 750 million compounds, DrugBank with 13,443 drugs (Wishart et al., 2018), NCI (https://cactus.nci.nih.gov/ncidb2.2/) with 250,250 structures, Maybridge (https://www.maybridge.com) with over 53,000 compounds and Chembank database (Seiler et al., 2008). Top screened compounds were then subjected to molecular docking that identified new scaffolds with high potency and selectivity against ERK2 (Larif et al., 2014).

It can be concluded that VS strategies, especially Pharmacophore-based VS and combined use of VS and molecular docking, can be a reliable tool for future discovery of new KIs and have a potential to replace a HTS that is costly and time consuming process.

TABLE 3 | Selected studies that have used LBVS in the design of kinase inhibitors.

Target kinase	Software package	References
EGFR	PHASE	Sudha et al., 2015
CDK2	SYBYL	Zhang and Ren, 2018
ERK-1/2	QSAR-Co	Halder et al., 2019
VEGFR 2	Discovery studio	Sobhy et al., 2019
ALK	PHASE	James et al., 2019
CDK9/Cyclin T1	LigandScout	Hussain and Verma, 2019
FGFR1	Discovery studio	Liu et al., 2020

Case Studies

Application of quantitative structure-activity relationship in structure elucidation of Lyn kinase inhibitors

The generalized linear model (GLM) and the artificial neural network (ANN) QSAR models were combined with structural analysis in order to define pharmacophore of Lyn kinase inhibitors (Naboulsi et al., 2018). Lyn kinase is a member of the Src family of tyrosine kinases that was found to be correlated with chemotherapeutic resistance of cancer cells in patients with CML (Chakraborty et al., 2013; Aira et al., 2018). Derived pharmacophore for the inhibition of Lyn kinase suggested the presence of planar heterocyclic ring that contains HBD and HBA, a spacer that allows free bond rotation and central hydrophobic area that is linked to the aromatic ring substituted with lipophilic groups. These structural futures can be found in nilotinib and dasatinib that are approved for treatment of CML (**Figure 1**). Pyrimidine moiety of nilotinib has the role of the hydrogen bonding region; the attached amino group serves as a spacer that is linked to hydrophobic benzyl group connected with another aromatic ring that is substituted with lipophilic trifluoromethyl group and methylimidazole. Aminopyrimidine moiety is also present in dasatinib that is indicated in CML patients that developed resistance to nilotinib (Okabe et al., 2011). Dasatinib, instead of central hydrophobic benzene ring, contains thiazole connected to an aromatic ring with lipophilic substituents. Results of these QSAR studies can be of great help in future design and lead to optimization of new, more potent Lyn kinase inhibitors for treatment of patients with imatinib and nilotinib-resistant CML.

Quantum mechanical based quantitative structure-activity relationship of N-phenylquinazolin-4-amine derivatives as epidermal growth factor receptor inhibitors

Recently, Simeon at al. applied several 2D- and 3D-QSAR methodologies on a series of EGFR inhibitors, derivatives of N-phenylquinazolin-4-amine (Simeon et al., 2019). 2D QSAR models were created using physico-chemical descriptors, e-state indices and molecular fingerprints, while 3D-QSAR models were developed using CoMFA, CoMSIA, and quantum mechanical (QM) methods. Based on the calculated statistical parameters, the QM-QSAR model displayed better predictive power compared to the other models. Development of this model started with docking of N-phenylquinazolin-4-amine analogs to the EGFR active site and calculation of pairwise interaction energies between each inhibitor and amino acid residues using quantum mechanics. Distances that hold information about the position of the quinazoline ring and the aniline pharmacophores within the active site of the EGFR were extracted and used as descriptors for the QM-QSAR model. Combined 2D- physico-chemical and QM-QSAR model showed even better predictivity and provided more precise information about structural characteristics that are important for EGFR inhibitory activity. Based on the results of this study, it can be concluded that a combination of classical and more advanced quantum mechanical QSAR methodologies represents a good concept for future design of new EGFR inhibitors.

Discovery of potential FGFR1 inhibitors using pharmacophore-based virtual screening

Pharmacophore-based VS protocol was developed in Maestro 9.0 software package (https://www.schrodinger.com/) and used to screen SPECS database (http://www.specs.net) for potential FGFR1 inhibitors (Zhou et al., 2015). Database was previously filtered to extract only compounds with drug-like properties that comply with the Lipinski's rule of 5. Activities of top ranked compounds were predicted with constructed atom-based 3D-QSAR model, and those with highest activities were purchased for experimental enzyme assay. Nineteen hits exhibited moderate inhibitory activity with more than 50% FGFR1 inhibition at 50 μM concentration and IC_{50} values of most active compounds were 7.9 and 55.5 μM. It should be mentioned that the identified compounds had low structural similarity with previously reported FGFR1 inhibitors and offered novel chemical scaffolds for future optimization of FGFR1 inhibitors.

Structure based methods in drug design

Recent progresses in the field of X-ray crystallography, Nuclear Magnetic Resonance (NMR) techniques, and cryo-electron microscopy (CEM) caused a significant increase in the number of known 3D structures of proteins (Sun et al., 2011). With known 3D structures of proteins, docking became a method of choice in drug design.

Molecular docking predicts the most probable orientation of one molecule toward another (Lengauer and Rarey, 1996). It can be performed between a small molecule and a target protein (ligand-protein docking) or between two proteins (protein-protein docking). In ligand-protein docking, which will be discussed here, the samples of conformations of small molecules–ligands are placed into the binding sites of protein, where scoring functions are used to calculate which of these conformations best fits the target protein binding site (Sousa et al., 2006; Warren et al., 2006). Overall, docking protocols include search algorithm and a scoring function. Initially, the search algorithm is used to orient small molecules in the target binding site (Taylor et al., 2002). Sampling of conformational space has to be carried out with acceptable accuracy to determine the conformation that best fits the binding site, but fast enough to evaluate a large number of docked ligands. With today's computer power it would be impossible to explore all the degrees of freedom for ligand and protein complex. Therefore, there are different ways to overcome this problem. Search algorithms can be systematic and stochastic and deterministic (Novič et al., 2016). Systematic search algorithms sample the search space at predefined intervals while stochastic make random changes until a user-defined termination criterion is met, and because of that outcome can vary (Morris and Lim-Wilby, 2008). Search algorithm is then followed by scoring function that estimates the affinity of ligand through the assessment of interactions between ligands and potential targets (Kitchen et al., 2004). Scoring functions can be physics-based, empirical, knowledge-based, and machine learning-based (Liu and Wang, 2015; Li et al., 2019). The physic-based scoring function computes the free energy of binding by summing up the van der Waals and electrostatic interactions between the protein–ligand (enthalpy), and adding the torsion

entropy of ligand as well as the solvation/desolvation effect described by explicit and implicit solvent models (Huang et al., 2006; Liu and Wang, 2015). Empirical scoring function estimates the binding affinity of a complex by accumulating significant energetic factors for protein–ligand binding (hydrogen bonds, hydrophobic effects, steric clashes, etc.). It uses a training set with known binding affinities of protein–ligand complex and optimizes the weights of the energetic factors by the means of regression analysis (Eldridge et al., 1997; Liu and Wang, 2015). The knowledge-based scoring functions also uses structural information of large set of known protein–ligand complexes and converts it into distance-dependent Helmholtz free interaction energies (Muegge and Martin, 1999; Li et al., 2019). Machine-learning based scoring functions for docking are getting more interests nowadays. These methods combine QSAR analysis and protein–ligand interaction evaluation. They combine QSAR analysis and protein–ligand interaction evaluation. The training set of protein–ligand complexes with known structures and binding affinities is required for a model calculation. Structural interaction fingerprints between a protein and a ligand are coded with certain descriptors (electrostatic interactions, hydrogen bonds, or aromatic stacking, surface or shape properties, molecular weight, number of rotatable single bonds, etc.). Then, different machine-learning algorithms are employed for variable selection (Deng et al., 2004; Zhang et al., 2006).

Molecular docking can be employed in many parts of drug discovery process, such as structure–activity studies, lead optimization, structure based virtual screening, binding modes defining, chemical mechanism studies, etc. (Nikolic et al., 2013; Bautista-Aguilera et al., 2014; Oluic et al., 2017; Albert et al., 2019). Most popular docking programs are DOCK (Kuntz et al., 1982), Autodock (Morris et al., 2009), AD Vina (Trott and Olson, 2010), GOLD (Verdonk et al., 2003), GLIDE from Schrödinger suite (Halgren et al., 2004), and they mostly differ in search algorithms and scoring functions they use. It is always recommendable to explore several different docking programs and then decide on the best one for the specific protein-ligand complexes.

For the last decade, molecular docking has been widely used in design of protein kinase inhibitors (**Table 4**). Tsou et al. designed 4-(phenylaminomethylene) isoquinoline-1, 3(2H, 4H)-dione derivatives, an original class of potent inhibitors that selectively inhibit CDK4 over CDK2 and CDK1 activities. They used SAR and docking to identify interactions between the ligands and residues of the protein's ATP binding pocket and to find interactions with amino acids unique to CDK4 (His82, Val83, and Asp84) and to optimize compounds with improved activity and selectivity toward CDK4 (Tsou et al., 2008). Gopalsamy et al. identified a compound as B-Raf inhibitor from high throughput screening (HTS) and used docking into the crystal structure of B-Raf-Sorafenib complex (1UWH) (Wan et al., 2004) to identify important protein–ligand interactions (two hydrogen bonds with Glu500 and Asp593, and hydrophobic interactions with Ile462, Trp530, Phe582, Ile 512, His 573, and Ile 571) and to optimize the scaffold to obtain compound with improved potency (Gopalsamy et al., 2009). In 2018, Amr et al. synthetized a series of macrocyclic pyrido-pentapeptide candidates, and identified their activity *in vitro* on several kinases. Following docking study of the best compound into VEGFR-2, EGFR, PDGFR, provided information of the binding mode and important protein-ligand interactions which can be further used as a guideline for future design (Amr et al., 2018). In their efforts to design 2-phenazinamine derivatives as Bcr-Abl tyrosine kinase inhibitors, Kale and Sonwane combined molecular docking studies with G-QSAR (Group-Based QSAR). Their *in silico* studies predicted better activity for the thiazolidones and benzenesulfonyl derivatives of phenazinamines than doxorubicin. However, *in vitro* cytotoxic activity was good, though still less than of doxorubicin (Kale and Sonwane, 2018).

TABLE 4 | Selected studies that have used docking in the design of kinase inhibitors.

Target kinase	Software package	References
EGFR	Maestro	Hu et al., 2017
VEGFR-2, CDK-2 and PDGFRβ	MOE	Amr et al., 2018
Bcr-Abl	Autodock	Kale and Sonwane, 2018
PKMYT1	GOLD	Platzer et al., 2018
EGFR, PDGFR-β	GOLD	Fischer et al., 2018
Pim-1	MOE	Mohareb et al., 2019
PAK4	Glide	Gao et al., 2019
Bcr-Abl	Discovery studio	Melge et al., 2019
EGFR	Glide	Debnath et al., 2019
PI3K	AutoDock	Wang et al., 2019b
Pim-1	AutoDock	Hazhazi et al., 2019
VEGFR-2	GOLD	Zhao et al., 2019b
PKC	Glide	Wang et al., 2019c
EGFR	MOE	Khodair et al., 2019

Molecular dynamics (MD) is a simulation technique for studying time dependent evolution of molecular system. Relying on principles of classical mechanics, in MD simulations, positions, and velocities of atoms are computed by classical (Newtonian) laws of motion (Klepeis et al., 2009). The forces acting on these atoms are computed using potential energy functions known as force fields. All common force fields express potential energy through bonded terms (covalent bond-stretching, angle-bending, torsion potential, improper torsions) and non-bonded terms (Lenard Jones repulsion and dispersion and Coulomb electrostatics) (Vanommeslaeghe et al., 2014). Several force fields were found to provide quite accurate representations of the structure and dynamics of a number of small globular proteins on the sub-microsecond timescale (Beauchamp et al., 2012). Most commonly used force fields today are CHARMM (Yin and MacKerell, 1998), AMBER (Weiner et al., 1984; Cornell et al., 1995), GROMOS (Oostenbrink et al., 2004), OPLS (Jorgensen et al., 1996), and COMPASS (Sun, 1998) force fields since they include various chemical groups present in macromolecules and drug-like entities.

Recent algorithmic advances and increase in computational power have enabled simulation studies of protein systems on biophysically-relevant timescales. Combined with modern

improvements in the quality of force field parameters, protein structure prediction and modeling has advanced impressively (Beauchamp et al., 2012; Raval et al., 2012; Piana et al., 2014). Providing structural and dynamical insight into the studied molecular system difficult to obtain experimentally, as well as thermodynamics and kinetic understanding of the system, MD simulations are usually referred to as "computational microscopes" (Dror et al., 2012). In this review, we discuss the usefulness of MD and MD-based methods in the discovery of kinase inhibitors through different case studies presented below.

Structure-based virtual screening (SBVS) is based on the knowledge of the 3D structure of the target protein, obtained by X-ray crystallography, NMR, cryo-EM or homology modeling (Lionta et al., 2014). Nowadays, the SBVS methods are enabled thanks to a large number of 3D structural information deposited in the PDB. As described above, by using the 3D structural information of the protein target, we are now able to investigate the basic molecular interactions involved in ligand-protein binding and understand experimental results up to atomic levels. In SBVS, large libraries of commercially available drug-like compounds that are computationally screened against proteins of known structure and those that are predicted to bind well can be experimentally tested (Benod et al., 2013; Vucicevic et al., 2016; Oluic et al., 2017).

Case Studies

Structure-based design of imidazo [4,5-b]pyridin-2-one-based p38 mitogen-activated protein kinase inhibitors

Using structure-based drug design, Kaieda et al. have identified a series of potent p38 mitogen-activated protein kinase inhibitors. First they identified the lead compound with moderate inhibitory activity toward p38 MAP kinase by means of high-throughput screening. The lead compound was then crystalized with the MAP kinase. The X-ray crystallographic results showed that carbonyl group of the compound forms two hydrogen bonds with the backbone amide of Met109 and Gly110 of the enzyme (**Figure 5A**). The hinge backbone conformation of their crystal structure was different from that typically seen in protein kinases. Namely, usually the backbone amide and carbonyl group of Met109 are directed toward the ATP binding site and accessible for creation of hydrogen bonding with ligand. In the obtained crystal structure a flip of the peptide bond between Met109 and Gly110 was noticed which led to a switching of the hydrogen-bond acceptor and donor distribution around the peptide plane, instead. It was assumed that this flip could be responsible for the high kinase selectivity. After switching the scaffold of the carbonylpiperidine group while maintaining this binding mode, a series of synthetized imidazo[4,5-b]pyridin-2-one derivatives were identified as potent inhibitors of the p38 MAP kinase (**Figure 5B**; Kaieda et al., 2019).

Discovery of novel Pim-1 kinase inhibitors by support vector machine, pharmacophore modeling and molecular docking

In 2011 Ren et al. reported the discovery of novel potent Pim-1 inhibitors by combining ligand- and structure-based filtering methods. In order to find new molecules, a pipeline was created that consisted of support vector machine-based VS (SVM-based VS), pharmacophore-based VS (PB-VS), and docking-based VS (DB-VS) and screened approximately 20 million molecules. Protocol was evaluated by using the library which contained 203 known Pim-1 inhibitors and around 117,000 generated decoys. For validation of the performance of VS, the percentage of predicted compounds in known inhibitors, percentage of known inhibitors in predicted compounds, as well as enrichment factor were calculated. The combined protocol showed much better performance than solely SB-VS, PB-VS, and DB-VS. Finally, 47 compounds were selected for further *in vitro* Pim-1 kinase inhibitory assay for an inhibitor concentration of 10 μM, and 15 compounds showed nanomolar level or low micromolar inhibition potency against Pim-1. In conclusion, new scaffolds with the potential for the future chemical development were found (Ren et al., 2011).

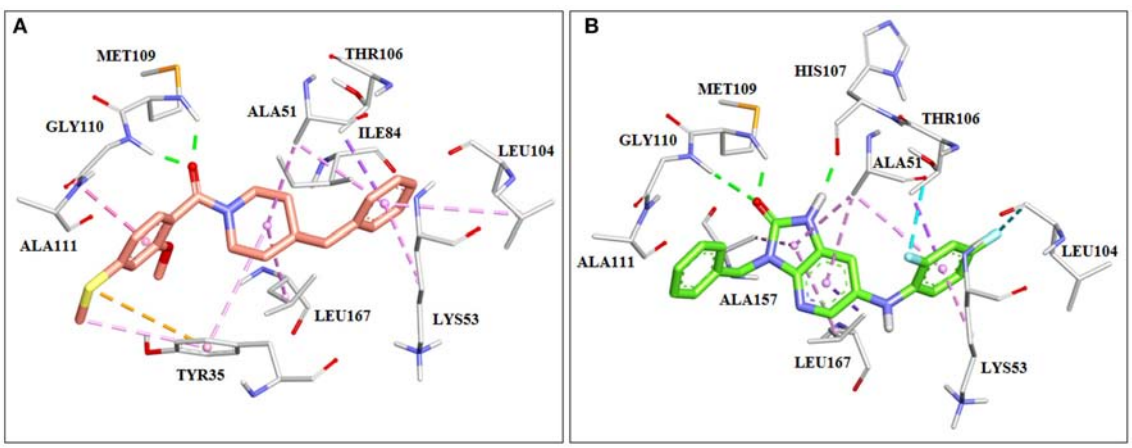


FIGURE 5 | Crystal structures of p38 mitrogen-activated protein (MAP) kinase with imidazo[4,5-b]pyridin-2-one derivatives. **(A)** Lead compound found by HTS (PDB: 6M95). **(B)** Potent p38 MAP kinase inhibitor designed using structure-based drug design (SBDD) approach (PDB: 6M9L).

Discovery of pazopanib, vascular endothelial growth factor family of receptor inhibitor

In 2008 Harris et al. published a paper explaining their discovery of pazopanib. That was a good example of usage of homology modeling and SBDD in the discovery of a drug that is today on the market. Since the crystal structure of VEGFR2 was not available at that time, a homology model of the VEGFR2 enzyme based on FGFR crystal structures was created to predict the binding mode of dimethoxyquinazoline analogs. It was noticed that the pyrimidine and the quinazoline bound similarly in the ATP binding site, making the hydrogen bonds with the Cys919 of the backbone (**Figure 6**). Crystallization of these compounds with VEGFR2 confirmed *in silico* results (PDB: 1Y6A, 1Y6B). Finally a series of new analogs was designed, synthetized, and tested *in vitro*, which led to the discovery of pazopanib (Harris et al., 2005, 2008).

Rational discovery of dual-indication multitarget phosphodiesterase/ kinase inhibitor

One of the latest studies published this year by Lim et al. combined molecular docking with other bioinformatics tools, with the goal of finding multi-target-multi-indication drugs (Lim et al., 2019). They have used structural and chemical genomics data and combined tools from bioinformatics, chemoinformatics, protein-ligand docking, and machine learning to create a novel structural systems pharmacology platform—3D-REMAP. It used four networks as input: 1. protein–ligand association, 2. off-target, 3. ligand–ligand similarity, and 4. protein–protein similarity. The protein–ligand associations were obtained from ChEMBL, DrugBank, and from other publications about kinome assays (Christmann-Franck et al., 2016; Drewry et al., 2017; Klaeger et al., 2017; Merget et al., 2017) and protein structure-based off-target prediction from binding pocket similarity search and protein–ligand docking. Ligand–ligand similarity was calculated in MadFast software from ChemAxon, and protein–protein similarity was run through BLAST. Moreover, to validate and show advantages of their platform, they searched for marketed drugs that could be dual-indication agents. In their study, they focused on drugs that could reduce the cardiotoxicity of anti-cancer therapy. They predicted that levosimendan, a phosphodiesterase (PDE) inhibitor which is used for heart failure, also inhibits serine/threonine-protein kinase RIO kinase 1 (RIOK1) and several other kinases [Ca2+/calmodulin-dependent protein kinase II (CAMK2), FMS-like tyrosine kinase 3 (FLT3), RIOK3, etc.]. To validate their results they tested anti-cancer activity of levosimendan for more than 200 cancer cell lines. Their experimental results showed that levosimendan is active against several cancers, particularly lymphoma, through the inhibition of RIOK1 and its RNA processing pathway (Lim et al., 2019). Since this study is brand new, the time will tell whether levosimendan will be a candidate for clinical research.

Fragment-based drug design of kinase inhibitors

Discovery of kinase inhibitors is a highly competitive process wherein teams of experienced researchers, both from academia and industry, use all the previous knowledge and new ideas to provide more effective therapies for patients. Depending on the available methodologies, one research group may start their drug discovery project with a high-throughput screening (HTS) campaign and search for the bioactive (HIT) compounds against the studied kinase. Selected HIT molecules usually possess drug-like properties and should be further optimized with the aid of lead optimization techniques. Contrary to drug-like molecules, fragments have a smaller number of heavy atoms (HA) and they should comply with Rule of Three (RO3), in which molecular weight is <300 Da, number of hydrogen bond donors and acceptors should be ≤ 3 and clogP is ≤ 3 (Congreve et al., 2003).

Fragment molecules tend to show high micromolar to millimolar affinities for a certain biological target. The advantages of using fragments in drug design studies of novel kinase inhibitors are numerous:

- Fragments displaying affinity to the examined biological target can overcome the entropy barrier and their binding is related to the favorable enthalpy contribution (Murray and Verdonk, 2002);

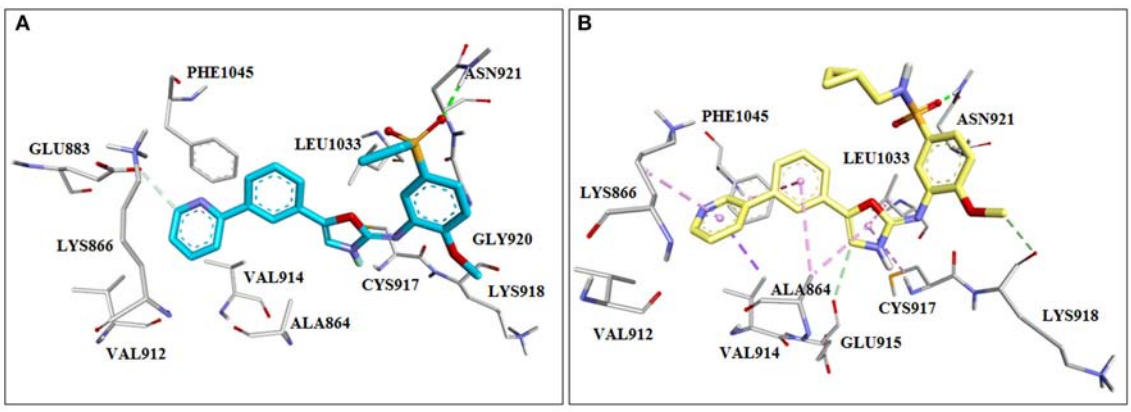


FIGURE 6 | Crystal structures of initial screening hits for inhibitors of the vascular endothelial growth factor (VEGF) that lead to the discovery of pazopanib. **(A)** PDB: 1Y6A, **(B)** PDB: 1Y6B.

- Comparing to drug-sized molecules, pharmacokinetic and physicochemical properties of fragments could be more efficiently optimized (Leach and Hann, 2011);
- Drug-sized molecules may suffer from a potential loss of complementarity with the studied targets, whereas the fragments seldom possess functional groups that establish ligand–protein intermolecular clashes (Hann et al., 2001);
- Given all the above, FBDD projects can lead to increased HIT rates and discovery of novel fragments interesting from different points of view (binding affinity, synthetic accessibility, intellectual property).

Historically, first FBDD projects were applied by a technique named "SAR by NMR" (structure-activity relationship by nuclear magnetic resonance) (Shuker et al., 1996). In this paper, authors successfully developed a potent compound with nanomolar affinity to FK506 binding protein (FKBP) by merging two building blocks. Except for NMR, protein-fragment interacting patterns are characterized by other biophysical methods such as X-ray crystallography, surface plasmon resonance (SRC), high concentration screening (HCS) assays, isothermal titration calorimetry (ITC), fluorescence correlation spectroscopy and many more (Sun et al., 2011). The choice of a particular method depends on the previous experience in FBDD projects and also the size of fragment libraries.

Until now, fragment-based drug discovery (FBDD) method resulted in FDA approval of three kinase inhibitors—vemurafenib (Bollag et al., 2012), venetoclax (Deeks, 2016), and erdafitinib (Markham, 2019). These excellent textbook examples of FBDD are developed by different biophysical methods; nevertheless, the present review focuses on various *in silico* techniques frequently used in fragment identification and optimization.

In recent years, experimental screening procedures may be replaced by computational methods to reduce the costs and time for early stages of FBDD project (Alves Avelar et al., 2019; Ruzic et al., 2019). It appears that *in silico* studies may support kinase drug discovery at almost every stage of fragment-based drug design projects. Various ligand-based virtual screening (Giordanetto et al., 2011), structure-based (Warner et al., 2006; Zhao et al., 2012), and quantum mechanical (Machrouhi et al., 2010) techniques have been proved as successful in novel fragment identification. Before running any virtual screening protocol, computational chemists must pay attention to the valid preparation of fragment library database. The fragment library databases should obey the aforementioned Rule of 3 (RO3); additionally, their chemical properties are filtered through certain software which removes possible toxicophores and pan-assay interference compounds (PAINS) (Baell and Walters, 2014). Nowadays, computational chemists may use kinase fragment libraries which may assist faster identification of novel hinge binding motifs. Moreover, fragments that target distal pockets from the ATP binding pocket could be scanned by allosteric kinase library, such as Enamine Allosteric Kinase Library (https://enamine.net).

Case Studies

Identification of PI3K p110β selective fragment

Intracellular lipid kinases that transfer a phosphate group from ATP to certain cell membrane's phospholipids (Phosphoinositide-4,5-biphosphate, PIP_2) belong to the family of phosphoinositide 3-kinases (PI3K). These enzymes regulate important cellular events and present interesting drug targets in anticancer drug discovery. Giordanetto et al. (2011) successfully identified fragments that showed selective p110β inhibition. At the time this study was performed, the crystal structure of p110β isoform was not available. Consequently, the homology model was built in MODELLER (Webb and Sali, 2016) by using the crystal structure of p110γ isoform. In this study, authors used AstraZeneca's virtual fragment database and subjected 183,330 fragments to a molecular docking study in GLIDE software (Schrödinger, New York). The poses and orientation of the fragments in the ATP binding pocket were inspected, as well as hydrogen bonding interactions with amino acid residues in the hinge region, affinity and selectivity pocket. The authors reported five chemical classes of fragments (**Figure 7A**) based on the different heterocyclic rings interacting with the hinge region in p110β and their *in vitro* enzymatic profiles against four human PI3K isoforms (p110α, p110β, p110γ, and p110δ). Overall, the hit rate achieved from this screening was 8.57%, indicating good performance of the molecular docking-based search for novel and chemically interesting fragments as PI3K hinge binders. The authors continued this study with the morpholine derivative, compound (**1**) (**Figure 7B**), which showed moderate potency against p110β (IC_{50} = 34 μM), but its inhibition of the other p110 isoforms was not determined at the tested concentrations.

In the following study, authors aimed to improve the affinity of the compound (**1**) by substituting the dimethylamino group with a more voluminous 2-(benzylamino) moiety (Giordanetto et al., 2012). The novel compound (**2**) showed improved potency (IC_{50} = 1.9 μM) and efficiency (LE = 0.37 and LLE = 4.52) toward p110β. The rationale for this chemical modification relies on the observation that the bulkier substituents might target amino acid residues M804 and W812 in the proximal selectivity pocket. Finally, compound (**3**) was synthesized by introducing the naphthyl group, which in turn attributed to the nanomolar potency (IC_{50} = 0.093 μM) and improved p110β selectivity profile of compound (**3**).

Identification of mitogen-activated protein kinase-interacting kinase 1 inhibitors

Mitogen-activated protein (MAP) kinase interacting kinases 1 and 2 (**MNK1** and **MNK2**) carry out phosphorylation reaction of eukaryotic translation initiation factor 4E (eIF4E) on serine 209 (Wendel et al., 2007). This translation factor is involved in different cellular pathways, such as Ras/Raf/MEK/ERK and PI3-kinase/protein kinase B (Akt) signaling pathways (Proud, 2015). The overexpression of phosphorylated eIF4E leads to several malignant diseases, such as lymphomas, breast cancer, and glioblastoma (Astanehe et al., 2012). The significance of MNK1/2 enzymes in malignant transformation of the cell has led to high demand for drug design of MNK1/2 inhibitors.

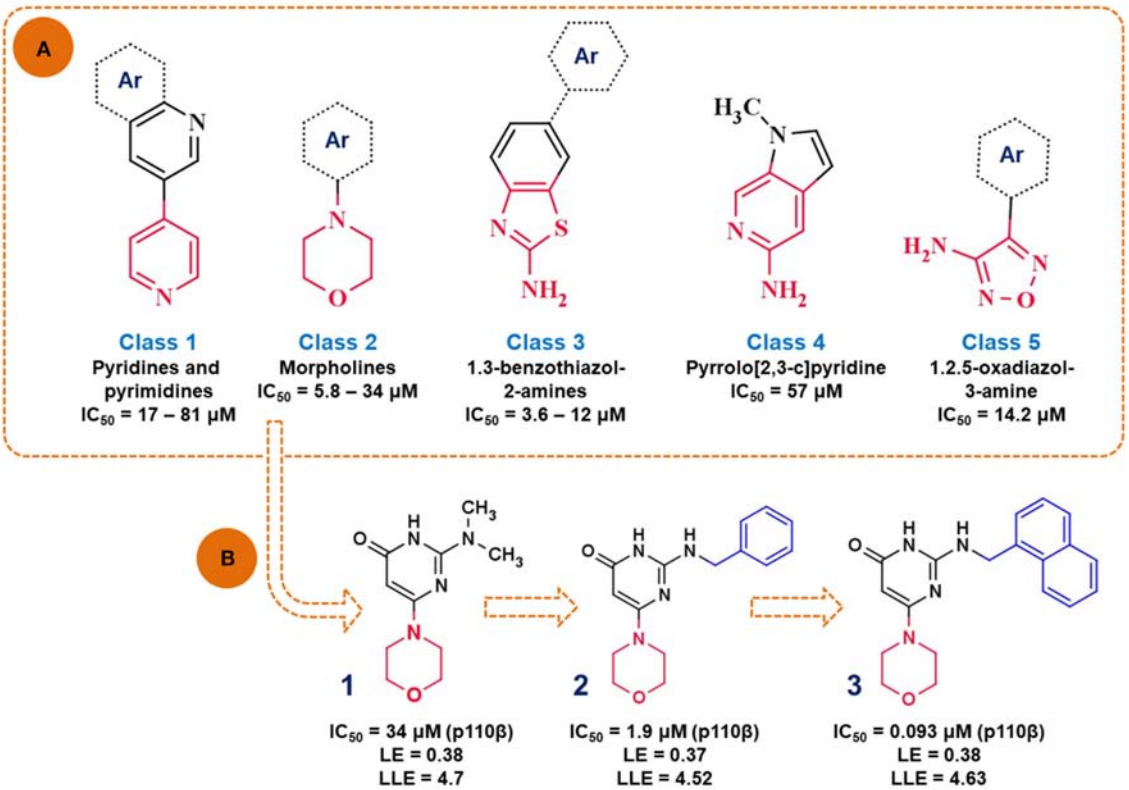


FIGURE 7 | (A) Presentation of the identified scaffolds by *in vitro* biochemical screening as PI3K inhibitors (hinge interacting moieties are labeled in red), *Ar*—presents carbocyclic or heterocyclic aromatic rings; **(B)** the scheme of lead optimization of the selected morpholine derivatives.

One remarkable study was performed in 2010, where Oyarzabal et al. identified a highly potent and efficient fragment entirely by *in silico* modeling. In this comprehensive study, authors combined different virtual screening techniques to identify pharmacological tools for MNK1 inhibition. Initially, the Centro Nacional de Investigaciones Oncológicas (CNIO) database was filtered according to the molecular weight (<300 Da) and calculated solubility values (threshold −4 mol/L). By performing this prefiltering procedure, the authors extracted 42,168 fragment-like compounds for virtual screenings (Oyarzabal et al., 2010).

Availability of the crystal structure of MNK2 complexed with staurosporine (PDB: 2HW7) enabled creating minimal substructure, required for crucial interactions with MNK2 (**Figure 8**). The GOLD software (Jones et al., 1997) used in this study was able to reproduce the binding mode of staurosporine in MNK2. The virtual substructure was docked in the crystal structure of MNK1 (PDB: 2HW6) to similarly elucidate crucial amino acid interactions in the ATP binding pocket. MNK1 pharmacophore prepared in this way was used for pharmacophore fitting study, as a molecular docking alternative and 92 compounds were extracted according to their goodness of fit with the pre-defined substructure.

Structure-based virtual screening protocols were combined with ligand-based virtual screenings of CNIO database and external virtual database of compounds collected by the authors who performed this study. These strategies involved 2D-substructural searches, 2D Tanimoto structural similarity, Feature Trees similarity, and three-dimensional shape and electrostatic similarities based on two reported MNK1 inhibitors. Finally, the authors selected 1,236 compounds for biochemical MNK1 assay and 26 of them were active. The hit ratio of this screening was 2.10% and 10 different scaffolds were represented. Interestingly, one compound (**Figure 8**, compound 29) demonstrated nanomolar MNK1 (IC_{50} = 646 nM) and MNK2 (IC_{50} = 575 nM) inhibition. Additionally, at the cellular level, compound 29 showed an antiproliferative effect against acute myeloid leukemia cell line (MV4:11, EC_{50} = 17 μM) with dose-dependent decrease in phosphorylation on serine 209 in eIF4E. In conclusion, this study identified 26 hit molecules as MNK1 inhibitors, with 19 of them as fragments with high ligand efficiency values. Among the 26 identified hits, there were 10 diverse chemotypes represented for further drug design studies.

Researchers from A-STAR were particularly interested in imidazopyridazine scaffold (chemotype III in the study of Oyarzabal et al.) as a starting fragment for lead optimization of MNK1/2 inhibitors (Yang et al., 2018). Extensive SAR study of imidazopyridazine derivatives was based on *in silico* conclusions defined in their previous computational study (Kannan et al., 2017). Concisely, researchers in this study aimed to modify the

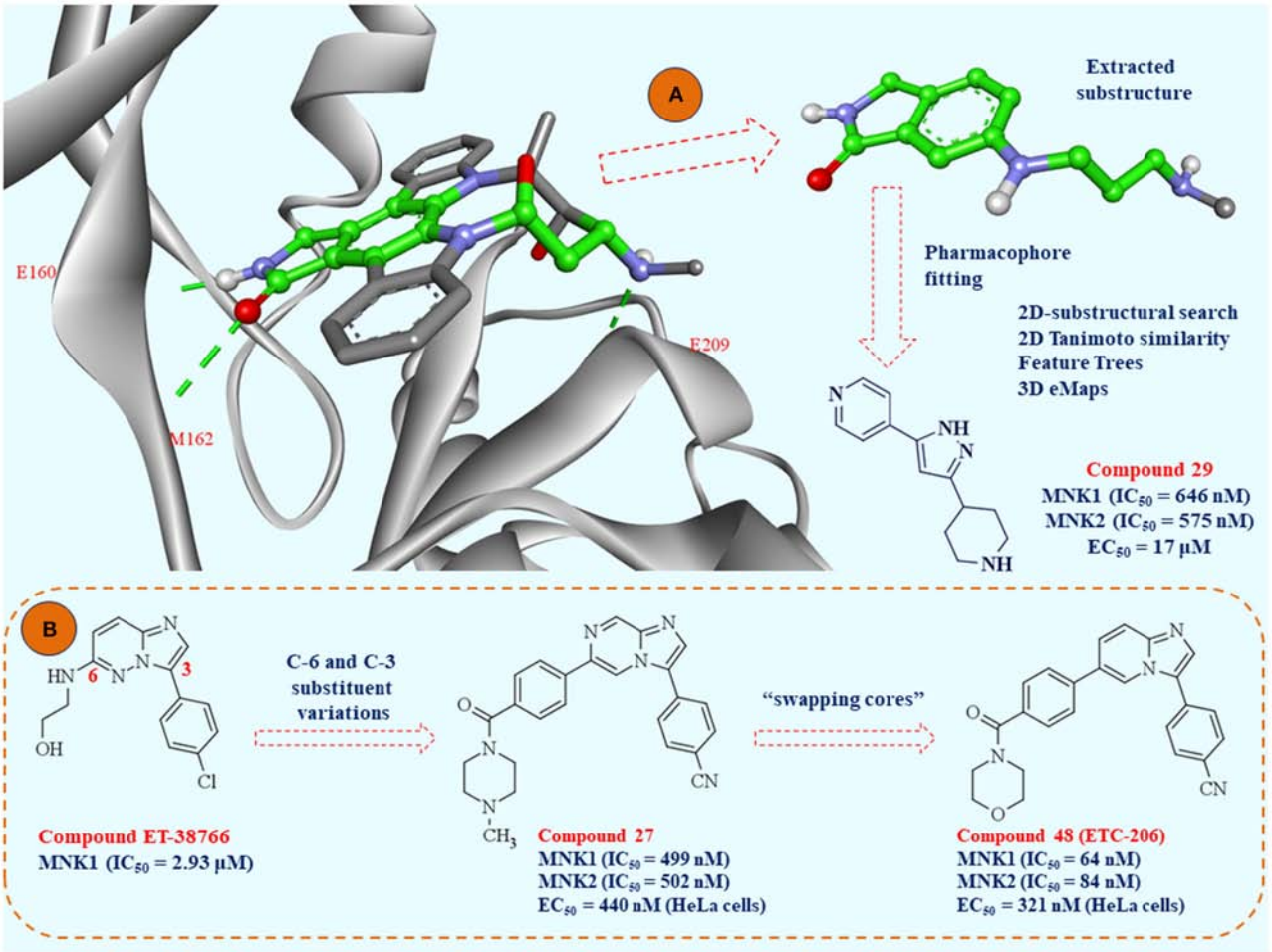


FIGURE 8 | (A) Crystal structure of MNK-2 complexed with staurosporine (PDB: 2HW7); the atoms of the substructure used in the study of pharmacophore fitting screening are labeled in green; **(B)** the lead optimization strategies starting from fragment ET-38766 to clinical candidate ETC-206.

heterocyclic core in positions 3 and 6, with later modification of the imidazopyridazine scaffold (swapping cores strategy). All the derivatives synthesized in this study were firstly examined by molecular docking studies in Glide 2017-3 software (www.schrodinger.com). By detailed computational analysis of the important amino acid residues in the ATP pocket of MNK1/2 kinases, researchers performed initial lead optimization of the fragment (compound ET-38766) to compound 27 (**Figure 8B**). Novel compound 27 bears imidazopyrazine scaffold, with improved potency against MNK-1 and MNK-2, cell permeability and improved pharmacokinetic properties. After finding optimal substituents in positions 3 and 6, the final step of lead optimization was focused on detailed DFT study to select the final heterocyclic core of MNK1/2 inhibitors. Initially, it was unclear from molecular dynamics (MD) simulations whether the imidazopyrazine N-7 contributes favorably to the binding affinity of MNK inhibitors. To examine this, the authors performed DFT study and demonstrated that N-7 is mostly solvent exposed, thus the final selected heterocyclic core was imidazopyridine. The most promising compound 48 (**Figure 8B**) later designated as ETC-206, was presented as superior compared to other derivatives in the study. This compound was investigated for the synergism with dasatinib *in vivo* and currently is in phase I clinical trial for the blast crisis chronic myeloid leukemia (BC-CML).

Computational approaches in rational discovery of allosteric kinase inhibitors

Although targeting of highly conserved ATP-binding site by Type I and Type II inhibitors provides limited selectivity, inhibiting multiple kinases with a single small-molecule inhibitor was proven to be a useful strategy for therapeutic intervention. However, development of highly selective small-molecule kinase inhibitors remains a pressing concern where targeting of allosteric sites emerged as a promising approach (Wu et al., 2015a). Some of the advantages of targeting allosteric sites include increased selectivity and low toxicity of such inhibitors due to low evolutional conservation of allosteric sites compared to orthosteric (ATP-binding) sites (Fang et al., 2013). Additionally, overcoming of point mutation-associated drug resistance,

especially for mutations in the ATP-binding site reported for almost all of ATP-competitive inhibitors, could be another advantage of developing allosteric kinase inhibitors (Gibbons et al., 2012).

While exploitation of allosteric sites represents a very promising strategy, it remains challenging from the aspect of rational drug discovery. Some of the major obstacles include identification of allosteric binding sites, which are usually hidden in less populated higher energy conformations of the proteins. Those conformations are poorly accessible to current experimental methods of structural biology (Lu et al., 2018a). Additionally, allosteric effectors are susceptible to "mode switching," where minor chemical modification of ligand induces critical change in activity (Wenthur et al., 2014). Although known CADD workflows for discovery of drugs directed to orthosteric binding sites are being used in allosteric inhibitors discovery (Rastelli et al., 2014; Schoepfer et al., 2018), they provide limited utility rising the need for development of more spatialized tools and workflows (Greener and Sternberg, 2018).

Identification of allosteric pockets is a crucial first step in rational discovery of allosteric inhibitors. As will discussed below, a plethora of computationally inexpensive methodologies have been developed for this purpose and many of them are even implemented as web servers. While these methodologies provide fast and inexpensive highway in the discovery of druggable allosteric pockets, proper understanding of the allosteric mechanism is impossible without considering underlying conformational landscape and free-energy profiles where more computationally demanding molecular dynamics based approaches have a predominant role. In this review, we discuss few examples of computational methodologies used for direct discovery of novel allosteric sites and/or allosteric kinase inhibitors. For detailed description of recent breakthroughs in computational methodologies used for allosteric inhibitors discovery in general, the interested reader is referred to the recent reviews (Wagner et al., 2016; Lu et al., 2019).

Automatic Computational Tools/Web Servers to Investigate Allostery

Structure-based computational tools AlloSite and recently advanced descendant AllositePro (http://mdl.shsmu.edu.cn/AST/) are intended for fast detection of allosteric site in input PDB structures. Initial detection of allosteric sites is based on Fpocket, a fast open source protein pocket detection software package based on Voronoi tessellation (Le Guilloux et al., 2009). While Allosite uses a machine-learning model to re-rank detected pockets in terms of their allosteric character, AllositePro additionally implements normal-mode analysis (NMA) perturbation with elastic network models to account for protein flexibility. NMA is a technique developed for investigation of the vibrational motion of a harmonic oscillating system in the immediate vicinity of its equilibrium. Under assumption that the potential energy landscape in the vicinity of a minimized atomic structure is approximately harmonic, NMA eliminates the need to integrate the equations of motion and makes NMA much less computationally demanding compared to MD (Bahar and Rader, 2005). Zhang et al. demonstrated utility of AllositePro in identification of novel allosteric site on CDK2 kinase. Existence of novel site was validated in mutagenic analysis (Song et al., 2017). Recently, the same group developed AlloFinder, integrated allosterome mapping, and virtual screening workflow implemented as web server (http://mdl.shsmu.edu.cn/ALF/). AlloFinder relies on AllositePro algorithm for detection of allosteric sites, Allolike filter for pre-filtering of ligand library to enrich allosteric-like compounds (Wang et al., 2012), AutoDock Vina algorithm for docking (Trott and Olson, 2010), and Alloscore empirical scoring function for scoring allosteric modulator-protein complexes (Li et al., 2016). In the final step, alosterome mapping is used to detect highly similar allosteric sites among known human allosteric sites and to rule out selective ligands. This approach was retrospectively validated on several kinase targets (Huang et al., 2018).

CavityPlus (http://www.pkumdl.cn:8000/cavityplus/index.php) is another web server for detection of potential allosteric sites that works on similar principle (Xu et al., 2018). CavityPlus is aimed to detect potential binding sites on the surface of a given protein and rank them based on ligandability and druggability scores. This server integrates several functionalities: CAVITY for detection and scoring of potential binding sites (Yuan et al., 2013); CavPharmer for generation of receptor-based pharmacophores (Chen et al., 2014); CorrSite for prediction of allostery based on NMA motion correlation analysis between allosteric and orthosteric sites (Ma et al., 2016); CivCys for detection of binding sites for covalent inhibitors (Zhang et al., 2017). Functionalities of CavityPlus were successfully used for identification of allosteric binding site on Polo-like kinase 1 (Plk1). Subsequent molecular-docking-based virtual screening on allosteric site resulted in identification of few potent Plk1 inhibitors (Yun et al., 2016).

Another successful implementation of web server based tools for allosteric drug discovery is Kinase Atlas (https://kinase-atlas.bu.edu/) (Yueh et al., 2019). Kinase Atlas is systematic collection of mostly unexplored allosteric sites (binding hot spots) calculated for 4,910 PDB structures of 376 distinct kinases. The hot spots are identified by FTMap. This method places molecular probes (small organic molecules) on a dense grid around the protein and finds favorable positions using an empirical energy function and CHARMM potential. After clustering of obtained positions for each probe, regions that bind several probe clusters are marked as hot spots (Kozakov et al., 2015). Authors of the study identified novel allosteric site on CDK2 and screened library of 1,280 molecules using disulphide-based fragment screening. Two potent and novel allosteric inhibitors were described.

Molecular Dynamics-Based Approaches to Investigate Allostery

Molecular dynamics-based approaches in rational discovery of allosteric kinase inhibitors have potential to provide exclusive insight in atomic-level dynamical mechanism of allostery, to explore conformational landscape and capture kinase conformational states inaccessible to current experimental methodologies. Therefore, molecular-dynamics-based approaches, even though being computationally intensive,

could detect previously unknown conformations and hidden allosteric binding pockets (Guo and Zhou, 2016; Lu et al., 2018a).

Combination of conventional MD simulations with other standard SBDD approaches resulted in identification of novel allosteric sites and discovery of novel allosteric ligands in several cases. For example, Perez et al. identified novel inhibitory allosteric site and inhibitors of p38α by using MD simulations starting from the X-ray structure of binary complex of p38α and its interacting partner MAPK-activated protein kinase 2 (MK2). MD simulations permitted definition of pharmacophoric features of small peptide inhibitors derived from sequence of MK2. Subsequent virtual screening study resulted in first small molecule allosteric inhibitor for identified binding site (Gomez-Gutierrez et al., 2016). Cournia et al. verified existence of allosteric site on human PI3Kα previously described in murine PI3Kα using combination of FTMap, MD, and *in vitro* assays. Intriguingly, MD simulations revealed different binding mode of studied allosteric inhibitor in murine, WT, and mutant forms of PI3Kα and consequent differences in propagation of allosteric signal to orthosteric ATP-binding site (Gkeka et al., 2015).

Computational costs of insufficient conformational sampling often limit application of conventional MD simulations in investigating allostery phenomena. Currently, there is a large gap between the time scale which can be reached in MD simulations and that observed in experiments. Several strategies for enhancing the sampling of MD simulations have been proposed (Aci-Seche et al., 2016; Yang et al., 2019). Two recently reported studies demonstrating full power of enhanced sampling methods (Markov-state modeling based adaptive sampling and parallel tempering in the well-tempered ensemble) are discussed below with special reference to atomic-level description of allosteric communication and discovery of cryptic allosteric pockets.

Pande et al. investigated activation pathway of c-Src kinase using massively distributed MD simulations (550 μs) on Folding@HOME (Shirts and Pande, 2000) Markov-state modeling (MSM) and adaptive sampling algorithms in order to provide description of factors underlying thermodynamics and kinetics of c-Src activation and to identify key structural intermediates (Shukla et al., 2014). Briefly, MSM models represent kinetical description of a system's underlying free-energy landscape, useful for characterization of probability of dynamical transitions between conformational states identified in many independent MD simulations and for extrapolation of long time system's behavior (Sengupta and Strodel, 2018). In this study intermediate conformational state which could be stabilized to block the c-Src activation pathway, was described through MSM analysis for the first time. Further analysis on identified c-Src conformational state revealed the existence of allosteric pocket and surprisingly high structural similarity to known complex of CDK2 bound to allosteric inhibitor—ANS (Betzi et al., 2011). Further simulations confirmed binding of ANS to the novel allosteric site of c-Src and blockage of activation process by stabilization of intermediate states. Additionally, the long-range residues coupling analysis identified myristate-binding pocket as another potential target for development of allosteric modulators of c-Src. Taken together, results of this study highlighted large-scale MD coupled with MSM modeling as an indispensable tool for identification of novel conformational states, potential allosteric pockets, and study of mechanisms of allostery in kinases.

In another example, authors explored the possibility of bidirectional communication between allosteric so-called PIF-pocket and ATP-binding site in PDK1 protein kinase using a combination of experimental techniques and enhanced-sampling simulations [parallel tempering simulations in the well-tempered ensemble (PT-WTE)] (Schulze et al., 2016). Results of PT-WTE MD revealed bidirectional mechanisms of communication between the ATP-binding site and allosteric site. Interestingly, this study for the first time demonstrated how different ligands which bind to the ATP-binding site differently modulate responses of allosteric site in interaction with a partner protein (e.g., enhance or inhibit interaction). Providing computer platform for rational design of allosteric modulators, the authors of this study opened an exciting avenue for future discovery of novel class of kinase inhibitors with less on-target side effects and more specific modulation of signaling pathways.

Case Study

Rational design of clinical candidate Asciminib—allosteric Bcr-Abl1 inhibitor

Asciminib belongs to a class of drugs designed to inhibit Bcr-Abl by binding to an allosteric pocket known as myristate-binding pocket. Rational development of Asciminib started with fragment-based screening using NMR assay (Schoepfer et al., 2018). Although determined NMR-based dissociation constants (Kd) for fragment hits were satisfactory, none of the fragments were active in biochemical and cellular assays. Subsequent X-ray studies revealed inability of fragment hits to induce assembled inactive state by bending of helix I, previously reported as conformational change important or autoinhibition of Abl by myristoilation (Nagar et al., 2003). Following this finding, the authors established another screening assay, the NMR-based conformational assay, which monitors the conformational state of C-terminal helix I (Jahnke et al., 2010). NMR-based conformational assay was used to investigate identified fragments and series of known allosteric modulators—derivatives of GNF-2 (Adrian et al., 2006; **Figure 9**). Results of the study revealed that compounds which bind to myristoyl pocket and do not induce helix I bending were actually functional activators of Abl1 (by interfering with autoinhibition mechanism of Abl1). Critical bending of helix I was found to be induced by the presence of CF_3O– group from GNF-2. Based on these findings, CADD techniques (molecular docking, similarity and pharmacophore searches) were used to design compound **X** in respect to X-ray structure with bent helix I conformation. Subsequent introduction of CF_3O– group finally led to the first active allosteric inhibitor. Molecular modeling techniques were used in combination with X-ray crystallography in order to optimize potency and drug-like properties of the compound. Although only standard CADD techniques were reported in the discovery of Asciminib, recent application of molecular dynamics-based approaches demonstrated utility of such techniques in examination of mechanisms of resistance

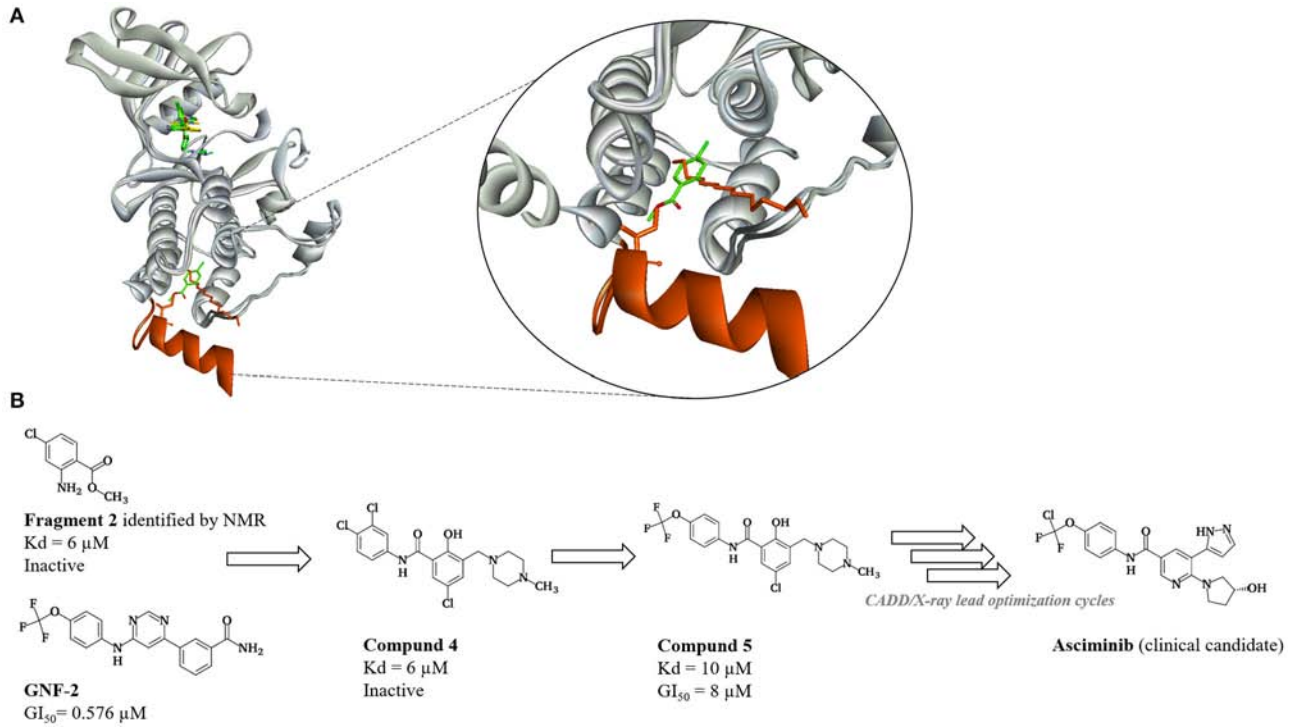


FIGURE 9 | Discovery of Asciminib (Schoepfer et al., 2018). **(A)** Representation of initial hypothesis regarding bending of helix I (orange). Assembled inactive state of ABL1 kinase in complex with myristic acid (orange sticks) (PDB: 1OPL, SH3, and SH2 domains omitted for clarity) is superimposed on ABL1 in complex with fragment 2 (green sticks) (PDB: 3MS9). Steric clash between isoleucin I521 (orange sticks) on helix I and fragment 2 prevents full bending of helix I and formation of assembled inactive state of ABL1. Helix I is not visible in the PDB: 3MS9. **(B)** Medicinal chemistry progression from fragment 2 to fragment derived hit 4, first active hit–compound 5 and finally clinical candidate—Asciminib.

and effects of dual targeting of ATP-binding and allosteric site providing rationale for development of novel drugs (El Rashedy et al., 2018; Meng et al., 2018; Zhan et al., 2019).

Machine learning methods to predict kinase-compound interactions

Nowadays, we are seeing the widespread use of machine learning in many areas, including pharmaceutical industry, especially in drug design. Popular computational methods initially used in pharmaceutical research were quantitative structure activity relationships (QSAR) and quantitative structure property relationship (QSPR), which were adequate for small datasets. However, with the rapid growth of databases (thanks to methods such as high-throughput *in vitro* screening and X-ray crystallography), it became inevitable to develop different *in silico* tools that can manage bigger data (Ekins, 2016). Today, many different machine learning methods such as support vector machines (SVM), k-Nearest Neighbors, Artificial Neural Networks (ANN), Deep Learning (DL), *etc*. are used in pharmaceutical research and they can be applied in various processes of drug design from virtual screening to *de novo* drug design (Buchwald et al., 2011; Drewry et al., 2017; Konze et al., 2019; Kuthuru et al., 2019; Lee et al., 2019; Zhavoronkov et al., 2019).

Many different machine learning models were created for the prediction of drug–target interactions (DTI), and many DTI methods have been applied to the protein kinases family (Kuthuru et al., 2019). Unlike LB and SB methods, DTI prediction uses the information from both protein and ligand and these methods can be similarity based or descriptor-based. One of the first similarity-based methods for identification of drug–target interactions was introduced by Yamanishi et al. in 2008. It used the known drug structure, protein sequence and drug–target interaction network to determine unknown ligand–target interactions. The main hypothesis is that two compounds that have high structure similarity might probably interact with similar target proteins, and *vice versa* two proteins with high sequence similarity might probably interact with similar drugs (Yamanishi et al., 2008). On the other hand, descriptor-based models use feature vectors from known drug structures and protein sequences, as inputs for machine learning methods, such SVM, AAN, DL, etc. In 2011, Buchwald et al. used SVM to prepare the model for prediction of protein kinases–ligand interactions. They used a set of binding data obtained from 113 different protein kinases and 20 inhibitors obtained through ATP site-dependent binding competition assays. They focused on vector features that describe the structure of molecules that are connected with certain chemical environment–protein active site sequence and

created a SVM model with good predictivity (Buchwald et al., 2011).

Recently, the use of ANN, especially deep learning methods saw a significant increase in the process of drug design (Ekins, 2016; Merk et al., 2018; Putin et al., 2018; Konze et al., 2019). Deep generative models are utilizing neural networks to generate new objects (drugs) with desired properties (for example activity, Ki, IC_{50}). These methods should be able to produce chemically correct structures without the need for including fragment libraries and/or rules for their combination (Merk et al., 2018). The ability to produce novel chemical structures with certain properties makes deep generative models suitable for the discovery of novel possible therapeutics (Zhavoronkov et al., 2019). In 2018, Merk et al. applied generative models to come up with novel bioactive, synthesizable drugs. They trained the model with more than 500,000 SMILES of bioactive compounds with their activity properties extracted from the ChEMBL (K_D, K_i, IC/EC_{50} values $<1\,\mu M$). Additionally, the model was fine-tuned to enable the *de novo* generation of target-specific ligands on retinoid X receptors (RXR) and/or peroxisome proliferator-activated receptors (PPAR). Finally, none of the generated compounds was identical to compounds from the training sets, and they were residing within the RXR/PPAR region of the fine-tuning set (Merk et al., 2018).

Case Studies

Predictive proteochemometric models for kinases derived from 3D protein field-based descriptors

Subramanian et al. described the development of proteochemometric models for 1,572 inhibitors and 95 kinases obtained from Kinase SARfari (https://chembl.gitbook.io/chembl-interface-documentation/legacy-resources#kinase-sarfari) and CHEMBL database, using 3D structure of proteins and active and inactive ligands. Proteins were described with molecular interaction fields derived from Schrödinger's WaterMaps, while different 1D, 2D, and 3D descriptors were used to describe the ligands. Separate training sets were created for the ligands and targets. Different methods were used for preparation of the proteochemometric models: support vector machines (SVM) and random forests (RF). The ligand prediction model was trained on the ligand training set and was used for ligand prediction model and target training set for target predicting model. In the end, they validated all the models using internal and external validation. This approach allows creation of not only predictive proteochemometrics model for protein kinases, but also preparation of visually interpretable models. This allows interpretation of kinase–ligand interactions, which can be used, for example, for optimization of ligand in order to achieve optimal activity and/or selectivity. Having visually interpretable models is the advantage compared to classical DTI methods that use only 2D information (Subramanian et al., 2013, 2016).

Deep learning model for identification of potent discoidin domain receptor 1 kinase inhibitors

Recently, Zhavoronkov et al. created a deep generative model for *de novo* small-molecule design—GENTRL (GENerative Tensorial Reinforcement Learning). Besides the effectiveness of a compound against a given biological target, GENTRL also takes into account its dissimilarity from other molecules in the literature and patent space, as well as its synthetic feasibility. For the proof-of-concept GENTRL was used to design potential Discoidin domain receptor 1 (DDR1) kinase inhibitors. Data was collected from different data sets: ZINC data set, known DDR1 kinase inhibitors data set, common kinase inhibitors, molecules with activity on non-kinase targets, patent data, and used to train the model. The model was generated by combining reinforcement learning with a reward, variational inference, and tensor decompositions. Finally the randomly elected six compounds that have not been previously published or patented were designed, synthesized, and experimentally tested. The whole process lasted only 46 days, which suggests that the application of drug design methods such as this will reduce the time and cost of drug discovery process (Zhavoronkov et al., 2019).

CONCLUDING REMARKS

In silico approaches are viable, usually cheaper and faster alternative to experimental drug discovery techniques. This review summarizes the most important computational tools that have led to the discovery of kinase inhibitors, many of which are in clinical use today as promising anticancer drugs. Computational approaches, such as QSAR modeling, ligand-based and structure-based virtual screening, molecular docking, molecular dynamics, quantum mechanics, fragment-based drug design, and machine learning methods, provide unique insight in the conformational landscape of kinases, structural requirements for inhibitory activity, binding modes and atomistic mechanisms of allostery, which represent indispensable information for rational *de novo* design. One of the main advantages of computational approaches is the possibility of introduction of new groups on the known scaffolds and *in silico* prediction of activities and binding affinities. Known scaffolds of the approved KIs include pyrimidine (imatinib, dasatinib, nilotinib), quinazoline (erlotinib, gefitinib, afatinib, vandetanib), pyridine (sorafenib), pirrolopyridine (vemurafenib), pyrazolopyridine (ibrutinib) etc. *In silico* modification of these scaffolds resulted in the design of many kinase inhibitors with enhanced predicted activities and binding affinities which can serve as lead compounds for further synthesis and preclinical testing. New chemical scaffolds that possess kinase inhibitory activity (imidazopyridazine, imidazopyridine, isoquinoline, phenazinamine, etc.) have also been proposed by computational approach and represent a good starting point for discovery of new kinase inhibitors. Due to increases in computational power, algorithmic improvements and increased accuracy, *in silico* approaches are yet expected to radically shape the era of kinase inhibitor discovery. Of note is to emphasize that not all drug discovery projects could be initiated and guided only with computational studies. The computational chemist must be aware of the structural biology of the studied targets, their dynamical changes influenced upon fragment/ligand binding. Whenever possible, it is advised to start CADD studies with

experimental data and continue *in silico* optimization with combined modeling approaches, as much as possible. This review highlights the recent advances in discovery of kinase inhibitors by *in silico* approaches and can be useful for future design and synthesis of new kinase inhibitors as anticancer drugs.

AUTHOR CONTRIBUTIONS

ZG wrote the introduction and LB methods. TD wrote the structure of protein kinases, machine learning methods, and SB methods. DR wrote FB methods. ND wrote modeling of allosteric kinase inhibitors. KN contributed to the conception of the manuscript, collection of the data, drafting and revising. All authors have contributed to the bibliographical research and interpretation of the work, to its critical revision, and approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors kindly acknowledge national project number 172033 supported by the Ministry of Education, Science and Technological development of the Republic of Serbia.

REFERENCES

Aci-Seche, S., Ziada, S., Braka, A., Arora, R., and Bonnet, P. (2016). Advanced molecular dynamics simulation methods for kinase drug discovery. *Future Med. Chem.* 8, 545–566. doi: 10.4155/fmc.16.9

Adams, J. A. (2001). Kinetic and catalytic mechanisms of protein kinases. *Chem. Rev.* 101, 2271–2290. doi: 10.1021/cr000230w

Adrian, F. J., Ding, Q., Sim, T., Velentza, A., Sloan, C., Liu, Y., et al. (2006). Allosteric inhibitors of Bcr-abl-dependent cell proliferation. *Nat. Chem. Biol.* 2, 95–102. doi: 10.1038/nchembio760

Agafonov, R. V., Wilson, C., and Kern, D. (2015). Evolution and intelligent design in drug development. *Front. Mol. Biosci.* 2:27. doi: 10.3389/fmolb.2015.00027

Aira, L. E., Villa, E., Colosetti, P., Gamas, P., Signetti, L., Obba, S., et al. (2018). The oncogenic tyrosine kinase Lyn impairs the pro-apoptotic function of Bim. *Oncogene* 37, 2122–2136. doi: 10.1038/s41388-017-0112-0

Albert, L., Penalver, A., Djokovic, N., Werel, L., Hoffarth, M., Ruzic, D., et al. (2019). Modulating protein-protein interactions with visible-light-responsive peptide backbone photoswitches. *Chembiochem* 20, 1417–1429. doi: 10.1002/cbic.201800737

Alves Avelar, L. A., Ruzic, D., Djokovic, N., Kurz, T., and Nikolic, K. (2019). Structure-based design of selective histone deacetylase 6 zinc binding groups. *J. Biomol. Struct. Dyn.* 1–12. doi: 10.1080/07391102.2019.1652687. [Epub ahead of print].

Amr, A. E. E., Abo-Ghalia, M. H., Moustafa, G. O., Al-Omar, M. A., Nossier, E. S., and Elsayed, E. A. (2018). Design, synthesis and docking studies of novel macrocyclic pentapeptides as anticancer multi-targeted kinase inhibitors. *Molecules* 23:2416. doi: 10.3390/molecules23102416

Astanehe, A., Finkbeiner, M. R., Krzywinski, M., Fotovati, A., Dhillon, J., Berquin, I. M., et al. (2012). MKNK1 is a YB-1 target gene responsible for imparting trastuzumab resistance and can be blocked by RSK inhibition. *Oncogene* 31, 4434–4446. doi: 10.1038/onc.2011.617

Bacilieri, M., and Moro, S. (2006). Ligand-based drug design methodologies in drug discovery process: an overview. *Curr. Drug Discov. Technol.* 3, 155–165. doi: 10.2174/157016306780136781

Baell, J., and Walters, M. A. (2014). Chemistry: chemical con artists foil drug discovery. *Nature* 513, 481–483. doi: 10.1038/513481a

Bahar, I., and Rader, A. J. (2005). Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.* 15, 586–592. doi: 10.1016/j.sbi.2005.08.007

Bajorath, J. (2017). Molecular similarity concepts for informatics applications. *Methods Mol. Biol.* 1526, 231–245. doi: 10.1007/978-1-4939-6613-4_13

Bajusz, D., Racz, A., and Heberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* 7:20. doi: 10.1186/s13321-015-0069-3

Banegas-Luna, A. J., Ceron-Carrasco, J. P., and Perez-Sanchez, H. (2018). A review of ligand-based virtual screening web tools and screening algorithms in large molecular databases in the age of big data. *Future Med. Chem.* 10, 2641–2658. doi: 10.4155/fmc-2018-0076

Bautista-Aguilera, O. M., Esteban, G., Bolea, I., Nikolic, K., Agbaba, D., Moraleda, I., et al. (2014). Design, synthesis, pharmacological evaluation, QSAR analysis, molecular modeling and ADMET of novel donepezil-indolyl hybrids as multipotent cholinesterase/monoamine oxidase inhibitors for the potential treatment of Alzheimer's disease. *Eur. J. Med. Chem.* 75, 82–95. doi: 10.1016/j.ejmech.2013.12.028

Beauchamp, K. A., Lin, Y. S., Das, R., and Pande, V. S. (2012). Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *J. Chem. Theory Comput.* 8, 1409–1414. doi: 10.1021/ct2007814

Benod, C., Carlsson, J., Uthayaruban, R., Hwang, P., Irwin, J. J., Doak, A. K., et al. (2013). Structure-based discovery of antagonists of nuclear receptor LRH-1. *J. Biol. Chem.* 288, 19830–19844. doi: 10.1074/jbc.M112.411686

Bethune, G., Bethune, D., Ridgway, N., and Xu, Z. (2010). Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. *J. Thorac. Dis.* 2, 48–51.

Betzi, S., Alam, R., Martin, M., Lubbers, D. J., Han, H., Jakkaraj, S. R., et al. (2011). Discovery of a potential allosteric ligand binding site in CDK2. *ACS Chem. Biol.* 6, 492–501. doi: 10.1021/cb100410m

Bhullar, K. S., Lagaron, N. O., McGowan, E. M., Parmar, I., Jha, A., Hubbard, B. P., et al. (2018). Kinase-targeted cancer therapies: progress, challenges and future directions. *Mol. Cancer* 17:48. doi: 10.1186/s12943-018-0804-2

Blume-Jensen, P., and Hunter, T. (2001). Oncogenic kinase signalling. *Nature* 411, 355–365. doi: 10.1038/35077225

Bollag, G., Tsai, J., Zhang, J., Zhang, C., Ibrahim, P., Nolop, K., et al. (2012). Vemurafenib: the first drug approved for BRAF-mutant cancer. *Nat. Rev. Drug Discov.* 11, 873–886. doi: 10.1038/nrd3847

Buchdunger, E., Cioffi, C. L., Law, N., Stover, D., Ohno-Jones, S., Druker, B. J., et al. (2000). Abl protein-tyrosine kinase inhibitor STI571 inhibits *in vitro* signal transduction mediated by c-kit and platelet-derived growth factor receptors. *J. Pharmacol. Exp. Ther.* 295, 139–145.

Buchwald, F., Richter, L., and Kramer, S. (2011). Predicting a small molecule-kinase interaction map: a machine learning approach. *J. Cheminform.* 3:22. doi: 10.1186/1758-2946-3-22

Burgess, M. R., Skaggs, B. J., Shah, N. P., Lee, F. Y., and Sawyers, C. L. (2005). Comparative analysis of two clinically active BCR-ABL kinase inhibitors reveals the role of conformation-specific binding in resistance. *Proc. Natl. Acad. Sci. U.S.A.* 102, 3395–3400. doi: 10.1073/pnas.0409770102

Caballero, J., and Alzate-Morales, J. H. (2012). Molecular dynamics of protein kinase-inhibitor complexes: a valid structural information. *Curr. Pharm. Des.* 18, 2946–2963. doi: 10.2174/138161212800672705

Carles, F., Bourg, S., Meyer, C., and Bonnet, P. (2018). PKIDB: a curated, annotated and updated database of protein kinase inhibitors in clinical trials. *Molecules* 23:908. doi: 10.3390/molecules23040908

Chakraborty, S., Lin, Y. H., Leng, X., Miranda, R. N., Medeiros, L. J., Shpall, E., et al. (2013). Activation of Jak2 in patients with blast crisis chronic myelogenous leukemia: inhibition of Jak2 inactivates Lyn kinase. *Blood Cancer J.* 3:e142. doi: 10.1038/bcj.2013.41

Chen, J., Ma, X., Yuan, Y., Pei, J., and Lai, L. (2014). Protein-protein interface analysis and hot spots identification for chemical ligand design. *Curr. Pharm. Des.* 20, 1192–1200. doi: 10.2174/13816128113199990065

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., et al. (2014). QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* 57, 4977–5010. doi: 10.1021/jm4004285

Christmann-Franck, S., van Westen, G. J., Papadatos, G., Beltran Escudie, F., Roberts, A., Overington, J. P., et al. (2016). Unprecedently large-scale kinase

inhibitor set enabling the accurate prediction of compound-kinase activities: a way toward selective promiscuity by design? *J. Chem. Inf. Model.* 56, 1654–1675. doi: 10.1021/acs.jcim.6b00122

Cohen, P. (2002). Protein kinases — the major drug targets of the twenty-first century? *Nat. Rev. Drug Discov.* 1, 309–315. doi: 10.1038/nrd773

Collins, D. M., Conlon, N. T., Kannan, S., Verma, C. S., Eli, L. D., Lalani, A. S., et al. (2019). Preclinical characteristics of the irreversible pan-HER kinase inhibitor neratinib compared with lapatinib: implications for the treatment of HER2-positive and HER2-mutated breast cancer. *Cancers* 11:737. doi: 10.3390/cancers11060737

Congreve, M., Carr, R., Murray, C., and Jhoti, H. (2003). A 'Rule of Three' for fragment-based lead discovery? *Drug Discov. Today* 8, 876–877. doi: 10.1016/S1359-6446(03)02831-9

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., et al. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117, 5179–5197. doi: 10.1021/ja00124a002

Cramer, R. D., Patterson, D. E., and Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* 110, 5959–5967. doi: 10.1021/ja00226a005

Dar, A. C., and Shokat, K. M. (2011). The evolution of protein kinase inhibitors from antagonists to agonists of cellular signaling. *Annu. Rev. Biochem.* 80, 769–795. doi: 10.1146/annurev-biochem-090308-173656

de Angelo, R. M., Almeida, M. O., de Paula, H., and Honorio, K. M. (2018). Studies on the dual activity of EGFR and HER-2 inhibitors using structure-based drug design techniques. *Int. J. Mol. Sci.* 19:3728. doi: 10.3390/ijms19123728

Debnath, S., Kanakaraju, M., Islam, M., Yeeravalli, R., Sen, D., and Das, A. (2019). *In silico* design, synthesis and activity of potential drug-like chrysin scaffold-derived selective EGFR inhibitors as anticancer agents. *Comput. Biol. Chem.* 83:107156. doi: 10.1016/j.compbiolchem.2019.107156

Deeks, E. D. (2016). Venetoclax: first global approval. *Drugs* 76, 979–987. doi: 10.1007/s40265-016-0596-x

Deng, W., Breneman, C., and Embrechts, M. J. (2004). Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods. *J. Chem. Inf. Comput. Sci.* 44, 699–703. doi: 10.1021/ci034246+

Drewry, D. H., Wells, C. I., Andrews, D. M., Angell, R., Al-Ali, H., Axtman, A. D., et al. (2017). Progress towards a public chemogenomic set for protein kinases and a call for contributions. *PLoS ONE* 12:e0181585. doi: 10.1371/journal.pone.0181585

Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H., and Shaw, D. E. (2012). Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.* 41, 429–452. doi: 10.1146/annurev-biophys-042910-155245

Duong-Ly, K. C., and Peterson, J. R. (2013). The human kinome and kinase inhibition. *Curr. Protoc. Pharmacol.*60, 2.9.1–2.9.14. doi: 10.1002/0471141755.ph0209s60

Duran, A., Zamora, I., and Pastor, M. (2009). Suitability of GRIND-based principal properties for the description of molecular similarity and ligand-based virtual screening. *J. Chem. Inf. Model.* 49, 2129–2138. doi: 10.1021/ci900228x

Ekins, S. (2016). The next era: deep learning in pharmaceutical research. *Pharm. Res.* 33, 2594–2603. doi: 10.1007/s11095-016-2029-7

El Rashedy, A. A., Olotu, F. A., and Soliman, M. E. S. (2018). Dual drug targeting of mutant Bcr-Abl induces inactive conformation: new strategy for the treatment of chronic myeloid leukemia and overcoming monotherapy resistance. *Chem. Biodivers.* 15:e1700533. doi: 10.1002/cbdv.201700533

Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., and Mee, R. P. (1997). Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* 11, 425–445. doi: 10.1023/A:1007996124545

Fang, Z., Grutter, C., and Rauh, D. (2013). Strategies for the selective regulation of kinases with allosteric modulators: exploiting exclusive structural features. *ACS Chem. Biol.* 8, 58–70. doi: 10.1021/cb300663j

FDA (2019). *Novel Drug Approvals for 2019.* Available online at: https://www.fda.gov/drugs/new-drugs-fda-cders-new-molecular-entities-and-new-therapeutic-biological-products/novel-drug-approvals-2019

Ferreira, L. G., Dos Santos, R. N., Oliva, G., and Andricopulo, A. D. (2015). Molecular docking and structure-based drug design strategies. *Molecules* 20, 13384–13421. doi: 10.3390/molecules200713384

Fischer, T., Najjar, A., Totzke, F., Schachtele, C., Sippl, W., Ritter, C., et al. (2018). Discovery of novel dual inhibitors of receptor tyrosine kinases EGFR and PDGFR-beta related to anticancer drug resistance. *J. Enzyme Inhib. Med. Chem.* 33, 1–8. doi: 10.1080/14756366.2017.1370583

Flaherty, K. T., Infante, J. R., Daud, A., Gonzalez, R., Kefford, R. F., Sosman, J., et al. (2012). Combined BRAF and MEK inhibition in melanoma with BRAF V600 mutations. *N. Engl. J. Med.* 367, 1694–1703. doi: 10.1056/NEJMoa1210093

Force, T., and Kolaja, K. L. (2011). Cardiotoxicity of kinase inhibitors: the prediction and translation of preclinical models to clinical outcomes. *Nat. Rev. Drug Discov.* 10, 111–126. doi: 10.1038/nrd3252

Gagic, Z., Ivkovic, B., Srdic-Rajic, T., Vucicevic, J., Nikolic, K., and Agbaba, D. (2016a). Synthesis of the vitamin E amino acid esters with an enhanced anticancer activity and *in silico* screening for new antineoplastic drugs. *Eur. J. Pharm. Sci.* 88, 59–69. doi: 10.1016/j.ejps.2016.04.008

Gagic, Z., Nikolic, K., Ivkovic, B., Filipic, S., and Agbaba, D. (2016b). QSAR studies and design of new analogs of vitamin E with enhanced antiproliferative activity on MCF-7 breast cancer cells. *J. Taiwan Inst. Chem. Eng.* 59, 33–44. doi: 10.1016/j.jtice.2015.07.019

Gao, Y., Wang, H., Wang, J., and Cheng, M. (2019). *In silico* studies on p21-activated kinase 4 inhibitors: comprehensive application of 3D-QSAR analysis, molecular docking, molecular dynamics simulations, and MM-GBSA calculation. *J. Biomol. Struct. Dyn.* 1–15. doi: 10.1080/07391102.2019.1673823. [Epub ahead of print].

Gibbons, D. L., Pricl, S., Kantarjian, H., Cortes, J., and Quintas-Cardama, A. (2012). The rise and fall of gatekeeper mutations? The BCR-ABL1 T315I paradigm. *Cancer* 118, 293–299. doi: 10.1002/cncr.26225

Giordanetto, F., Kull, B., and Dellsén, A. (2011). Discovery of novel class 1 phosphatidylinositide 3-kinases (PI3K) fragment inhibitors through structure-based virtual screening. *Bioorg. Med. Chem. Lett.* 21, 829–835. doi: 10.1016/j.bmcl.2010.11.087

Giordanetto, F., Wallberg, A., Cassel, J., Ghosal, S., Kossenjans, M., Yuan, Z. Q., et al. (2012). Discovery of 4-morpholino-pyrimidin-6-one and 4-morpholino-pyrimidin-2-one-containing Phosphoinositide 3-kinase (PI3K) p110beta isoform inhibitors through structure-based fragment optimisation. *Bioorg. Med. Chem. Lett.* 22, 6665–6670. doi: 10.1016/j.bmcl.2012.08.101

Gkeka, P., Papafotika, A., Christoforidis, S., and Cournia, Z. (2015). Exploring a non-ATP pocket for potential allosteric modulation of PI3Kalpha. *J. Phys. Chem. B* 119, 1002–1016. doi: 10.1021/jp506423e

Gomez-Gutierrez, P., Campos, P. M., Vega, M., and Perez, J. J. (2016). Identification of a Novel Inhibitory Allosteric Site in p38alpha. *PLoS ONE* 11:e0167379. doi: 10.1371/journal.pone.0167379

Gopalsamy, A., Ciszewski, G., Hu, Y., Lee, F., Feldberg, L., Frommer, E., et al. (2009). Identification of pyrazolo[1,5-a]pyrimidine-3-carboxylates as B-Raf kinase inhibitors. *Bioorg. Med. Chem. Lett.* 19, 2735–2738. doi: 10.1016/j.bmcl.2009.03.129

Graves, J. D., and Krebs, E. G. (1999). Protein phosphorylation and signal transduction. *Pharmacol. Ther.* 82, 111–121. doi: 10.1016/S0163-7258(98)00056-4

Greener, J. G., and Sternberg, M. J. (2018). Structure-based prediction of protein allostery. *Curr. Opin. Struct. Biol.* 50, 1–8. doi: 10.1016/j.sbi.2017.10.002

Guo, J., and Zhou, H. X. (2016). Protein allostery and conformational dynamics. *Chem. Rev.* 116, 6503–6515. doi: 10.1021/acs.chemrev.5b00590

Halder, A. K., Giri, A. K., and Cordeiro, M. (2019). Multi-target chemometric modelling, fragment analysis and virtual screening with ERK inhibitors as potential anticancer agents. *Molecules* 24:E3909. doi: 10.3390/molecules24213909

Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* 47, 1750–1759. doi: 10.1021/jm030644s

Hann, M. M., Leach, A. R., and Harper, G. (2001). Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* 41, 856–864. doi: 10.1021/ci000403i

Hansch, C., Maloney, P. P., Fujita, T., and Muir, R. M. (1962). Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* 194, 178–180. doi: 10.1038/194178b0

Harris, P. A., Boloor, A., Cheung, M., Kumar, R., Crosby, R. M., Davis-Ward, R. G., et al. (2008). Discovery of 5-[[4-[(2,3-dimethyl-2H-indazol-6-yl)methylamino]-2-pyrimidinyl]amino]-2-methyl-b enzenesulfonamide (Pazopanib), a novel and potent vascular endothelial growth factor receptor inhibitor. *J. Med. Chem.* 51, 4632–4640. doi: 10.1021/jm800566m

Harris, P. A., Cheung, M., Hunter, R. N. III, Brown, M. L., Veal, J. M., Nolte, R. T., et al. (2005). Discovery and evaluation of 2-anilino-5-aryloxazoles as a novel class of VEGFR2 kinase inhibitors. *J. Med. Chem.* 48, 1610–1619. doi: 10.1021/jm049538w

Hasinoff, B. B. (2010). The cardiotoxicity and myocyte damage caused by small molecule anticancer tyrosine kinase inhibitors is correlated with lack of target specificity. *Toxicol. Appl. Pharmacol.* 244, 190–195. doi: 10.1016/j.taap.2009.12.032

Hazhazi, H., Melkemi, N., Salah, T., and Bouachrine, M. (2019). DFT-based reactivity and combined QSAR, molecular docking of 1,2,4,5-Tetrazine derivatives as inhibitors of Pim-1 kinase. *Heliyon* 5:e02451. doi: 10.1016/j.heliyon.2019.e02451

Holohan, C., Van Schaeybroeck, S., Longley, D. B., and Johnston, P. G. (2013). Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer* 13, 714–726. doi: 10.1038/nrc3599

Hu, X., Wang, D., Tong, Y., Tong, L., Wang, X., Zhu, L., et al. (2017). Design, synthesis, and evaluation of ribose-modified anilinopyrimidine derivatives as EGFR tyrosine kinase inhibitors. *Front. Chem.* 5:101. doi: 10.3389/fchem.2017.00101

Huang, M., Song, K., Liu, X., Lu, S., Shen, Q., Wang, R., et al. (2018). AlloFinder: a strategy for allosteric modulator discovery and allosterome analyses. *Nucleic Acids Res.* 46, W451–W458. doi: 10.1093/nar/gky374

Huang, N., Kalyanaraman, C., Irwin, J. J., and Jacobson, M. P. (2006). Physics-based scoring of protein-ligand complexes: enrichment of known inhibitors in large-scale virtual screening. *J. Chem. Inf. Model.* 46, 243–253. doi: 10.1021/ci0502855

Hubbard, S. R., Wei, L., Ellis, L., and Hendrickson, W. A. (1994). Crystal structure of the tyrosine kinase domain of the human insulin receptor. *Nature* 372, 746–754. doi: 10.1038/372746a0

Hussain, A., and Verma, C. K. (2019). Ligand- and structure-based pharmacophore modeling, docking study reveals 2-[[4-[6-(isopropylamino) pyrimidin-4-yl]-1H-pyrrolo[2,3-b] pyridin-6-yl] amino] ethanol as a potential anticancer agent of CDK9/cyclin T1 kinase. *J. Cancer Res. Ther.* 15, 1131–1140. doi: 10.4103/jcrt.JCRT_47_18

Iqbal, N., and Iqbal, N. (2014). Imatinib: a breakthrough of targeted therapy in cancer. *Chemother. Res. Pract.* 2014:357027. doi: 10.1155/2014/357027

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. (2012). ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* 52, 1757–1768. doi: 10.1021/ci3001277

Jahnke, W., Grotzfeld, R. M., Pelle, X., Strauss, A., Fendrich, G., Cowan-Jacob, S. W., et al. (2010). Binding or bending: distinction of allosteric Abl kinase agonists from antagonists by an NMR-based conformational assay. *J. Am. Chem. Soc.* 132, 7043–7048. doi: 10.1021/ja101837n

James, N., Shanthi, V., and Ramanathan, K. (2019). Density functional theory and molecular simulation studies for prioritizing anaplastic lymphoma kinase inhibitors. *Appl. Biochem. Biotechnol.* doi: 10.1007/s12010-019-03156-1. [Epub ahead of print].

Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* 267, 727–748. doi: 10.1006/jmbi.1996.0897

Jorgensen, W. L., Maxwell, D. S., and Tirado-Rives, J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* 118, 11225–11236. doi: 10.1021/ja9621760

Kaieda, A., Takahashi, M., Fukuda, H., Okamoto, R., Morimoto, S., Gotoh, M., et al. (2019). Structure-based design, synthesis, and biological evaluation of imidazo[4,5-b]pyridin-2-one-based p38 MAP kinase inhibitors: part 1. *ChemMedChem* 14, 1022–1030. doi: 10.1002/cmdc.201900129

Kale, M. A., and Sonwane, G. M. (2018). Molecular docking, G-QSAR studies, synthesis and anticancer screening of some new 2-phenazinamines as Bcr-Abl tyrosine kinase inhibitors. *Curr. Drug Discov. Technol.* doi: 10.2174/1570163815666180913122542. [Epub ahead of print].

Kannan, S., Poulsen, A., Yang, H. Y., Ho, M., Ang, S. H., Eldwin, T. S. W., et al. (2017). Correction to probing the binding mechanism of Mnk inhibitors by docking and molecular dynamics simulations. *Biochemistry* 56, 5964–5965. doi: 10.1021/acs.biochem.7b00996

Khodair, A. I., Alsafi, M. A., and Nafie, M. S. (2019). Synthesis, molecular modeling and anti-cancer evaluation of a series of quinazoline derivatives. *Carbohydr. Res.* 486:107832. doi: 10.1016/j.carres.2019.107832

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2019). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* 47, D1102–D1109. doi: 10.1093/nar/gky1033

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* 3, 935–949. doi: 10.1038/nrd1549

Klaeger, S., Heinzlmeir, S., Wilhelm, M., Polzer, H., Vick, B., Koenig, P. A., et al. (2017). The target landscape of clinical kinase drugs. *Science* 358:eaan4368. doi: 10.1126/science.aan4368

Klebe, G., Abraham, U., and Mietzner, T. (1994). Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* 37, 4130–4146. doi: 10.1021/jm00050a010

Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O., and Shaw, D. E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* 19, 120–127. doi: 10.1016/j.sbi.2009.03.004

Knight, J. D., Qian, B., Baker, D., and Kothary, R. (2007). Conservation, variability and the modeling of active protein kinases. *PLoS ONE* 2:e982. doi: 10.1371/journal.pone.0000982

Knighton, D. R., Zheng, J. H., Ten Eyck, L. F., Ashford, V. A., Xuong, N. H., Taylor, S. S., et al. (1991). Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* 253, 407–414. doi: 10.1126/science.1862342

Koneru, J. K., Sinha, S., and Mondal, J. (2019). *In Silico* reoptimization of binding affinity and drug-resistance circumvention ability in kinase inhibitors: a case study with RL-45 and Src kinase. *J. Phys. Chem. B* 123, 6664–6672. doi: 10.1021/acs.jpcb.9b02883

Konze, K. D., Bos, P. H., Dahlgren, M. K., Leswing, K., Tubert-Brohman, I., Bortolato, A., et al. (2019). Reaction-based enumeration, active learning, and free energy calculations to rapidly explore synthetically tractable chemical space and optimize potency of cyclin-dependent kinase 2 inhibitors. *J. Chem. Inf. Model.* 59, 3782–3793. doi: 10.1021/acs.jcim.9b00367

Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., et al. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* 10, 733–755. doi: 10.1038/nprot.2015.043

Kumar, V., Krishna, S., and Siddiqi, M. I. (2015). Virtual screening strategies: recent advances in the identification and design of anti-cancer agents. *Methods* 71, 64–70. doi: 10.1016/j.ymeth.2014.08.010

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* 161, 269–288. doi: 10.1016/0022-2836(82)90153-X

Kuthuru, S., Szafran, A. T., Stossi, F., Mancini, M. A., and Rao, A. (2019). Leveraging image-derived phenotypic measurements for drug-target interaction predictions. *Cancer Inform.* 18:1176935119856595. doi: 10.1177/1176935119856595

La Sala, G., Riccardi, L., Gaspari, R., Cavalli, A., Hantschel, O., and De Vivo, M. (2016). HRD motif as the central hub of the signaling network for activation loop autophosphorylation in Abl kinase. *J. Chem. Theory Comput.* 12, 5563–5574. doi: 10.1021/acs.jctc.6b00600

Lamba, V., and Ghosh, I. (2012). New directions in targeting protein kinases: focusing upon true allosteric and bivalent inhibitors. *Curr. Pharm. Des.* 18, 2936–2945. doi: 10.2174/138161212800672813

Larif, S., Salem, C. B., Hmouda, H., and Bouraoui, K. (2014). *In silico* screening and study of novel ERK2 inhibitors using 3D QSAR, docking and molecular dynamics. *J. Mol. Graph. Model.* 53, 1–12. doi: 10.1016/j.jmgm.2014.07.001

Laurini, E., Posocco, P., Fermeglia, M., Gibbons, D. L., Quintas-Cardama, A., and Pricl, S. (2013). Through the open door: preferential binding of dasatinib to the active form of BCR-ABL unveiled by *in silico* experiments. *Mol. Oncol.* 7, 968–975. doi: 10.1016/j.molonc.2013.06.001

Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* 10:168. doi: 10.1186/1471-2105-10-168

Leach, A. R., and Hann, M. M. (2011). Molecular complexity and fragment-based drug discovery: ten years on. *Curr. Opin. Chem. Biol.* 15, 489–496. doi: 10.1016/j.cbpa.2011.05.008

Lee, H. W., Choi, H. Y., Joo, K. M., and Nam, D. H. (2015). Tumor progression locus 2 (Tpl2) kinase as a novel therapeutic target for cancer: double-sided effects of Tpl2 on cancer. *Int. J. Mol. Sci.* 16, 4471–4491. doi: 10.3390/ijms16034471

Lee, I., Keum, J., and Nam, H. (2019). DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol.* 15:e1007129. doi: 10.1371/journal.pcbi.1007129

Lengauer, T., and Rarey, M. (1996). Computational methods for biomolecular docking. *Curr. Opin. Struct. Biol.* 6, 402–406. doi: 10.1016/S0959-440X(96)80061-3

Li, J., Fu, A., and Zhang, L. (2019). An overview of scoring functions used for protein-ligand interactions in molecular docking. *Interdiscip. Sci.* 11, 320–328. doi: 10.1007/s12539-019-00327-w

Li, S., Shen, Q., Su, M., Liu, X., Lu, S., Chen, Z., et al. (2016). Alloscore: a method for predicting allosteric ligand-protein interactions. *Bioinformatics* 32, 1574–1576. doi: 10.1093/bioinformatics/btw036

Lim, H., He, D., Qiu, Y., Krawczuk, P., Sun, X., and Xie, L. (2019). Rational discovery of dual-indication multi-target PDE/Kinase inhibitor for precision anti-cancer therapy using structural systems pharmacology. *PLoS Comput. Biol.* 15:e1006619. doi: 10.1371/journal.pcbi.1006619

Lionta, E., Spyrou, G., Vassilatis, D. K., and Cournia, Z. (2014). Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr. Top. Med. Chem.* 14, 1923–1938. doi: 10.2174/1568026614666140929124445

Liu, J., and Wang, R. (2015). Classification of current scoring functions. *J. Chem. Inf. Model.* 55, 475–482. doi: 10.1021/ci500731a

Liu, J., Wen, Y., Gao, L., Gao, L., He, F., Zhou, J., et al. (2020). Design, synthesis and biological evaluation of novel 1H-1,2,4-triazole, benzothiazole and indazole-based derivatives as potent FGFR1 inhibitors viafragment-based virtual screening. *J. Enzyme Inhib. Med. Chem.* 35, 72–84. doi: 10.1080/14756366.2019.1673745

Lu, S., He, X., Ni, D., and Zhang, J. (2019). Allosteric modulator discovery: from serendipity to structure-based design. *J. Med. Chem.* 62, 6405–6421. doi: 10.1021/acs.jmedchem.8b01749

Lu, S., Ji, M., Ni, D., and Zhang, J. (2018a). Discovery of hidden allosteric sites as novel targets for allosteric drug design. *Drug Discov. Today* 23, 359–365. doi: 10.1016/j.drudis.2017.10.001

Lu, W., Zhang, R., Jiang, H., Zhang, H., and Luo, C. (2018b). Computer-aided drug design in epigenetics. *Front. Chem.* 6:57. doi: 10.3389/fchem.2018.00057

Ma, X., Meng, H., and Lai, L. (2016). Motions of allosteric and orthosteric ligand-binding sites in proteins are highly correlated. *J. Chem. Inf. Model.* 56, 1725–1733. doi: 10.1021/acs.jcim.6b00039

Machrouhi, F., Ouhamou, N., Laderoute, K., Calaoagan, J., Bukhtiyarova, M., Ehrlich, P. J., et al. (2010). The rational design of a novel potent analogue of the 5'-AMP-activated protein kinase inhibitor compound C with improved selectivity and cellular activity. *Bioorg. Med. Chem. Lett.* 20, 6394–6399. doi: 10.1016/j.bmcl.2010.09.088

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science* 298, 1912–1934. doi: 10.1126/science.1075762

Markham, A. (2019). Erdafitinib: first global approval. *Drugs* 79, 1017–1021. doi: 10.1007/s40265-019-01142-9

Martin, M. P., Alam, R., Betzi, S., Ingles, D. J., Zhu, J. Y., and Schonbrunn, E. (2012). A novel approach to the discovery of small-molecule ligands of CDK2. *Chembiochem* 13, 2128–2136. doi: 10.1002/cbic.201200316

Melge, A. R., Kumar, L. G., Pavithran, K., Nair, S. V., Manzoor, K., Gopi Mohan, C. (2019). Predictive models for designing potent tyrosine kinase inhibitors in chronic myeloid leukemia for understanding its molecular mechanism of resistance by molecular docking and dynamics simulations. *J. Biomol. Struct. Dyn.* 37, 4747–4766. doi: 10.1080/07391102.2018.1559765

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Felix, E., et al. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* 47, D930–D940. doi: 10.1093/nar/gky1075

Meng, Y., Gao, C., Clawson, D. K., Atwell, S., Russell, M., Vieth, M., et al. (2018). Predicting the conformational variability of Abl tyrosine kinase using molecular dynamics simulations and markov state models. *J. Chem. Theory Comput.* 14, 2721–2732. doi: 10.1021/acs.jctc.7b01170

Merget, B., Turk, S., Eid, S., Rippmann, F., and Fulle, S. (2017). Profiling prediction of kinase inhibitors: toward the virtual assay. *J. Med. Chem.* 60, 474–485. doi: 10.1021/acs.jmedchem.6b01611

Merk, D., Friedrich, L., Grisoni, F., and Schneider, G. (2018). *De novo* design of bioactive small molecules by artificial intelligence. *Mol. Inform.* 37. doi: 10.1002/minf.201700153

Mobitz, H. (2015). The ABC of protein kinase conformations. *Biochim. Biophys. Acta* 1854(10 Pt B), 1555–1566. doi: 10.1016/j.bbapap.2015.03.009

Modi, V., and Dunbrack, R. L. Jr. (2019). Defining a new nomenclature for the structures of active and inactive kinases. *Proc. Natl. Acad. Sci. U.S.A.* 116, 6818–6827. doi: 10.1073/pnas.1814279116

Mohamed, T. K., Batran, R. Z., Elseginy, S. A., Ali, M. M., and Mahmoud, A. E. (2019). Synthesis, anticancer effect and molecular modeling of new thiazolylpyrazolyl coumarin derivatives targeting VEGFR-2 kinase and inducing cell cycle arrest and apoptosis. *Bioorg. Chem.* 85, 253–273. doi: 10.1016/j.bioorg.2018.12.040

Mohareb, R. M., Samir, E. M., and Halim, P. A. (2019). Synthesis, and anti-proliferative, Pim-1 kinase inhibitors and molecular docking of thiophenes derived from estrone. *Bioorg. Chem.* 83, 402–413. doi: 10.1016/j.bioorg.2018.10.067

Morabito, A., De Maio, E., Di Maio, M., Normanno, N., and Perrone, F. (2006). Tyrosine kinase inhibitors of vascular endothelial growth factor receptors in clinical trials: current status and future directions. *Oncologist* 11, 753–764. doi: 10.1634/theoncologist.11-7-753

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* 30, 2785–2791. doi: 10.1002/jcc.21256

Morris, G. M., and Lim-Wilby, M. (2008). Molecular docking. *Methods Mol. Biol.* 443, 365–382. doi: 10.1007/978-1-59745-177-2_19

Muegge, I., and Martin, Y. C. (1999). A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* 42, 791–804. doi: 10.1021/jm980536j

Murray, C. W., and Verdonk, M. L. (2002). The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J. Comput. Aided Mol. Des.* 16, 741–753. doi: 10.1023/A:1022446720849

Naboulsi, I., Aboulmouhajir, A., Kouisni, L., Bekkaoui, F., and Yasri, A. (2018). Combining a QSAR approach and structural analysis to derive an SAR map of Lyn kinase inhibition. *Molecules* 23:3271. doi: 10.3390/molecules23123271

Nagar, B., Hantschel, O., Young, M. A., Scheffzek, K., Veach, D., Bornmann, W., et al. (2003). Structural basis for the autoinhibition of c-Abl tyrosine kinase. *Cell* 112, 859–871. doi: 10.1016/S0092-8674(03)00194-6

Najjar, A., Platzer, C., Luft, A., Assmann, C. A., Elghazawy, N. H., Erdmann, F., et al. (2019). Computer-aided design, synthesis and biological characterization of novel inhibitors for PKMYT1. *Eur. J. Med. Chem.* 161, 479–492. doi: 10.1016/j.ejmech.2018.10.050

Neves, B. J., Braga, R. C., Melo-Filho, C. C., Moreira-Filho, J. T., Muratov, E. N., and Andrade, C. H. (2018). QSAR-based virtual screening: advances and applications in drug discovery. *Front. Pharmacol.* 9:1275. doi: 10.3389/fphar.2018.01275

Nikolic, K., Agbaba, D., and Stark, H. (2015). Pharmacophore modeling, drug design and virtual screening on multi-targeting procognitive agents approaching histaminergic pathways. *J. Taiwan Inst. Chem. Eng.* 46, 15–29. doi: 10.1016/j.jtice.2014.09.017

Nikolic, K., Veljkovic, N., Gemovic, B., Srdic-Rajic, T., and Agbaba, D. (2013). Imidazoline-1 receptor ligands as apoptotic agents: pharmacophore modeling and virtual docking study. *Comb. Chem. High Throughput Screen* 16, 298–319. doi: 10.2174/1386207311316040004

Noolvi, M. N., and Patel, H. M. (2013). A comparative QSAR analysis and molecular docking studies of quinazoline derivatives as tyrosine kinase (EGFR)

inhibitors: a rational approach to anticancer drug design. *J. Saudi Chem. Soc.* 17, 361–379. doi: 10.1016/j.jscs.2011.04.017

Novič, M., Tibaut, T., Anderluh, M., Borišek, J., and Tomašič, T. (2016). "The comparison of docking search algorithms and scoring functions," in *Methods and Algorithms for Molecular Docking-Based Drug Design and Discovery,* ed S. Dastmalchi, M. Hamzeh-Mivehroud, B. Sokouti. (Hershey, PA: IGI Global), 99–127.

Ohren, J. F., Chen, H., Pavlovsky, A., Whitehead, C., Zhang, E., Kuffa, P., et al. (2004). Structures of human MAP kinase kinase 1 (MEK1) and MEK2 describe novel noncompetitive kinase inhibition. *Nat. Struct. Mol. Biol.* 11, 1192–1197. doi: 10.1038/nsmb859

Okabe, S., Tauchi, T., Tanaka, Y., and Ohyashiki, K. (2011). Dasatinib preferentially induces apoptosis by inhibiting Lyn kinase in nilotinib-resistant chronic myeloid leukemia cell line. *J. Hematol. Oncol.* 4:32. doi: 10.1186/1756-8722-4-32

Oluic, J., Nikolic, K., Vucicevic, J., Gagic, Z., Filipic, S., and Agbaba, D. (2017). 3D-QSAR, virtual screening, docking and design of dual PI3K/mTOR inhibitors with enhanced antiproliferative activity. *Comb. Chem. High Throughput Screen* 20, 292–303. doi: 10.2174/1386207320666170427143858

Oostenbrink, C., Villa, A., Mark, A. E., and van Gunsteren, W. F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* 25, 1656–1676. doi: 10.1002/jcc.20090

Oyarzabal, J., Zarich, N., Albarran, M. I., Palacios, I., Urbano-Cuadrado, M., Mateos, G., et al. (2010). Discovery of mitogen-activated protein kinase-interacting kinase 1 inhibitors by a comprehensive fragment-oriented virtual screening approach. *J. Med. Chem.* 53, 6618–6628. doi: 10.1021/jm1005513

Park, E., Kim, N., Ficarro, S. B., Zhang, Y., Lee, B. I., Cho, A., et al. (2015). Structure and mechanism of activity-based inhibition of the EGF receptor by Mig6. *Nat. Struct. Mol. Biol.* 22, 703–711. doi: 10.1038/nsmb.3074

Pastor, M., Cruciani, G., McLay, I., Pickett, S., and Clementi, S. (2000). GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* 43, 3233–3243. doi: 10.1021/jm000941m

Piana, S., Klepeis, J. L., and Shaw, D. E. (2014). Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.* 24, 98–105. doi: 10.1016/j.sbi.2013.12.006

Platzer, C., Najjar, A., Rohe, A., Erdmann, F., Sippl, W., and Schmidt, M. (2018). Identification of PKMYT1 inhibitors by screening the GSK published protein kinase inhibitor set I and II. *Bioorg. Med. Chem.* 26, 4014–4024. doi: 10.1016/j.bmc.2018.06.027

Proud, C. G. (2015). Mnks, eIF4E phosphorylation and cancer. *Biochim. Biophys. Acta* 1849, 766–773. doi: 10.1016/j.bbagrm.2014.10.003

Putin, E., Asadulaev, A., Ivanenkov, Y., Aladinskiy, V., Sanchez-Lengeling, B., Aspuru-Guzik, A., et al. (2018). Reinforced adversarial neural computer for *de novo* molecular design. *J. Chem. Inf. Model.* 58, 1194–1204. doi: 10.1021/acs.jcim.7b00690

Rastelli, G., Anighoro, A., Chripkova, M., Carrassa, L., and Broggini, M. (2014). Structure-based discovery of the first allosteric inhibitors of cyclin-dependent kinase 2. *Cell Cycle* 13, 2296–2305. doi: 10.4161/cc.29295

Raval, A., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins* 80, 2071–2079. doi: 10.1002/prot.24098

Ren, J. X., Li, L. L., Zheng, R. L., Xie, H. Z., Cao, Z. X., Feng, S., et al. (2011). Discovery of novel Pim-1 kinase inhibitors by a hierarchical multistage virtual screening approach based on SVM model, pharmacophore, and molecular docking. *J. Chem. Inf. Model.* 51, 1364–1375. doi: 10.1021/ci100464b

Rettenmaier, T. J., Fan, H., Karpiak, J., Doak, A., Sali, A., Shoichet, B. K., et al. (2015). Small-molecule allosteric modulators of the protein kinase PDK1 from structure-based docking. *J. Med. Chem.* 58, 8285–8291. doi: 10.1021/acs.jmedchem.5b01216

Roskoski, R. Jr. (2015). A historical overview of protein kinases and their targeted small molecule inhibitors. *Pharmacol. Res.* 100, 1–23. doi: 10.1016/j.phrs.2015.07.010

Roskoski, R. Jr. (2016). Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. *Pharmacol. Res.* 103, 26–48. doi: 10.1016/j.phrs.2015.10.021

Roskoski, R. Jr. (2019a). Cyclin-dependent protein serine/threonine kinase inhibitors as anticancer drugs. *Pharmacol. Res.* 139, 471–488. doi: 10.1016/j.phrs.2018.11.035

Roskoski, R. Jr. (2019b). Properties of FDA-approved small molecule protein kinase inhibitors. *Pharmacol. Res.* 144, 19–50. doi: 10.1016/j.phrs.2019.03.006

Ruzic, D., Petkovic, M., Agbaba, D., Ganesan, A., and Nikolic, K. (2019). Combined ligand and fragment-based drug design of selective histone deacetylase - 6 inhibitors. *Mol. Inform.* 38:e1800083. doi: 10.1002/minf.201800083

Schoepfer, J., Jahnke, W., Berellini, G., Buonamici, S., Cotesta, S., Cowan-Jacob, S. W., et al. (2018). Discovery of Asciminib (ABL001), an allosteric inhibitor of the tyrosine kinase activity of BCR-ABL1. *J. Med. Chem.* 61, 8120–8135. doi: 10.1021/acs.jmedchem.8b01040

Schulze, J. O., Saladino, G., Busschots, K., Neimanis, S., Suss, E., Odadzic, D., et al. (2016). Bidirectional allosteric communication between the ATP-binding site and the regulatory PIF pocket in PDK1 protein kinase. *Cell Chem. Biol.* 23, 1193–1205. doi: 10.1016/j.chembiol.2016.06.017

Seiler, K. P., George, G. A., Happ, M. P., Bodycombe, N. E., Carrinski, H. A., Norton, S., et al. (2008). ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* 36, D351–D359. doi: 10.1093/nar/gkm843

Sengupta, U., and Strodel, B. (2018). Markov models for the elucidation of allosteric regulation. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 373:20170178. doi: 10.1098/rstb.2017.0178

Shah, N. P., Nicoll, J. M., Nagar, B., Gorre, M. E., Paquette, R. L., Kuriyan, J., et al. (2002). Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. *Cancer Cell* 2, 117–125. doi: 10.1016/S1535-6108(02)00096-X

Shah, N. P., Tran, C., Lee, F. Y., Chen, P., Norris, D., and Sawyers, C. L. (2004). Overriding imatinib resistance with a novel ABL kinase inhibitor. *Science* 305, 399–401. doi: 10.1126/science.1099480

Shi, Z., Resing, K. A., and Ahn, N. G. (2006). Networks for the allosteric control of protein kinases. *Curr. Opin. Struct. Biol.* 16, 686–692. doi: 10.1016/j.sbi.2006.10.011

Shiri, F., Pirhadi, S., and Ghasemi, J. B. (2016). Alignment independent 3D-QSAR, quantum calculations and molecular docking of Mer specific tyrosine kinase inhibitors as anticancer drugs. *Saudi Pharm. J.* 24, 197–212. doi: 10.1016/j.jsps.2015.03.012

Shirts, M., and Pande, V. S. (2000). COMPUTING: screen savers of the world unite! *Science* 290, 1903–1904. doi: 10.1126/science.290.5498.1903

Shuker, S. B., Hajduk, P. J., Meadows, R. P., and Fesik, S. W. (1996). Discovering high-affinity ligands for proteins: SAR by NMR. *Science* 274, 1531–1534. doi: 10.1126/science.274.5292.1531

Shukla, D., Meng, Y., Roux, B., and Pande, V. S. (2014). Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nat. Commun.* 5:3397. doi: 10.1038/ncomms4397

Simeon, S., Jongkon, N., Chotpatiwetchkul, W., and Gleeson, M. P. (2019). Insights into the EGFR SAR of N-phenylquinazolin-4-amine-derivatives using quantum mechanical pairwise-interaction energies. *J. Comput. Aided Mol. Des.* 33, 745–757. doi: 10.1007/s10822-019-00221-z

Smajić, M., Nikoli,ć, K., Vujić, Z., Ahmetović, L., and Kuntić, V. (2015). 3D-QSAR studies and pharmacophore identification of AT1 receptor antagonists. *Med. Chem. Res.* 25, 51–61. doi: 10.1007/s00044-015-1470-1

Sobhy, M. K., Mowafy, S., Lasheen, D. S., Farag, N. A., and Abouzid, K. A. M. (2019). 3D-QSAR pharmacophore modelling, virtual screening and docking studies for lead discovery of a novel scaffold for VEGFR 2 inhibitors: design, synthesis and biological evaluation. *Bioorg. Chem.* 89:102988. doi: 10.1016/j.bioorg.2019.102988

Song, K., Liu, X., Huang, W., Lu, S., Shen, Q., Zhang, L., et al. (2017). Improved method for the identification and validation of allosteric sites. *J. Chem. Inf. Model.* 57, 2358–2363. doi: 10.1021/acs.jcim.7b00014

Sousa, S. F., Fernandes, P. A., and Ramos, M. J. (2006). Protein-ligand docking: current status and future challenges. *Proteins* 65, 15–26. doi: 10.1002/prot.21082

Subramanian, V., Prusis, P., Pietila, L. O., Xhaard, H., and Wohlfahrt, G. (2013). Visually interpretable models of kinase selectivity related features derived from field-based proteochemometrics. *J. Chem. Inf. Model.* 53, 3021–3030. doi: 10.1021/ci400369z

Subramanian, V., Prusis, P., Xhaard, H., and Wohlfahrt, G. (2016). Predictive proteochemometric models for kinases derived from 3D protein field-based descriptors. *Medchemcomm* 7, 1007–1015. doi: 10.1039/C5MD00556F

Sudha, A., Srinivasan, P., and Rameshthangam, P. (2015). Exploration of potential EGFR inhibitors: a combination of pharmacophore-based virtual screening, atom-based 3D-QSAR and molecular docking analysis. *J. Recept. Signal Transduct. Res.* 35, 137–148. doi: 10.3109/10799893.2014.942461

Sun, C., Petros, A. M., and Hajduk, P. J. (2011). Fragment-based lead discovery: challenges and opportunities. *J. Comput. Aided Mol. Des.* 25, 607–610. doi: 10.1007/s10822-011-9451-z

Sun, H. (1998). COMPASS: An *ab initio* force-field optimized for condensed-phase applicationsoverview with details on alkane and benzene compounds. *J. Phys. Chem. B* 102, 7338–7364. doi: 10.1021/jp980939v

Taylor, R. D., Jewsbury, P. J., and Essex, J. W. (2002). A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.* 16, 151–166. doi: 10.1023/A:1020155510718

Taylor, S. S., and Kornev, A. P. (2011). Protein kinases: evolution of dynamic regulatory proteins. *Trends Biochem. Sci.* 36, 65–77. doi: 10.1016/j.tibs.2010.09.006

Taylor, S. S., Radzio-Andzelm, E., and Hunter, T. (1995). How do protein kinases discriminate between serine/threonine and tyrosine? Structural insights from the insulin receptor protein-tyrosine kinase. *FASEB J.* 9, 1255–1266. doi: 10.1096/fasebj.9.13.7557015

Teli, M. K., and Rajanikant, G. K. (2012). Pharmacophore generation and atom-based 3D-QSAR of novel quinoline-3-carbonitrile derivatives as Tpl2 kinase inhibitors. *J. Enzyme Inhib. Med. Chem.* 27, 558–570. doi: 10.3109/14756366.2011.603128

Trott, O., and Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455–461. doi: 10.1002/jcc.21334

Tsou, H. R., Otteng, M., Tran, T., Floyd, M. B. Jr., Reich, M., Birnberg, G., et al. (2008). 4-(Phenylaminomethylene)isoquinoline-1,3(2H,4H)-diones as potent and selective inhibitors of the cyclin-dependent kinase 4 (CDK4). *J. Med. Chem.* 51, 3507–3525. doi: 10.1021/jm800072z

Ung, P. M., Rahman, R., and Schlessinger, A. (2018). Redefining the protein kinase conformational space with machine learning. *Cell Chem. Biol.* 25, 916–924.e912. doi: 10.1016/j.chembiol.2018.05.002

Vanderpool, D., Johnson, T. O., Ping, C., Bergqvist, S., Alton, G., Phonephaly, S., et al. (2009). Characterization of the CHK1 allosteric inhibitor binding site. *Biochemistry* 48, 9823–9830. doi: 10.1021/bi900258v

Vanommeslaeghe, K., Guvench, O., and MacKerell, A. D. Jr. (2014). Molecular mechanics. *Curr. Pharm. Des.* 20, 3281–3292. doi: 10.2174/13816128113199990600

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins* 52, 609–623. doi: 10.1002/prot.10465

Vittorio, S., Seidel, T., Germano, M. P., Gitto, R., Ielo, L., Garon, A., et al. (2019). A combination of pharmacophore and docking-based virtual screening to discover new tyrosinase inhibitors. *Mol. Inform.* doi: 10.1002/minf.201900054. [Epub ahead of print].

Vucicevic, J., Nikolic, K., and Mitchell, J. B. O. (2017). Rational drug design of antineoplastic agents using 3D-QSAR, cheminformatic, and virtual screening approaches. *Curr. Med. Chem.* 26, 3874–3889. doi: 10.2174/0929867324666170712115411

Vucicevic, J., Srdic-Rajic, T., Pieroni, M., Laurila, J. M., Perovic, V., Tassini, S., et al. (2016). A combined ligand- and structure-based approach for the identification of rilmenidine-derived compounds which synergize the antitumor effects of doxorubicin. *Bioorg. Med. Chem.* 24, 3174–3183. doi: 10.1016/j.bmc.2016.05.043

Wagner, J. R., Lee, C. T., Durrant, J. D., Malmstrom, R. D., Feher, V. A., and Amaro, R. E. (2016). Emerging computational methods for the rational discovery of allosteric drugs. *Chem. Rev.* 116, 6370–6390. doi: 10.1021/acs.chemrev.5b00631

Wan, P. T. C., Garnett, M. J., Roe, S. M., Lee, S., Niculescu-Duvaz, D., Good, V. M., et al. (2004). Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell* 116, 855–867. doi: 10.1016/S0092-8674(04)00215-6

Wang, J. L., Li, L., Hu, M. B., Wu, B., Fan, W. X., Peng, W., et al. (2019a). *In silico* drug design of inhibitor of nuclear factor kappa B kinase subunit beta inhibitors from 2-acylamino-3-aminothienopyridines based on quantitative structure-activity relationships and molecular docking. *Comput. Biol. Chem.* 78, 297–305. doi: 10.1016/j.compbiolchem.2018.12.021

Wang, Q., Zheng, M., Huang, Z., Liu, X., Zhou, H., Chen, Y., et al. (2012). Toward understanding the molecular basis for chemical allosteric modulator design. *J. Mol. Graph. Model.* 38, 324–333. doi: 10.1016/j.jmgm.2012.07.006

Wang, T., Peng, T., Wen, X., Wang, G., Sun, Y., Liu, S., et al. (2019b). Design, synthesis and preliminary biological evaluation of benzylsulfone coumarin derivatives as anti-cancer agents. *Molecules* 24:4034. doi: 10.3390/molecules24224034

Wang, Y., Tang, C., Yao, S., Lai, H., Li, R., Xu, J., et al. (2019c). Discovery of a novel protein kinase C activator from Croton tiglium for inhibition of non-small cell lung cancer. *Phytomedicine* 65:153100. doi: 10.1016/j.phymed.2019.153100

Warner, S. L., Bashyam, S., Vankayalapati, H., Bearss, D. J., Han, H., Mahadevan, D., et al. (2006). Identification of a lead small-molecule inhibitor of the Aurora kinases using a structure-assisted, fragment-based approach. *Mol. Cancer Ther.* 5, 1764–1773. doi: 10.1158/1535-7163.MCT-05-0524

Warren, G. L., Andrews, C. W., Capelli, A. M., Clarke, B., LaLonde, J., Lambert, M. H., et al. (2006). A critical assessment of docking programs and scoring functions. *J. Med. Chem.* 49, 5912–5931. doi: 10.1021/jm050362n

Webb, B., and Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinformatics* 54, 5.6.1–5.6.37. doi: 10.1002/cpbi.3

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., et al. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* 106, 765–784. doi: 10.1021/ja00315a051

Wendel, H. G., Silva, R. L., Malina, A., Mills, J. R., Zhu, H., Ueda, T., et al. (2007). Dissecting eIF4E action in tumorigenesis. *Genes Dev.* 21, 3232–3237. doi: 10.1101/gad.1604407

Wenthur, C. J., Gentry, P. R., Mathews, T. P., and Lindsley, C. W. (2014). Drugs for allosteric sites on receptors. *Annu. Rev. Pharmacol. Toxicol.* 54, 165–184. doi: 10.1146/annurev-pharmtox-010611-134525

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Wong, M. L., Murphy, J., Harrington, E., Gower, C. M., Jain, R. K., Schirle, M., et al. (2017). Examining the influence of specificity ligands and ATP-competitive ligands on the overall effectiveness of bivalent kinase inhibitors. *Proteome Sci.* 15:17. doi: 10.1186/s12953-017-0125-1

Wu, P., Clausen, M. H., and Nielsen, T. E. (2015a). Allosteric small-molecule kinase inhibitors. *Pharmacol. Ther.* 156, 59–68. doi: 10.1016/j.pharmthera.2015.10.002

Wu, P., Nielsen, T. E., and Clausen, M. H. (2015b). FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol. Sci.* 36, 422–439. doi: 10.1016/j.tips.2015.04.005

Xu, Y., Wang, S., Hu, Q., Gao, S., Ma, X., Zhang, W., et al. (2018). CavityPlus: a web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. *Nucleic Acids Res.* 46, W374–W379. doi: 10.1093/nar/gky380

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. (2008). Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24, i232–i240. doi: 10.1093/bioinformatics/btn162

Yang, B., and Papoian, T. (2012). Tyrosine kinase inhibitor (TKI)-induced cardiotoxicity: approaches to narrow the gaps between preclinical safety evaluation and clinical outcome. *J. Appl. Toxicol.* 32, 945–951. doi: 10.1002/jat.2813

Yang, H., Chennamaneni, L. R., Ho, M. W. T., Ang, S. H., Tan, E. S. W., Jeyaraj, D. A., et al. (2018). Optimization of selective mitogen-activated protein kinase interacting kinases 1 and 2 inhibitors for the treatment of blast crisis leukemia. *J. Med. Chem.* 61, 4348–4369. doi: 10.1021/acs.jmedchem.7b01714

Yang, Y. I., Shao, Q., Zhang, J., Yang, L., and Gao, Y. Q. (2019). Enhanced sampling in molecular dynamics. *J. Chem. Phys.* 151:070902. doi: 10.1063/1.5109531

Yin, D., and MacKerell, A. D. (1998). Combinedab initio/empirical approach for optimization of Lennard-Jones parameters. *J. Comp. Chem.* 19, 334–348. doi: 10.1002/(SICI)1096-987X(199802)19:3<334::AID-JCC7>3.0.CO;2-U

Yu, C., Liu, X., Yang, J., Zhang, M., Jin, H., Ma, X., et al. (2019). Combination of immunotherapy with targeted therapy: theory and practice in metastatic melanoma. *Front. Immunol.* 10:990. doi: 10.3389/fimmu.2019.00990

Yuan, Y., Pei, J., and Lai, L. (2013). Binding site detection and druggability prediction of protein targets for structure-based drug design. *Curr. Pharm. Des.* 19, 2326–2333. doi: 10.2174/1381612811319120019

Yueh, C., Rettenmaier, J., Xia, B., Hall, D. R., Alekseenko, A., Porter, K. A., et al. (2019). Kinase Atlas: druggability analysis of potential allosteric sites in kinases. *J. Med. Chem.* 62, 6512–6524. doi: 10.1021/acs.jmedchem.9b00089

Yun, T., Qin, T., Liu, Y., and Lai, L. (2016). Discovery of non-ATP-competitive inhibitors of polo-like kinase 1. *ChemMedChem* 11, 713–717. doi: 10.1002/cmdc.201600051

Zhan, J. Y., Ma, J., and Zheng, Q. C. (2019). Molecular dynamics investigation on the Asciminib resistance mechanism of I502L and V468F mutations in BCR-ABL. *J. Mol. Graph. Model.* 89, 242–249. doi: 10.1016/j.jmgm.2019.03.018

Zhang, G., and Ren, Y. (2018). Molecular modeling and design studies of purine derivatives as novel CDK2 inhibitors. *Molecules* 23:2924. doi: 10.3390/molecules23112924

Zhang, S., Golbraikh, A., and Tropsha, A. (2006). Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. *J. Med. Chem.* 49, 2713–2724. doi: 10.1021/jm050260x

Zhang, W., Pei, J., and Lai, L. (2017). Statistical analysis and prediction of covalent ligand targeted cysteine residues. *J. Chem. Inf. Model.* 57, 1453–1460. doi: 10.1021/acs.jcim.7b00163

Zhao, B., Zhao, C., Hu, X., Xu, S., Lan, Z., Guo, Y., et al. (2019a). Design, synthesis and 3D-QSAR analysis of novel thiopyranopyrimidine derivatives as potential antitumor agents inhibiting A549 and Hela cancer cells. *Eur. J. Med. Chem.* 185:111809. doi: 10.1016/j.ejmech.2019.111809

Zhao, H., Dong, J., Lafleur, K., Nevado, C., and Caflisch, A. (2012). Discovery of a novel chemotype of tyrosine kinase inhibitors by fragment-based docking and molecular dynamics. *ACS Med. Chem. Lett.* 3, 834–838. doi: 10.1021/ml3001984

Zhao, Y., Liu, F., He, G., Li, K., Zhu, C., Yu, W., et al. (2019b). Discovery of arylamide-5-anilinoquinazoline-8-nitro derivatives as VEGFR-2 kinase inhibitors: synthesis, *in vitro* biological evaluation and molecular docking. *Bioorg. Med. Chem. Lett.* 29:126711. doi: 10.1016/j.bmcl.2019.126711

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 37, 1038–1040. doi: 10.1038/s41587-019-0224-x

Zhou, N., Xu, Y., Liu, X., Wang, Y., Peng, J., Luo, X., et al. (2015). Combinatorial pharmacophore-based 3D-QSAR analysis and virtual screening of FGFR1 inhibitors. *Int. J. Mol. Sci.* 16, 13407–13426. doi: 10.3390/ijms160613407

Zuccotto, F., Ardini, E., Casale, E., and Angiolini, M. (2010). Through the "gatekeeper door": exploiting the active kinase conformation. *J. Med. Chem.* 53, 2681–2694. doi: 10.1021/jm901443h

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Gagic, Ruzic, Djokovic, Djikic and Nikolic. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 10 January 2020
doi: 10.3389/fchem.2019.00899

Vienna LiverTox Workspace—A Set of Machine Learning Models for Prediction of Interactions Profiles of Small Molecules With Transporters Relevant for Regulatory Agencies

Floriane Montanari†, Bernhard Knasmüller, Stefan Kohlbacher, Christoph Hillisch, Christine Baierová, Melanie Grandits and Gerhard F. Ecker*

Pharmacoinformatics Research Group, Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria

OPEN ACCESS

Edited by:
Jose L. Medina-Franco, National Autonomous University of Mexico, Mexico

Reviewed by:
Vinicius M. Alves, University of North Carolina at Chapel Hill, United States
Rodolpho C. Braga, InsilicAll, Brazil

***Correspondence:**
Melanie Grandits
melanie.grandits@univie.ac.at

†Present address:
Floriane Montanari, Department of Digital Technologies, Bayer AG, Berlin, Germany

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *09 October 2019*
Accepted: *13 December 2019*
Published: *10 January 2020*

Citation:
Montanari F, Knasmüller B, Kohlbacher S, Hillisch C, Baierová C, Grandits M and Ecker GF (2020) Vienna LiverTox Workspace—A Set of Machine Learning Models for Prediction of Interactions Profiles of Small Molecules With Transporters Relevant for Regulatory Agencies. Front. Chem. 7:899. doi: 10.3389/fchem.2019.00899

Transporters expressed in the liver play a major role in drug pharmacokinetics and are a key component of the physiological bile flow. Inhibition of these transporters may lead to drug-drug interactions or even drug-induced liver injury. Therefore, predicting the interaction profile of small molecules with transporters expressed in the liver may help medicinal chemists and toxicologists to prioritize compounds in an early phase of the drug development process. Based on a comprehensive analysis of the data available in the public domain, we developed a set of classification models which allow to predict—for a small molecule—the inhibition of and transport by a set of liver transporters considered to be relevant by FDA, EMA, and the Japanese regulatory agency. The models were validated by cross-validation and external test sets and comprise cross validated balanced accuracies in the range of 0.64–0.88. Finally, models were implemented as an easy to use web-service which is freely available at https://livertox.univie.ac.at.

Keywords: Vienna LiverTox Workspace, web service, machine learning, ABC-transporter, OATP-transporter, toxicity, classification models

INTRODUCTION

Membrane transporters expressed in the liver play different, but interconnected roles: on the one hand, basolateral transporters pick up xenobiotics, and endogenous molecules from the portal vein to the liver, or excrete their substrates into the blood. Apical transporters, on the other hand, take care of the flux toward the bile duct network (**Figure 1**). Three main types of substrates are of interest with respect to liver toxicity: drugs, which enter the hepatocytes at the first hepatic pass or at the elimination stage; bilirubin, a product of the degradation of the heme; and bile salts, which circulate between the gastro-intestinal tract, and the liver.

Additionally, to the enzyme family of cytochromes, also the transporters expressed in the liver are crucial for a fully functional organ. Some of them are e.g., involved in the bilirubin cycle: OATP1B1 and OATP1B3 uptake bilirubin into the hepatocytes (Briz et al., 2003), where glucuronidation takes place. MRP2 then excretes the bilirubin conjugate to the bile (Kamisako et al., 1999). At the basolateral membrane, MRP3 might also excrete it back to the sinusoidal blood (Keppler, 2014). As a result, inhibition of the uptake OATP transporters or of MRP2 may

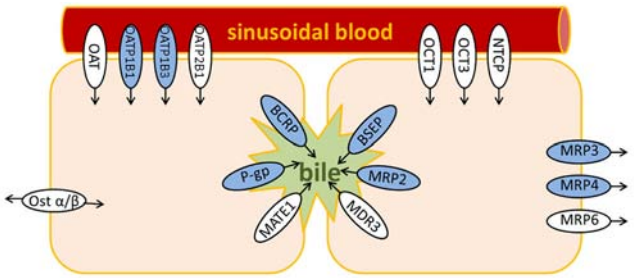


FIGURE 1 | Main hepatic transporters. In blue, transporters for which predictive models (inhibition and/or transport) are available in the Vienna LiverTox Workspace.

lead to the accumulation of bilirubin (conjugated or not) in the blood, which is referred to as hyperbilirubinemia. Conjugated hyperbilirubinemia is a marker of hepatobiliary injury (Dufour et al., 2000; Ozer et al., 2008; Padda et al., 2011), and predicting it may allow to flag compounds that could cause liver injury.

Bile acids are synthesized in the liver by catabolism of cholesterol and then excreted to the bile by the active bile salt export pump (BSEP) and by the multidrug resistance-associated protein 2 (MRP2) (Meier and Stieger, 2002). Bile salts have a pronounced detergent effect, which explains their toxicity when they accumulate in the liver (Attili et al., 1986). For their transport in the bile duct, bile salts form mixed micelles with phospholipids of the outer leaflet of the membrane. The multidrug resistance protein 3 (MDR3) allows lipid flopping at the apical membrane of the hepatocyte, and its function is necessary to avoid bile duct toxicity (Nicolaou et al., 2012). After reaching the intestine via the bile flow, bile acids are reabsorbed into the portal vein, and taken up again into the hepatocytes by the sodium taurocholate co-transporting polypeptide (NTCP) (Stieger, 2011). Impairment of bile flow leading to a toxic accumulation of bile salts in the hepatocytes might lead to drug-induced cholestasis, which is one of the main causes of drug-induced liver injury (DILI) (Padda et al., 2011).

Apart from playing a role in proper bile flow and bilirubin elimination, liver transporters also transport drugs that will then be metabolized and excreted. At this stage, drugs can inhibit different transporters and cause drug-drug interactions (König et al., 2013) (in case of co-administered inhibitor and substrate) or liver injury (by disrupting the bile flow for example). This is why predicting the inhibition and the substrate profile for liver transporters might be useful in identifying potentially problematic compounds. In addition, the Food and Drugs administration recommends experimental testing of the interactions between drugs and transporters (especially P-gp, BCRP, OATP1B1, and OATP1B3) to identify potential drug-drug interactions (U.S. Department of Health Human Services Food Drug Administration Center for Drug Evaluation Research, 2012). Thus, it definitely would be of value to have a suite of computational models available which allow the fast and easy assessment of compounds for their interaction profile with transporters expressed in the liver.

Here, we present the Vienna LiverTox Workspace, a web server for the prediction of interactions with liver transporters as well as selected liver toxicity endpoints. To the best of our knowledge, it is the first time that such an ensemble of predictive models for hepatotoxicity and liver transport is made available to the public. The predictions are made by individual machine learning models built on publicly available data for each target of interest.

Abbreviations: ADME-Tox, absorption distribution metabolism excretion toxicity; BCRP, breast cancer resistance protein; BSEP, bile salt export pump; DILI, drug-induced liver injury; ECFP, extended connectivity fingerprint, MDR1, multidrug resistance protein 1; MDR3, multidrug resistance protein 3; MRP2, multidrug resistance-associated protein 2; MRP3, multidrug resistance-associated protein 3; MRP4, multidrug resistance-associated protein 4; NTCP, sodium taurocholate co-transporting polypeptide; OATP1B1, organic anion transporting polypeptide 1B1; OATP1B3, organic anion transporting polypeptide 1B3; P-gp, P-glycoprotein; SMILES, simplified molecular input line entry system.

METHODS

Data Curation

The datasets for the training and testing of the models were collected from different sources (online tools as well as publications). The data were cleaned by using an in-house system combining the molecular operating environment (MOE 2014.09) (Molecular Operating Environment, 2014) wash option and the Atkinson Standardiser (https://github.com/flatkinson/standardiser). This approach was used for some of the datasets and for others the cleaning procedure of already published papers were used (Pinto et al., 2012; Kotsampasakou et al., 2015). In general, duplicates were removed from the dataset, including pairs of stereoisomers. Further, if these compounds share the same class label, one of the compounds was kept. A detailed list of references as well as the number of compounds revealed after the preprocessing of the data is given in **Table S1**. In the sections Transporter Models and Hepatotoxicity Models, the data curation as well as the model generation for the specific endpoints is given. The datasets are available in the Supplementary Material (**Data Sheet 2**).

Transporter Models

The web service allows for the prediction of interactions between a small molecule and eight different liver transporters (**Figure 1**, transporters marked in blue). The lack of publicly available data for the other transporters explains the absence of respective models in the Vienna LiverTox Workspace.

All the models predicting whether a compound will be a substrate of a transporter (BCRP, P-gp, BSEP, MRP2, and MRP3) were built upon a dataset correlating expression levels of 47 ABC-transporters with drug toxicity, which can then be used to infer transported vs. non-transported compounds (Szakács et al., 2004). For the transport inhibition models (BCRP, P-gp, BSEP, MRP3, MRP4, OATP1B1, and OATP1B3) the datasets were collected from literature and, if necessary, manually aggregated. In both cases, the models predict a binary outcome: the query compound is a substrate or not, or an inhibitor or not.

For chemistry encoding of the compounds, we used circular fingerprints or 2D molecular descriptors as implemented in

TABLE 1 | Overview of the parameters used for the models.

	Descriptors	Trainings set	Algorithm
INHIBITION			
P-gp (MDR1)	ECFP8-like fingerprints	Broccatelli et al., 2011	SVM
BSEP	Molecular descriptors	Warner et al., 2012 Dawson et al., 2012 Morgan et al., 2010	Naïve bayes
BCRP	ECFP8-like fingerprints	Montanari and Ecker, 2014	Logistic regression
MRP3	Molecular descriptors	Köck et al., 2014	BayesNet
MRP4	Molecular descriptors	Köck et al., 2014	AdaBoost (MetaCost)
OATP1B1	Molecular descriptors	De Bruyn et al., 2013	BayesNet
OATP1B3	Molecular descriptors	De Bruyn et al., 2013	BayesNet
TRANSPORT			
P-gp (MDR1)	Molecular descriptors	Szakács et al., 2004	Rotation Forest (MetaCost)
BSEP			SVM (MetaCost)
BCRP			k-nearest neighbors (MetaCost)
MRP2			
MRP3			
TOXICITY			
Hyperbilirubinemia	ECFP8-like fingerprints	Liu et al., 2011	SVM (MetaCost)
Cholestasis	Molecular descriptors	SIDER v2 database (Kuhn et al., 2010, 2016)	Tree model (MetaCost)
Drug-induced liver injury (DILI)	Molecular descriptors	Various sources*	Random Forest

see https://livertox.univie.ac.at/ for a detailed list.

RDKit version 2015.03.1 (https://www.rdkit.org/). Different machine learning algorithms were applied and the one giving the best cross-validation results was kept as final model. Especially for the transport prediction, a heavy class imbalancy (most of the drugs in the training set were non-substrates) was noted, which was handled by MetaCost (Domingos, 1999). The exact methodology and cross-validation performance for each individual transporter model is described in the documentation available at https://livertox.univie.ac.at, and an overview is given in **Table 1**. In some cases, external test sets were collected from Metrabase (Mak et al., 2015) or from recent publications (time-split evaluation).

Hepatotoxicity Models

Three models in the web service can be used to assess human liver damage potentially caused by a test compound: hyperbilirubinemia, cholestasis, and drug-induced liver injury (DILI).

For hyperbilirubinemia, 835 compounds were taken from Kotsampasakou et al. (2017a) and the modeling methodology was kept as in Kotsampasakou et al. (2017b): ECFP-like fingerprints were computed with RDKit (https://www.rdkit.org/), then a combination of feature selection, MetaCost, and support vector machines with RBF kernel was used for learning. The cholestasis model uses data from Kotsampasakou and Ecker (2017) and a combination of MetaCost and a tree algorithm to predict whether a compound is likely to cause cholestasis or not. Finally, the DILI model is based on a 966-compound dataset carefully compiled from literature (Kotsampasakou et al., 2017c). RDKit molecular descriptors and a random forest of 500 trees are used for modeling.

Web Service Implementation

The Vienna LiverTox Workspace has been implemented as Python/php based web service. It consists of two parts, namely the backend and the frontend. The backend consists of a docker image which runs the machine learning models on an input SD-File. It consists of a Python Flask server (https://palletsprojects.com/p/flask/) which processes the requests from the frontend. Each request consists of one or more input molecules and a list of models to run the predictions on. The frontend, also a docker image, is based on the CakePHP framework (https://cakephp.org/) and is responsible for the user interface (UI), which sends the request to the backend and displays the results. The web service provides, after a login, the possibility to upload a SD-File of 10 compounds. The service can also be used without logging in but then it is only possible to draw and predict a single molecule. JSME (Bienfait and Ertl, 2013) is used as drawing tool. The web service runs on an Ubuntu Linux based server with two twelve-core Intel Xeon 64bit processors and 128 GB RAM, and is hosted at the University of Vienna by the Pharmacoinformatics Research Group.

The models use the RDKit library (https://www.rdkit.org/) (version 2015.03.1) for computing the descriptors and handling the chemistry aspects, while the Weka (Hall et al., 2009) (version 3.7.11) and scikit-learn (Pedregosa et al., 2011) (version 0.14.1) libraries are used to train and run the predictive models. The models also include a compound cleaning step, implemented with the Atkinson Standardiser (https://github.com/flatkinson/standardiser) (**Figure 2**).

The steps performed during model building and a test compound prediction are shown in **Figure 2**. In both cases, the compound is standardized and the molecular descriptors are

calculated. In the case of the model generation, this allows the training of the model and its development (left workflow). For the prediction of a test compound, the descriptors are passed to the available model to predict its class affiliation (right workflow).

In general, the output of the model gives, in addition to the class prediction, the actual score. This score is a numerical value between 0 and 1, and roughly corresponds to a probability of being active (inhibitor, substrate, or toxic compound). Therefore, a value close to 1 indicates substrate/inhibitor/toxic properties, a value close to 0 annotates for non-substrates/non-inhibitors/non-toxic.

Applicability Domain

The Applicability Domain (AD) is used to validate the reliability of a given prediction model. It defines whether a dataset of interest is in or out of domain, meaning if it falls within the chemical space of the model or not. If it is out of Domain, the prediction cannot be regarded as reliable.

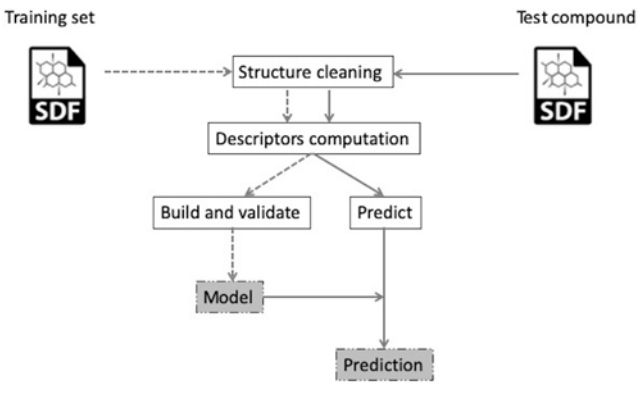


FIGURE 2 | General steps to build a model (left workflow, dotted line) or predict a property for a test compound (right workflow, solid line). Figure adapted from Carrió et al. (2015) published in Journal of Cheminformatics is licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

In our study, an Applicability Domain model, using the approach of Sahigara et al. (2013), was created for each transporter with the respective training set. RDKit descriptors were used as molecule representation. This approach combines the classical, widely used k-Nearest Neighbor (k-NN) method with a probability density function estimation. It uses three stages to determine the reliable space of a prediction model. First, a set of thresholds is defined depending on the diverse densities of the training set by considering the 15 k-nearest neighbors using Euclidean distances. This allows the AD to consider a dense and sparse training region (The threshold defines if a test sample can be reliable predicted). In a next step a decision rule is derived to filter out outlier molecules. Finally, the reliability of the AD is tested by looking at the model statistics and prediction errors. This feature is not yet implemented in the Web service, but will be available soon.

RESULTS

Model Results

The performance of the models was estimated by calculating statistical performance metrics using a 10-fold cross-validation. The results are provided in **Table 2**. The overall accuracy, corresponding to the rate of correct predictions, ranges from 0.59 to 0.87. Also, the sensitivity of the models was calculated (0.57–0.85). This parameter gives the number of actual positives that are correctly identified and is expressed by the number of true positives divided by the number of positive predictions. Further, the number of actual negatives was determined by the

TABLE 2 | Performance metrics of the transporter models.

	Accuracy	Sensitivity	Specificity	Balanced accuracy	MCC	ROC AUC
INHIBITION						
P-gp (MDR1)	0.87	0.85	0.90	0.88	0.76	0.94
BSEP	0.85	0.77	0.87	0.82	0.60	0.91
BCRP	0.83	0.77	0.87	0.82	0.64	0.90
MRP3	0.82	0.75	0.87	0.81	0.62	0.86
MRP4	0.73	0.86	0.65	0.76	0.49	0.74
OATP1B1	0.77	0.71	0.78	0.75	0.34	0.80
OATP1B3	0.81	0.78	0.81	0.80	0.35	0.81
TRANSPORT						
P-gp (MDR1)	0.81	0.81	0.81	0.81	0.44	0.85
BSEP	0.71	0.65	0.72	0.69	0.28	0.69
BCRP	0.73	0.57	0.75	0.66	0.20	0.71
MRP2	0.70	0.60	0.72	0.66	0.27	0.73
MRP3	0.72	0.60	0.74	0.67	0.28	0.72
TOXICITY						
Hyperbilirubinemia	0.67	0.64	0.67	0.66	0.20	0.69
Cholestasis	0.59	0.72	0.56	0.64	0.22	0.64
Drug-induced liver injury (DILI)	0.65	0.72	0.58	0.65	0.30	0.70

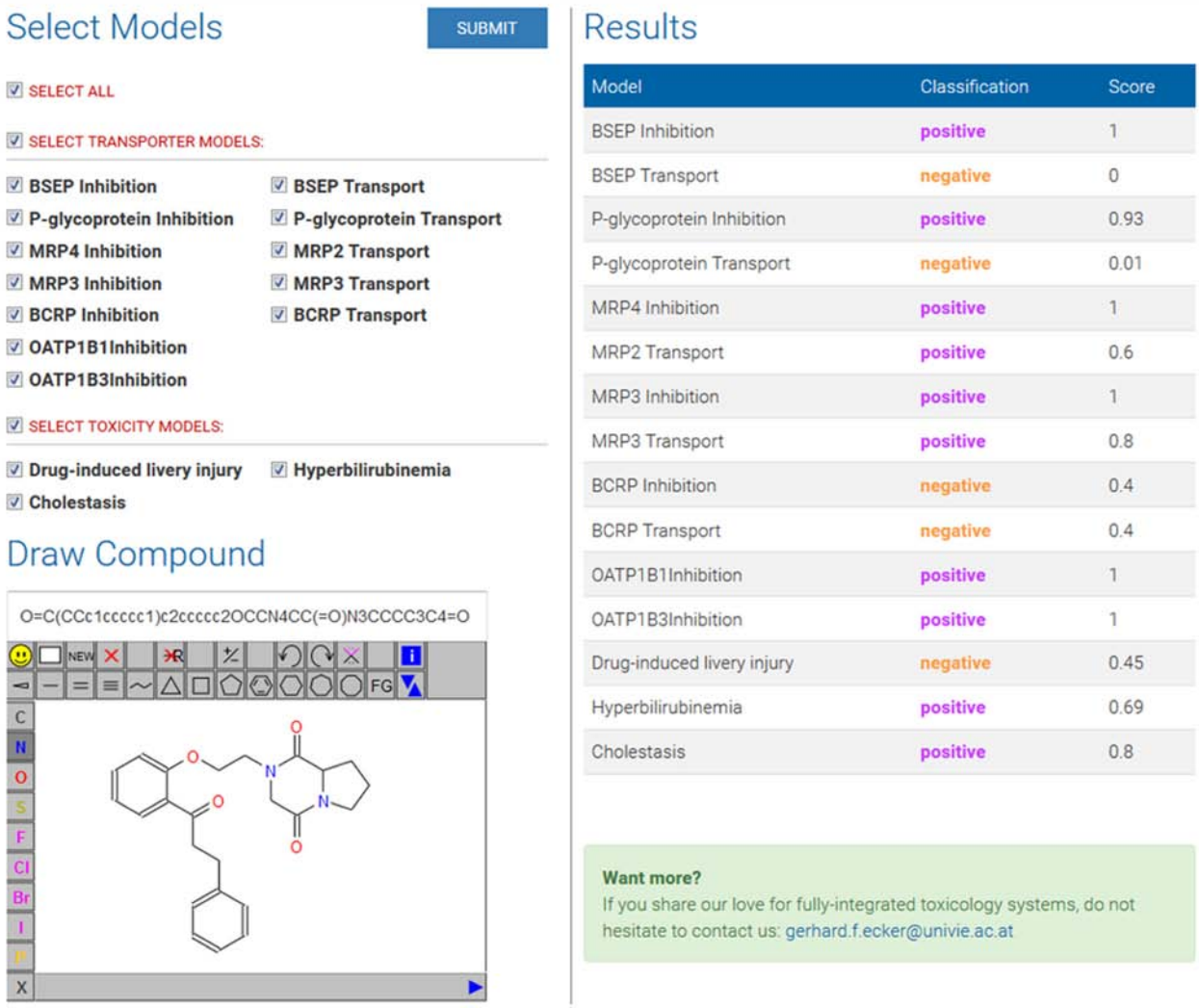


FIGURE 3 | Overview of the web service interface. On the left side, test compound is drawn and desired models are selected. On the right side, results table with the predictions and model scores.

number of true negatives divided by the number of negative predictions. The so-called specificity ranges from 0.56 to 0.90. To estimate a metric for the quality of the models, the Matthews correlation coefficient (MCC) and the Area under the Receiver Operating Characteristics curve (ROC AUC) were determined. The MCC is a number between −1 and 1 where 0 indicates a prediction equal to a random prediction and 1 indicates a perfect prediction, whereas −1 is a complete miss. The scores for our models are in the range of 0.20–0.76. The ROC AUC measures the ability of the model to distinguish between negatives and positives, while a higher value indicates a better performance. In the models provided on the Vienna LiverTox Workspace the ROC AUC ranges from 0.64 to 0.94. Furthermore, if data was available, the models were also validated with one or more external test sets. For more details see the documentation on the website.

Use Case: Prediction of Liver Interaction for a Propafenone Analog

In this section, we briefly detail how predictions can be generated for a given compound using the LiverTox web service. In first instance, the compound is drawn or its SMILES string is pasted in the Molecule Editor (Bienfait and Ertl, 2013). Then the models can be selected on the left panel, either one by one or all at the same time (**Figure 3**, left side).

By clicking on the SUBMIT button, the data is sent to the backend server where the predictions are running. Upon completion of the calculations, a table listing the different models and corresponding outputs will appear (**Figure 3**, right side). The second column in the results table corresponds to the binary classification, while the third column "Score" gives the actual output of the model, which corresponds to a probability of being active. For example, for BCRP inhibition and transport

TABLE 3 | Comparison of existing free online tools to predict ADME-Tox properties of compounds.

Web service	Transporters predictions	CYP450 predictions	Hepatotox. predictions	Batch prediction	Run time for 1 compound
ProTox-II (Drwal et al., 2014)	No	No	No	Yes (max. 100)	<5 s
BioZyne (Levatić et al., 2013)	P-gp	No	No	Not for free	∼5 s
QSAR DB (http://qsar.food.dtu.dk/)	No	Yes	No	Yes	N.A.
pkCSM (Pires et al., 2015)	P-gp	Yes	Yes	Yes (max. 100)	<5 s for 30 models
Lazar (Maunz et al., 2013)	No	No	No	No	∼10 s for 6 models
Vienna LiverTox Workspace	P-gp, BSEP, BCRP, MRP2, MRP3, MRP4, OATP1B1, OATP1B3	No	Yes	Not for free	∼30 s for 15 models

and for DILI, the output score is close to 0.5 (which is the threshold used to separate predicted actives from predicted inactives), which indicates an uncertainty of these three models for the particular query compound. Propafenone derivatives are frequently reported as inhibitors of P-gp, and indeed the P-gp inhibition model predicts this particular one to be an inhibitor with a high score (0.93).

DISCUSSION

Many systems already exist to predict *in silico* activities or properties of small molecules. **Table 3** compares freely available ones with our own web service in terms of model offer, submission and run time. For example, ProTox-II predicts oral drug toxicity in rodents (lethal dose LD_{50} and a category of toxicity between 1 and 6) using similarity to compounds with known LD_{50} and recognition of toxic fragments (Drwal et al., 2014). BioZyne proposes exclusively one model for P-gp transport prediction based on the same dataset as ours (Szakács et al., 2004; Levatić et al., 2013). It uses a Support Vector Machine classifier for the prediction of P-gp substrates. The Danish (Q)SAR Database contains pre-calculated properties combined from more than 200 models from both commercial and free tools (http://qsar.food.dtu.dk/). Predictions for environmental toxicity, blood-brain barrier permeation, cytochrome interactions, or human genotoxicity are available. Unfortunately, new predictions for compounds that are not part of the database cannot be made. PkCSM is another web service for predicting pharmacokinetics properties of compounds (Pires et al., 2015). Models such as P-gp inhibition and transport, blood-brain barrier permeation, interaction with cytochromes, renal clearance, or even liver toxicity are available.

In general, our models for the inhibitors show a better performance especially when looking at the correct prediction of the positives. The prediction of true negatives is for the inhibitor and transporter models quite similar which can be explained by the availability of more negatives if the training set is unbalanced. This is especially the case for the substrate models. The quality of the prediction (MCC) is higher for the inhibition models of P-gp, BSEP, BCRP, and MRP3 since the available dataset is more balanced. In comparison, the three toxicity models show a poorer performance due to the complexity of these endpoints and especially for hyperbilirubinemia and cholestasis which shows also a lack of positives.

The Transporters selected for this web service were chosen based on their importance for regulatory agencies such as FDA, EMA and the Japanese regulatory agency. They recommend or in some cases request these proteins to be routinely tested in inhibition—and substrate studies of new drugs.

CONCLUSION

We have presented the Vienna LiverTox Workspace, a web service dedicated to the prediction of liver toxicity and interactions between small molecules and liver transporters. It is easy to use, fast, web browser agnostic, and well-documented. Thanks to its modular system, it will be easy to integrate new models in the future, as well as re-implement existing models in case new training data becomes available. We hope that our models will help researchers to flag potentially dangerous compounds and shed light on the relationships between liver transporters and toxicity.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

FM, SK, CH, and MG developed the models. BK, CB, and MG implemented the web service. MG and GE supervised the study. FM and MG wrote the majority of the manuscript. All authors contributed to refining the manuscript.

FUNDING

We gratefully acknowledge financial support provided by the Austrian Science Fund, grant #F03502 (SFB35), and by the Innovative Medicines Initiative Joint Undertaking under grant agreement n°115002 (eTOX).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00899/full#supplementary-material

REFERENCES

Attili, A. F., Angelico, M., Cantafora, A., Alvaro, D., and Capocaccia, L. (1986). Bile acid-induced liver toxicity: relation to the hydrophobic-hydrophilic balance of bile acids. *Med. Hypotheses* 19, 57–69. doi: 10.1016/0306-9877(86)90137-4

Bienfait, B., and Ertl, P. (2013). JSME: a free molecule editor in JavaScript. *J. Cheminform.* 5:24. doi: 10.1186/1758-2946-5-24

Briz, O., Serrano, M. A., MacIas, R. I. R., Gonzalez-Gallego, J., and Marin, J. J. G. (2003). Role of organic anion-transporting polypeptides, OATP-A, OATP-C and OATP-8, in the human placenta-maternal liver tandem excretory pathway for foetal bilirubin. *Biochem. J.* 371, 897–905. doi: 10.1042/bj20030034

Broccatelli, F., Carosati, E., Neri, A., Frosini, M., Goracci, L., Oprea, T. I., et al. (2011). A novel approach for predicting P-glycoprotein (ABCB1) inhibition using molecular interaction fields. *J. Med. Chem.* 54, 1740–1751. doi: 10.1021/jm101421d

Carrió, P., López, O., Sanz, F., and Pastor, M. (2015). eTOXlab, an open source modeling framework for implementing predictive models in production environments. *J. Cheminform.* 7:8. doi: 10.1186/s13321-015-0058-6

Dawson, S., Stahl, S., Paul, N., Barber, J., and Kenna, J. G. (2012). *In vitro* inhibition of the bile salt export pump correlates with risk of cholestatic drug-induced liver injury in humans. *Drug Metab. Dispos.* 40, 130–8. doi: 10.1124/dmd.111.040758

De Bruyn, T., van Westen, G. J. P., IJzerman, A. P., Stieger, B., de Witte, P., Augustijns, P. F., et al. (2013). Structure-based identification of OATP1B1/3 inhibitors. *Mol. Pharmacol.* 83, 1257–1267. doi: 10.1124/mol.112.084152

Domingos, P. (1999). "MetaCost: a general method for making classifiers cost-sensitive," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York), 155–164. doi: 10.1145/312129.312220

Drwal, M. N., Banerjee, P., Dunkel, M., Wettig, M. R., and Preissner, R. (2014). ProTox: a web server for the *in silico* prediction of rodent oral toxicity. *Nucleic Acids Res.* 42, W53–W58. doi: 10.1093/nar/gku401

Dufour, D. R., Lott, J. A., Nolte, F. S., Gretch, D. R., Koff, R. S., and Seeff, L. B. (2000). Diagnosis and monitoring of hepatic injury. II. Recommendations for use of laboratory tests in screening, diagnosis, and monitoring. *Clin. Chem.* 46, 2050–68.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* 11, 10–18. doi: 10.1145/1656274.1656278

Kamisako, T., Leier, I., Cui, Y., König, J., Buchholz, U., Hummel-Eisenbeiss, J., et al. (1999). Transport of monoglucuronosyl and bisglucuronosyl bilirubin by recombinant human and rat multidrug resistance protein 2. *Hepatology* 30, 485–490. doi: 10.1002/hep.510300220

Keppler, D. (2014). The roles of MRP2, MRP3, OATP1B1, and OATP1B3 in conjugated hyperbilirubinemia. *Drug Metab. Dispos.* 42, 561–565. doi: 10.1124/dmd.113.055772

Köck, K., Ferslew, B. C., Netterberg, I., Yang, K., Urban, T. J., Swaan, P. W., et al. (2014). Risk factors for development of cholestatic drug-induced liver injury: inhibition of hepatic basolateral bile acid transporters multidrug resistance-associated proteins 3 and 4. *Drug Metab. Dispos.* 42, 665–74. doi: 10.1124/dmd.113.054304

König, J., Müller, F., and Fromm, M. F. (2013). Transporters and drug-drug interactions: important determinants of drug disposition and effects. *Pharmacol. Rev.* 65, 944–966. doi: 10.1124/pr.113.007518

Kotsampasakou, E., Brenner, S., Jäger, W., and Ecker, G. F. (2015). Identification of novel inhibitors of organic anion transporting polypeptides 1B1 and 1B3 (OATP1B1 and OATP1B3) using a consensus vote of six classification models. *Mol. Pharm.* 12, 4395–4404. doi: 10.1021/acs.molpharmaceut.5b00583

Kotsampasakou, E., and Ecker, G. F. (2017). Predicting drug-induced cholestasis with the help of hepatic transporters—an *in silico* modeling approach. *J. Chem. Inf. Model.* 57, 608–615. doi: 10.1021/acs.jcim.6b00518

Kotsampasakou, E., Escher, S. E., and Ecker, G. F. (2017a). Curated human hyperbilirubinemia data and the respective OATP1B1 and 1B3 inhibition predictions. *Data Br.* 11, 204–207. doi: 10.1016/j.dib.2017.02.009

Kotsampasakou, E., Escher, S. E., and Ecker, G. F. (2017b). Linking organic anion transporting polypeptide 1B1 and 1B3 (OATP1B1 and OATP1B3) interaction profiles to hepatotoxicity—The hyperbilirubinemia use case. *Eur. J. Pharm. Sci.* 100, 9–16. doi: 10.1016/j.ejps.2017.01.002

Kotsampasakou, E., Montanari, F., and Ecker, G. F. (2017c). Predicting drug-induced liver injury: the importance of data curation. *Toxicology* 389, 139–145. doi: 10.1016/j.tox.2017.06.003

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., and Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* 6:343. doi: 10.1038/msb.2009.98

Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic Acids Res.* 44, D1075–D1079. doi: 10.1093/nar/gkv1075

Levatić, J., Curak, J., Kralj, M., Šmuc, T., Osmak, M., and Supek, F. (2013). Accurate models for P-gp drug recognition induced from a cancer cell line cytotoxicity screen. *J. Med. Chem.* 56, 5691–5708. doi: 10.1021/jm400328s

Liu, Z., Shi, Q., Ding, D., Kelly, R., Fang, H., and Tong, W. (2011). Translating clinical findings into knowledge in drug safety evaluation - drug induced liver injury prediction system (DILIps). *PLoS Comput. Biol.* 7:e1002310. doi: 10.1371/journal.pcbi.1002310

Mak, L., Marcus, D., Howlett, A., Yarova, G., Duchateau, G., Klaffke, W., et al. (2015). Metrabase: a cheminformatics and bioinformatics database for small molecule transporter data analysis and (Q)SAR modeling. *J. Cheminform.* 7:31. doi: 10.1186/s13321-015-0083-5

Maunz, A., Gütlein, M., Rautenberg, M., Vorgrimmler, D., Gebele, D., and Helma, C. (2013). Lazar: a modular predictive toxicology framework. *Front. Pharmacol.* 4:38. doi: 10.3389/fphar.2013.00038

Meier, P. J., and Stieger, B. (2002). Bile salt transporters. *Annu. Rev. Physiol.* 64, 635–661. doi: 10.1146/annurev.physiol.64.082201.100300

Molecular Operating Environment (2014). *Molecular Operating Environment.* Available at: https://www.chemcomp.com/

Montanari, F., and Ecker, G. F. (2014). BCRP inhibition: from data collection to ligand-based modeling. *Mol. Inform.* 33, 322–331. doi: 10.1002/minf.201400012

Morgan, R. E., Trauner, M., van Staden, C. J., Lee, P. H., Ramachandran, B., Eschenberg, M., et al. (2010). Interference with bile salt export pump function is a susceptibility factor for human liver injury in drug development. *Toxicol. Sci.* 118, 485–500. doi: 10.1093/toxsci/kfq269

Nicolaou, M., Andress, E. J., Zolnerciks, J. K., Dixon, P. H., Williamson, C., and Linton, K. J. (2012). Canalicular ABC transporters and liver disease. *J. Pathol.* 226, 300–315. doi: 10.1002/path.3019

Ozer, J., Ratner, M., Shaw, M., Bailey, W., and Schomaker, S. (2008). The current state of serum biomarkers of hepatotoxicity. *Toxicology* 245, 194–205. doi: 10.1016/j.tox.2007.11.021

Padda, M. S., Sanchez, M., Akhtar, A. J., and Boyer, J. L. (2011). Drug-induced cholestasis. *Hepatology* 53, 1377–1387. doi: 10.1002/hep.24229

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Pinto, M., Trauner, M., and Ecker, G. F. (2012). An *in silico* classification model for putative ABCC2 substrates. *Mol. Inform.* 31, 547–553. doi: 10.1002/minf.201200049

Pires, D. E. V., Blundell, T. L., and Ascher, D. B. (2015). pkCSM: predicting small-molecule Pharmacokinetic And Toxicity Properties Using Graph-Based Signatures. *J. Med. Chem.* 58, 4066–4072. doi: 10.1021/acs.jmedchem.5b00104

Sahigara, F., Ballabio, D., Todeschini, R., and Consonni, V. (2013). Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. *J. Cheminform.* 5:27. doi: 10.1186/1758-2946-5-27

Stieger, B. (2011). "The role of the sodium-taurocholate cotransporting polypeptide (NTCP) and of the bile salt export pump (BSEP) in physiology and pathophysiology of bile formation," in *Handbook of Experimental Pharmacology* (Berlin; Heidelberg: Springer), 205–259. doi: 10.1007/978-3-642-14541-4_5

Szakács, G., Annereau, J.-P., Lababidi, S., Shankavaram, U., Arciello, A., Bussey, K. J., et al. (2004). Predicting drug sensitivity and resistance:

profiling ABC transporter genes in cancer cells. *Cancer Cell* 6, 129–37. doi: 10.1016/j.ccr.2004.06.026

U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (2012). *Guidance for Industry. Drug Interaction Studies Study Design, Data Analysis, Implications for Dosing, and Labeling Recommendations.* Rockville, MD.

Warner, D. J., Chen, H., Cantin, L.-D., Kenna, J. G., Stahl, S., Walker, C. L., et al. (2012). Mitigating the inhibition of human bile salt export pump by drugs: opportunities provided by physicochemical property modulation, *in silico* modeling, and structural modification. *Drug Metab. Dispos.* 40, 2332–41. doi: 10.1124/dmd.112.047068

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Montanari, Knasmüller, Kohlbacher, Hillisch, Baierová, Grandits and Ecker. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 04 February 2020
doi: 10.3389/fchem.2020.00046

ChEMBL-Likeness Score and Database GDBChEMBL

*Sven Bühlmann and Jean-Louis Reymond**

Department of Chemistry and Biochemistry, University of Bern, Bern, Switzerland

The generated database GDB17 enumerates 166.4 billion molecules up to 17 atoms of C, N, O, S and halogens following simple rules of chemical stability and synthetic feasibility. However, most molecules in GDB17 are too complex to be considered for chemical synthesis. To address this limitation, we report GDBChEMBL as a subset of GDB17 featuring 10 million molecules selected according to a ChEMBL-likeness score (CLscore) calculated from the frequency of occurrence of circular substructures in ChEMBL, followed by uniform sampling across molecular size, stereocenters and heteroatoms. Compared to the previously reported subsets FDB17 and GDBMedChem selected from GDB17 by fragment-likeness, respectively, medicinal chemistry criteria, our new subset features molecules with higher synthetic accessibility and possibly bioactivity yet retains a broad and continuous coverage of chemical space typical of the entire GDB17. GDBChEMBL is accessible at http://gdb.unibe.ch for download and for browsing using an interactive chemical space map at http://faerun.gdb.tools.

Keywords: chemical space exploration, molecular database, enumeration algorithm, chemical space mapping, virtual screening

OPEN ACCESS

Edited by:
Jose L. Medina-Franco, National Autonomous University of Mexico, Mexico

Reviewed by:
Cristian G. Bologa, University of New Mexico, United States
Rodrigo Ochoa, University of Antioquia, Colombia

****Correspondence:***
Jean-Louis Reymond
jean-louis.reymond@dcb.unibe.ch

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *25 October 2019*
Accepted: *15 January 2020*
Published: *04 February 2020*

Citation:
Bühlmann S and Reymond J-L (2020) ChEMBL-Likeness Score and Database GDBChEMBL. Front. Chem. 8:46. doi: 10.3389/fchem.2020.00046

INTRODUCTION

Innovation at the level of chemical structures is an essential part of drug discovery. Novelty often results from chemical intuition however this approach is increasingly difficult as the number of known molecules increases. Novelty is similarly limited in virtual combinatorial libraries (Leach and Hann, 2000; Hu et al., 2011; van Hilten et al., 2019) and generative models trained with known molecules (Chen et al., 2018; Elton et al., 2019) because these systems mostly shuffle known patterns, which produces many technically new but often not fundamentally innovative molecules. To circumvent this limitation, we have initiated the exhaustive enumeration of all possible organic molecules following simple rules of chemical stability and synthetic feasibility, and reported large databases enumerating molecules up to 11 (Fink et al., 2005; Fink and Reymond, 2007), 13 (Blum and Reymond, 2009), and 17 atoms (Ruddigkeit et al., 2012, 2013), as well as of possible ring systems up to 30 atoms (Visini et al., 2017a). Analyzing the resulting generated databases (GDBs) shows that there are many orders of magnitude more possible molecules spanning a much broader structural diversity than already known ones (Reymond, 2015; Awale et al., 2017b).

One of the defining features of the GDB databases is the exponential increase in the number of possible molecules as function of increasing molecular size and complexity elements, such as stereocenters and heteroatoms, implying that most possible molecules are in fact far too complex to be considered as realistic synthetic targets. To address this problem we have designed subsets of our largest database GDB17 by limiting complexity elements using simplification criteria, such as fragment-likeness (Congreve et al., 2003), producing the fragment database FDB17, and

medicinal chemistry rules for functional groups and complexity (Mignani et al., 2018), producing the medicinal chemistry aware database GDBMedChem (Visini et al., 2017b; Awale et al., 2019). These approaches however also constrain the diversity of GDB molecules, which partly defeats the purpose of exploring chemical space broadly.

Herein we report an alternative approach to create subsets of GDB17 based on the frequency of occurrence of substructures from known molecules independent of the overall molecular structure (**Figure 1A**). We define a "ChEMBL-likeness" score (CLscore) by considering which substructures in a molecule also occur in molecules from the public database ChEMBL (Gaulton et al., 2017), using a subset of molecules with reported high confidence datapoint of activity on single protein targets, a type of ChEMBL subset which we have used previously for target prediction (Awale and Reymond, 2019; Poirier et al., 2019). We then filter the entire GDB17 with a cut-off value for CLscore, followed by uniform sampling of the resulting subset across molecular size, stereocenters and heteroatoms as done previously with FDB17 and GDBMedChem, to obtain a ChEMBL-like subset of 10 million molecules forming the database GDBChEMBL. This database covers chemical space as broadly as but more continuously than FDB17 and GDBMedChem yet features a much higher synthetic accessibility as judged by a calculated synthetic accessibility score (Ertl and Schuffenhauer, 2009), might contain molecules with a higher probability of bioactivity, and in any case provides a very different starting point to serve as a source of inspiration for molecular design.

RESULTS AND DISCUSSION

ChEMBL-Likeness Score

Our definition of CLscore is related to the synthetic accessibility score (SAscore) (Ertl and Schuffenhauer, 2009) and natural product likeness score (NPscore) (Jayaseelan et al., 2012) of a molecule, which are calculated from the occurrence frequencies of its substructures in PubChem and fragments from natural products, respectively, combined with additional functional group rules. Here we focus on 457,139 compounds recorded in ChEMBL24 as being active on single protein targets (IC_{50} or $EC_{50} \leq 10\,\mu M$) with high confidence datapoints (Awale and Reymond, 2019; Poirier et al., 2019). To design our CLscore we consider circular substructures, called molecular shingles, because they form the basis for molecular fingerprints ECFP4 and MHFP6 which perform best in benchmarking studies (Riniker and Landrum, 2013; Probst and Reymond, 2018).

The frequency of occurrence of the 636,979 molecular shingles up to a diameter of six bonds found in our ChEMBL subset follows a power law distribution (**Figure 1C**). To compute the CLscore of a molecule, we assign to each of its shingles (S) a shingle value calculated from the logarithm of its frequency of occurrence f_S in our ChEMBL subset, considering only shingles occurring at least 100 times in this subset (141,261 shingles, 22.2% of the total). We then sum all shingle values and divide the sum by the total number of shingles in the molecule (Equation 1).

$$\mathrm{CLscore} = \frac{\sum_{i=1}^{m} \log_{10}(f_S)_i}{N}$$

$$\begin{aligned} &\mathrm{CLscore} := \mathrm{ChEMBL-likeness\ score} \\ &S := \mathrm{shingle in molecular\ structure} \\ &f_S := \mathrm{abundance of molecular shinglein\ ChEMBL} \\ &N := \mathrm{total number of shingles in molecular\ structure} \\ &m := \mathrm{number of shingles shared with\ ChEMBL} \end{aligned}$$

The histogram of CLscore for the 457,139 ChEMBL reference molecules is approximately Gaussian with a peak at CLscore = 3.9 (**Figure 1B**). DrugBank (Law et al., 2014) and particularly ZINC (Sterling and Irwin, 2015) peak at a similar CLscore, showing that these three databases consist of molecules built from the same type of substructures. By contrast GDB17 and its subsets FDB17 and GDBMedChem have a much lower CLscore distribution peaking at CLscore = 2.7, reflecting the fact that GDB molecules are very different from ChEMBL molecules. CLscore values correlate with SAscore values, reflecting the similar principles underlying both scores, and suggesting that molecules with high CLscore should also be synthetically accessible (**Figure 1D**).

GDBChEMBL Database

Calculating CLscores on the entire GDB17 (166.4 billion SMILES) and keeping molecules with CLscore $\geq$ 3.3, a cut-off value which retains 78.3% of our ChEMBL subset, eliminates 84.3% of GDB17. The remaining 26.2 billion molecules are then binned in triplet value bins considering heavy atom count (HAC 1-17), stereocenter count (0–4, $\geq$ 5) and heteroatom count (0–8, $\geq$ 8). There are 538 different triplet value bins, which are occupied by 1 to 1.6×10^9 molecules. Uniform sampling finally yields a final set of 10 million molecules evenly distributed across molecular size, stereochemical complexity and polarity, forming the database GDBChEMBL (**Figure 1E**).

As a consequence of uniform sampling, the heavy atom count (HAC) profile of GDBChEMBL resembles that of FDB17 and GDBMedChem and is relatively flat compared to the very steep peak at HAC = 17 in the parent database GDB17 (**Figure 2A**). Uniform sampling also explains the rotatable bond count (RBC) profile in GDB subsets compared to GDB17 (**Figure 2B**), as well as the fact that the profiles of the three GDB subsets across these parameters are generally more similar to the profile of molecules up to 17 atoms in ChEMBL (ChEMBL17) and to natural products (UNPD17) (Banerjee et al., 2015) than to the profile of GDB17.

GDBChEMBL displays a very broad distribution in terms of hydrogen bond donor atoms (HBD, **Figure 2C**), hydrogen bond acceptor atoms (HBA, **Figure 2D**) and nitrogen plus oxygen atom count (N+O, **Figure 2E**) due to the absence of heteroatom capping criteria in selecting GDBChEMBL compared to FDB17 and GDBMedChem, for which fragment-likeness criteria, respectively, caps on the number of functional groups were applied. Similar differences are visible in topological polar surface area (TPSA, **Figure 2F**) and calculated octanol/water partition coefficient (alogP, **Figure 2G**). The broader distribution of polarity parameters in GDGChEMBL compared to GDB17 results from uniform sampling since the procedure gives relatively more importance to molecules with extreme size and polarity values.

Synthetic accessibility is better (lower SAscore) in GDBChEMBL than for GDB17, FDB17, or GDBMedChem,

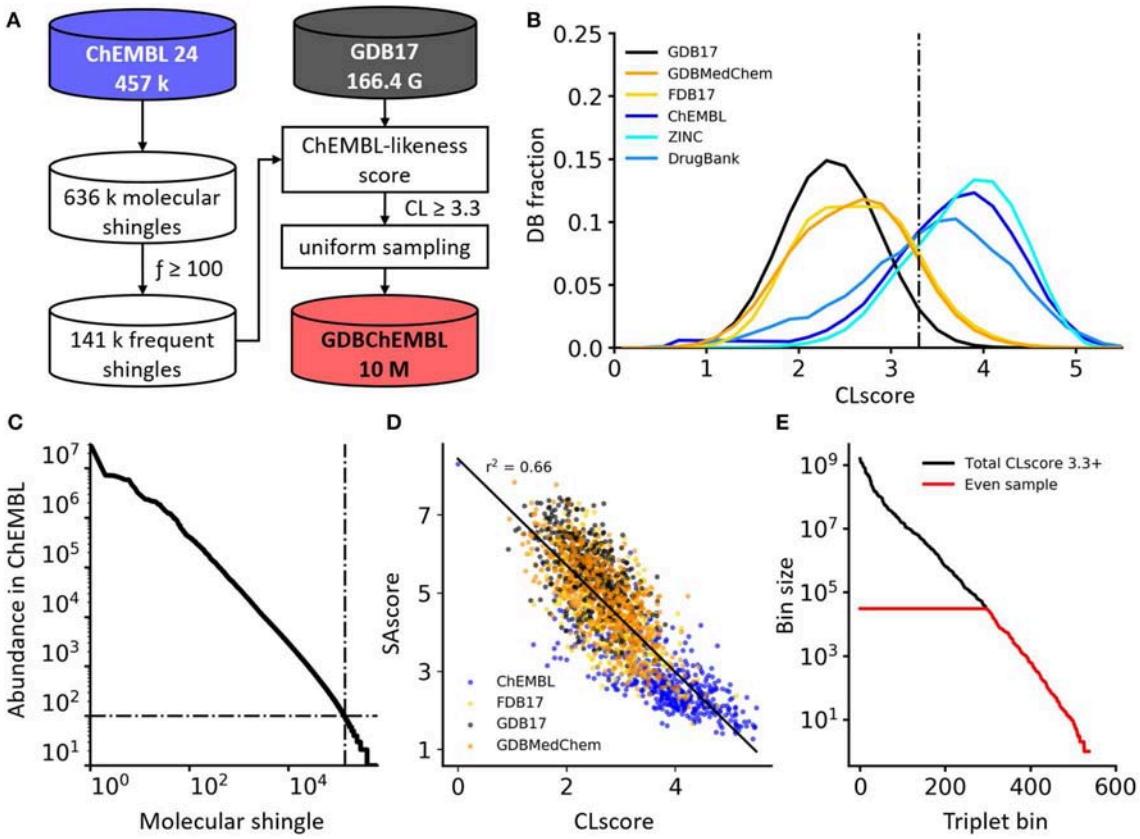


FIGURE 1 | (A) Generation process of GDBChEMBL. **(B)** CLscore distributions for GDB17, its subsets FDB17 and GDBMedChem, and public databases ChEMBL, ZINC, and DrugBank. **(C)** Frequency distribution of molecular shingles up to a diameter of 6 bonds in ChEMBL. **(D)** SAscore vs. CLscore in various databases. A lower SAscore indicates higher synthetic accessibility, and a higher CLscore indicates higher similarity to ChEMBL molecules. **(E)** Occupancy of triplet value bins (HAC, stereocenters, heteroatoms) in all GDB17 cpds with CLscore ≥3.3 (black line) and after uniform sampling forming GDBChEMBL (red line).

reflecting the correlation between CLscore and SAscore noted above (**Figure 2H**). Similar to GDB17 and its other subsets, GDBChEMBL displays a much higher fraction of sp^3 atoms than ChEMBL (fsp^3, **Figure 2I**). As a consequence GDB molecules are closer to natural products, which is reflected in the NPscore profile (**Figure 2J**). Despite of these differences and similarities in SAscore and NPscore, it must be noted that GDB17 and its subsets stand out by the fact that they contain fewer aromatic and more heterocyclic molecules than ChEMBL and natural products (**Figure 2K**).

Visualization and Similarity Searching

To gain an overview of GDBChEMBL we computed Molecular Quantum Number (MQN) fingerprint values (Nguyen et al., 2009), performed a principal component analysis (Rosén et al., 2009), and visualized the resulting 3D-map in the interactive web-based application faerun (Probst et al., 2018). In this 3D-map accessible at http://faerun.gdb.tools, each point represents one or more molecules present at the corresponding position and can be color-coded according to a molecular property selected from the faerun menu.

Comparing MQN maps of GDBChEMBL, FDB17 and GDBMedChem shows that each of the three GDB17 subset cover a similar range of properties, however coverage by GDBChEMBL is more continuous, as is well visible in the vertical stripe at right containing all acyclic molecules (**Figures 3A–C**). Note that CLscore values are not correlated with MQN properties, which is not surprising considering that ChEMBL substructure span a broad range of properties (**Figure 3D**). Color-coding by the calculated logP value (alogP, **Figure 3E**) and by rotatable bond count (RBC, **Figure 3F**) illustrate the distribution of molecules in the MQN map.

The fact that molecules in GDBChEMBL are substantially different from those in the other subsets FDB17 and GDBMedChem can be shown by retrieving 1,000 MQN-nearest neighbors of nicotine from each database, and representing each dataset in a similarity map (Medina-Franco et al., 2007; Raghavendra and Maggiora, 2007; Awale and Reymond, 2015) using the molecular shape and pharmacophore fingerprint Xfp (Awale and Reymond, 2014), computed with the web-based application WebMolCS (Awale et al., 2017a). This visualization shows that each database provides different types of nicotine analogs (**Figure 3G**) with a good number of high

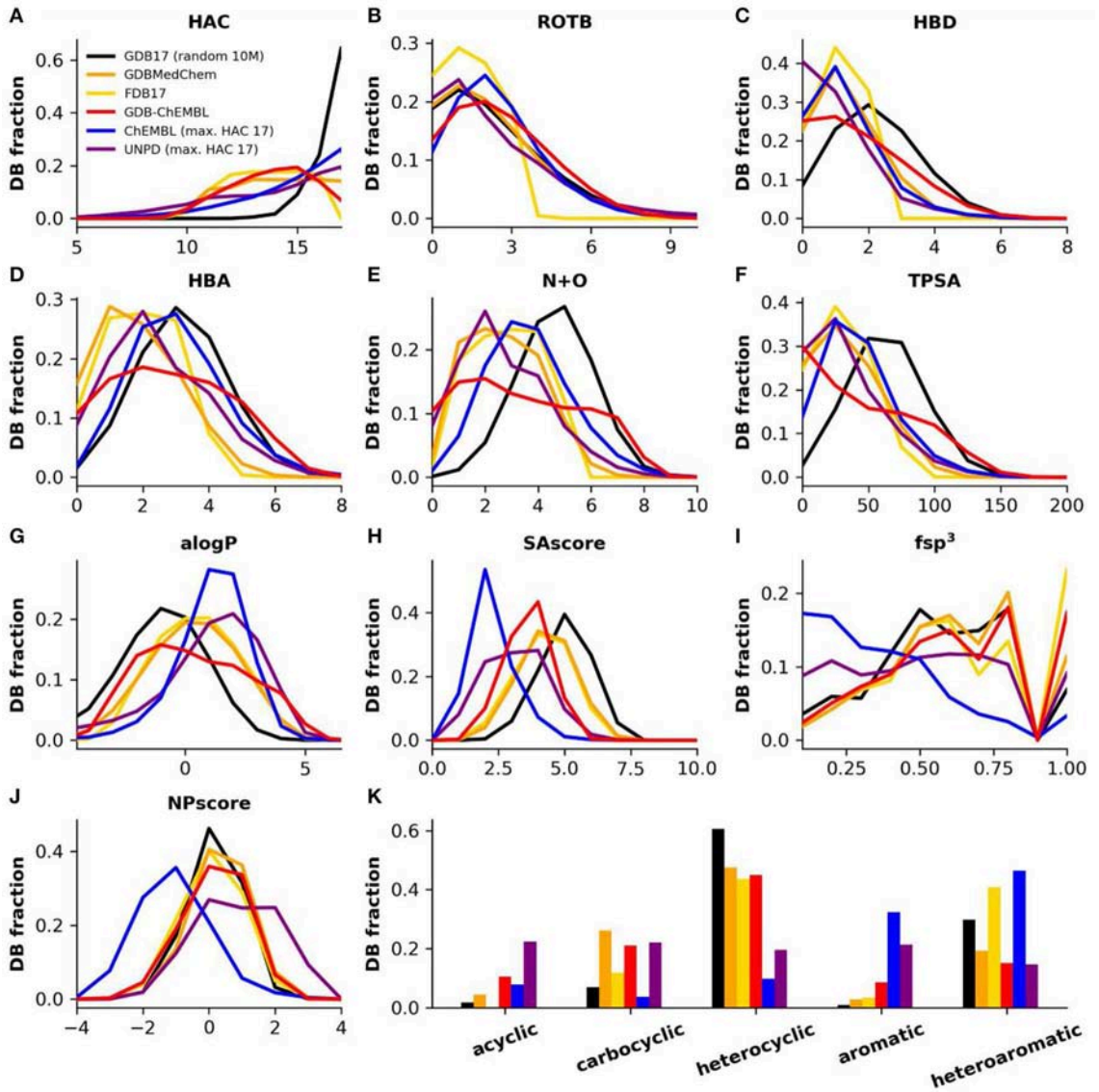


FIGURE 2 | (A) Heavy atom count (HAC), **(B)** rotatable bonds (ROTB), **(C)** hydrogen bond donors (HBD), **(D)** hydrogen bond acceptors (HBA), **(E)** nitrogen plus oxygen count (N+O), **(F)** topological polar surface area (TPSA), **(G)** computed partition coefficient (aLogP), **(H)** synthetic accessibility score (SAscore), **(I)** Fraction of sp^3 hybridized atoms (fsp^3), **(J)** natural product likeness score (NPscore), and **(K)** fraction of structures by ring class. Property histograms for GDB17 (black), GDBMedChem (orange), FDB17 (yellow), GDBChEMBL (red), ChEMBL (cpds with HAC $\leq$ 17, blue) and natural products (cpds with HAC $\leq$ 17, purple).

similarity analogs (**Figure 3H**). To facilitate similarity searches in GDBChEMBL, we have implemented a similarity search portal by which nearest neighbor searches of any molecule can be performed in GDBChEMBL using MQN, ECFP4, or a combined MQN-MHFP6 similarity, as described previously for GDBMedChem (Awale et al., 2019).

CONCLUSION

The data above demonstrate a substructure-based approach to select molecules from the generated database GDB17. As selection criterion we defined a ChEMBL-likeness score (CLscore) from the frequency occurrence of circular substructures, called molecular shingles, in a subset of the database ChEMBL consisting of compounds active on single protein targets with high confidence datapoints. This selection reduced GDB17 by 84.3%, leaving 26.2 billion molecules, which we sampled uniformly across molecular size, stereochemistry and heteroatoms to form GDBChEMBL comprising 10 million molecules.

Property profiles, chemical space maps and similarity searches show that GDBChEMBL is very different from our earlier GDB subsets FDB17 and GDBMedChem and spans chemical space more continuously. At the same time, the correlation between CLscore and the synthetic accessibility score (SAscore) implies that GDBChEMBL molecules will be on average easier

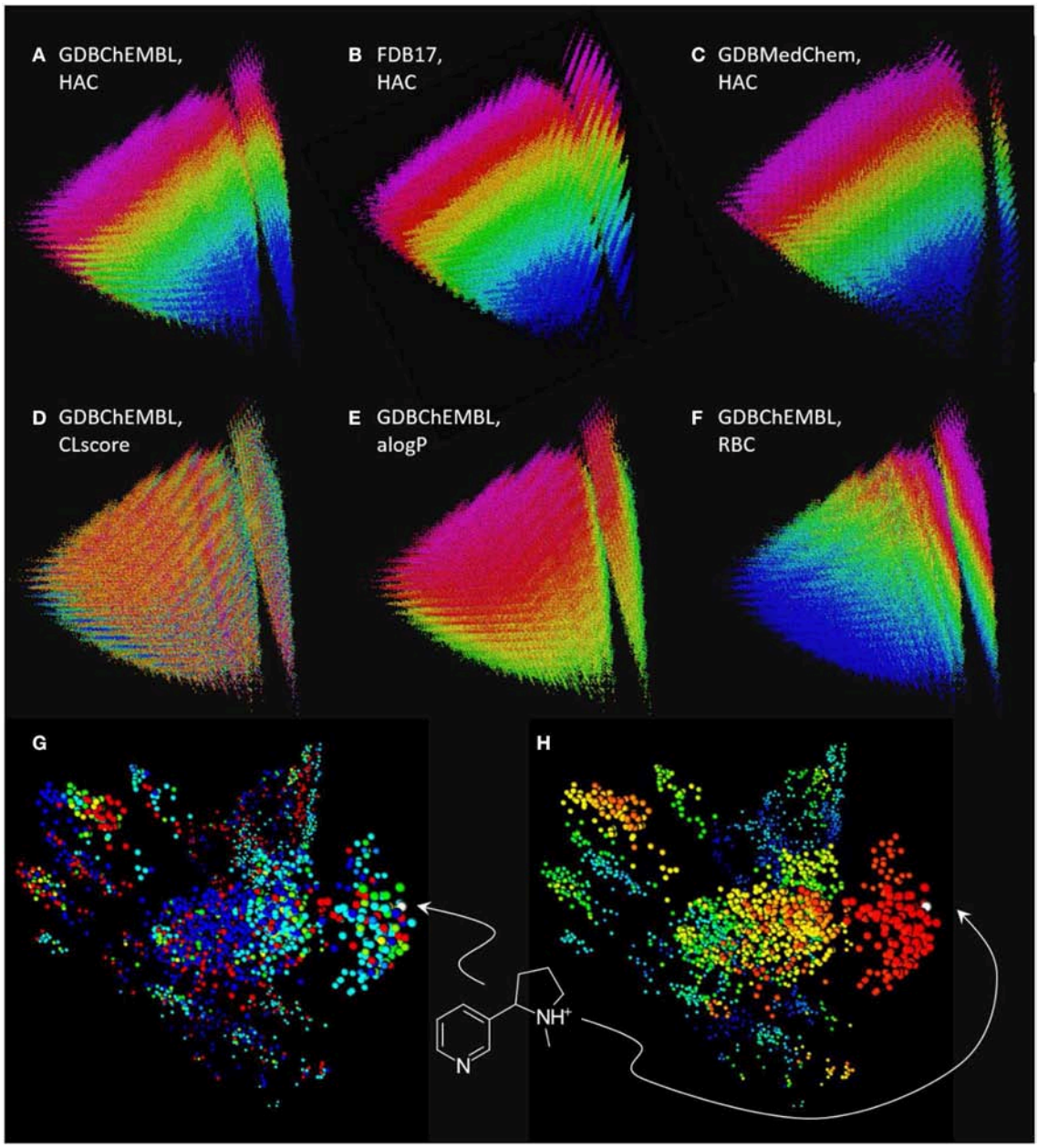


FIGURE 3 | Chemical space maps of GDBChEMBL, FDB17, and GDBMedChem. **(A)** PCA 3D-map of GDBChEMBL in MQN-space, color coded by heavy atom count; **(B)** same as a for FDB17; **(C)** same as a for GDBMedChem; **(D)** GDBChEMBL color-coded by CLscore value; **(E)** GDBChEMBL color-coded by calculated octanol/water partition coefficient alogP; **(F)** GDBChEMBL color-coded by rotatable bond count; **(G)** similarity map of MQN-nearest neighbors of nicotine from GDBChEMBL (red), FDB17 (cyan), and GDBMedChem (blue). Points in green and yellow indicate molecules shared by two databases. **(H)** Same as g color-coded by Xfp-similarity to nicotine. MQN maps a to f are accessible at http://faerun.gdb.tools. The similarity map of nicotine analogs g and h is accessible at: http://gdbtools.unibe.ch:8080/webMolCS/.

to synthesize than molecules from FDB17 and GDBMedChem, which have significantly lower CLscore and higher SAscores. We anticipate that the requirements for GDBChEMBL molecules to share a minimum number of substructures with molecules of known bioactivities from ChEMBL will also facilitate target prediction and the selection of interesting GDB molecules for synthesis and testing.

METHODS

Preparative Steps

ChEMBL Shingle Extraction

The ChEMBL (v 24.1) database was downloaded from https://www.ebi.ac.uk/chembl/. Data points for extraction of molecular shingles were selected by applying the same restrictions that were

used for extraction of training data for our Polypharmacology Browser PPB2 (Visini et al., 2017b). Structures were normalized to their major protonation state at pH 7.4 using ChemAxon cxcalc (v. 18.23.0). Molecular shingles for radii 1–3 were created using RDkit (2019.03.4) and converted to rooted, canonical, aromatic SMILES strings without retaining stereochemistry information. In association with abundancy in the ChEMBL, the SMILES substructures were stored as pickled python dictionary. Molecular substructures that were found <100 times were not stored.

CLscore Calculation

Scoring of GDB17 molecular structures was achieved by decomposition to molecular shingles in the exact same way as described for ChEMBL reference shingle extraction. For a specific query structure, all shingles are uniquely counted, then looked up in the ChEMBL reference database and upon match, logarithmic abundancy is summed up. The final CLscore is given by the ratio of total logarithmic abundancies of matched unique shingles to total unique shingles in the query structure. All respective scripts are accessible at: https://github.com/reymond-group/GDBChEMBL.

GDBChEMBL Generation

All 166.4 billion molecular structures of GDB17 were decomposed to unique substructures in the same way as described for ChEMBL reference molecules. Only structures with CLscore ≥3.3 were stored. The final GDBChEMBL was obtained by distribution of all filtered 26.2 billion structures to 538 property triplet bins (heavy atom, heteroatom and stereocenter count). Property information was gathered using RDKit. Bins with 5+ hetero atoms and/or 8+ chiral atoms were merged. The actual even sampling was performed by sorting all property bins by size and defining target structure count as 10 million. Iteratively, remaining target count was divided by count of remaining bins, keeping all bins of size smaller than the current number to sample randomly. For each step, number of previously selected structures was subtracted from target count until random sample per remaining bins was lower than bin size. At this point, sample size was kept constant for all further bins.

Visualizing GDBChEMBL in Faerun

Property color coded 3D maps for GDBChEMBL, FDB17, and GDBMedChem were generated using FUn (doc.gdb.tools/fun), an in-house developed framework for interactive visualization of chemical spaces on the web. Datasets were given as plain text, consisting of the four columns (space-separated): SMILES-string, numeric ID, 42 MQN descriptors (semicolon-separated) and further molecular properties used for map coloring (semicolon-separated). Next, the preprocessing toolchain was used to project the 42-dimensional MQN-space to 3D by applying Principal Component Analysis (PCA) and to generate all further files needed for visualization. Finally, the Underdark server was run using docker with Faerun visualization containers mapped.

Similarity Searching in GDBChEMBL

For better accessibility, GDBChEMBL is provided as a web-based interactive similarity search tool. The implementation uses HTML, Bootstrap, JavaScript, and the python Flask framework. Search times were reduced using Annoy trees (Approximate Nearest Neighbors Oh Yeah, https://github.com/spotify/annoy) which were created for the 42-dimensional MQN property space, as well as for 256-bit ECfp4. A third search option, MQN-MHFP6, initially searches using the MQN Annoy tree followed by resorting after Jaccard distance to query molecule in the MHFP6 fingerprint space (https://github.com/reymond-group/mhfp). The search tool is available at: gdb.unibe.ch/tools.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

SB designed and realized the study and wrote the paper. J-LR designed and supervised the study and wrote the paper.

FUNDING

This work was supported financially by the Swiss National Science Foundation.

REFERENCES

Awale, M., Probst, D., and Reymond, J. L. (2017a). WebMolCS: a web-based interface for visualizing molecules in three-dimensional chemical spaces. *J. Chem. Inf. Model.* 57, 643–649. doi: 10.1021/acs.jcim.6b00690

Awale, M., and Reymond, J.-L. (2019). Polypharmacology browser PPB2: target prediction combining nearest neighbors with machine learning. *J. Chem. Inf. Model.* 59, 10–17. doi: 10.1021/acs.jcim.8b00524

Awale, M., and Reymond, J. L. (2014). Atom pair 2D-fingerprints perceive 3D-molecular shape and pharmacophores for very fast virtual screening of ZINC and GDB-17. *J. Chem. Inf. Model.* 54, 1892–1897. doi: 10.1021/ci500232g

Awale, M., and Reymond, J. L. (2015). Similarity mapplet: interactive visualization of the directory of useful decoys and ChEMBL in high dimensional chemical spaces. *J. Chem. Inf. Model.* 55, 1509–1516. doi: 10.1021/acs.jcim.5b00182

Awale, M., Sirockin, F., Stiefl, N., and Reymond, J.-L. (2019). Medicinal chemistry aware database GDBMedChem. *Mol. Informatics* 38:1900031. doi: 10.1002/minf.201900031

Awale, M., Visini, R., Probst, D., Arus-Pous, J., and Reymond, J. L. (2017b). Chemical space: big data challenge for molecular diversity. *Chimia* 71, 661–666. doi: 10.2533/chimia.2017.661

Banerjee, P., Erehman, J., Gohlke, B.-O., Wilhelm, T., Preissner, R., and Dunkel, M. (2015). Super natural II—a database of natural products. *Nucleic Acids Res.* 43, D935–D939. doi: 10.1093/nar/gku886

Blum, L. C., and Reymond, J. L. (2009). 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* 131, 8732–8733. doi: 10.1021/ja902302h

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discovery Today* 23, 1241–1250. doi: 10.1016/j.drudis.2018.01.039

Congreve, M., Carr, R., Murray, C., and Jhoti, H. (2003). A rule of three for fragment-based lead discovery? *Drug Discovery Today* 8, 876–877. doi: 10.1016/S1359-6446(03)02831-9

Elton, D. C., Boukouvalas, Z., Fuge, M. D., and Chung, P. W. (2019). Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* 4, 828–849. doi: 10.1039/C9ME00039A

Ertl, P., and Schuffenhauer, A. (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* 1:8. doi: 10.1186/1758-2946-1-8

Fink, T., Bruggesser, H., and Reymond, J. L. (2005). Virtual exploration of the small-molecule chemical universe below 160 daltons. *Angew. Chem. Int. Ed. Engl.* 44, 1504–1508. doi: 10.1002/anie.200462457

Fink, T., and Reymond, J. L. (2007). Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* 47, 342–353. doi: 10.1021/ci600423u

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., et al. (2017). The ChEMBL database in 2017. *Nucleic Acids Res.* 45, D945–D954. doi: 10.1093/nar/gkw1074

Hu, Q., Peng, Z., Kostrowicki, J., and Kuki, A. (2011). LEAP into the Pfizer Global Virtual Library (PGVL) space: creation of readily synthesizable design ideas automatically. *Methods Mol. Biol.* 685, 253–276. doi: 10.1007/978-1-60761-931-4_13

Jayaseelan, K. V., Moreno, P., Truszkowski, A., Ertl, P., and Steinbeck, C. (2012). Natural product-likeness score revisited: an open-source, open-data implementation. *BMC Bioinformatics* 13:106. doi: 10.1186/1471-2105-13-106

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 42, D1091–D1097. doi: 10.1093/nar/gkt1068

Leach, A. R., and Hann, M. M. (2000). The *in silico* world of virtual libraries. *Drug Discov. Today* 5, 326–336. doi: 10.1016/S1359-6446(00)01516-6

Medina-Franco, J. L., Maggiora, G. M., Giulianotti, M. A., Pinilla, C., and Houghten, R. A. (2007). A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. *Chem. Biol. Drug Des.* 70, 393–412. doi: 10.1111/j.1747-0285.2007.00579.x

Mignani, S., Rodrigues, J., Tomas, H., Jalal, R., Singh, P. P., Majoral, J.-P., et al. (2018). Present drug-likeness filters in medicinal chemistry during the hit and lead optimization process: how far can they be simplified? *Drug Discovery Today* 23, 605–615. doi: 10.1016/j.drudis.2018.01.010

Nguyen, K. T., Blum, L. C., van Deursen, R., and Reymond, J.-L. (2009). Classification of organic molecules by molecular quantum numbers. *ChemMedChem* 4, 1803–1805. doi: 10.1002/cmdc.200900317

Poirier, M., Awale, M., Roelli, M. A., Giuffredi, G. T., Ruddigkeit, L., Evensen, L., et al. (2019). Identifying lysophosphatidic acid acyltransferase β (LPAAT-β) as the target of a nanomolar angiogenesis inhibitor from a phenotypic screen using the polypharmacology browser PPB2. *ChemMedChem* 14, 224–236. doi: 10.1002/cmdc.201800554

Probst, D., and Reymond, J.-L. (2018). A probabilistic molecular fingerprint for big data settings. *J. Cheminf.* 10:66. doi: 10.1186/s13321-018-0321-8

Probst, D., Reymond, J.-L., and Wren, J. (2018). FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. *Bioinformatics* 34, 1433–1435. doi: 10.1093/bioinformatics/btx760

Raghavendra, A. S., and Maggiora, G. M. (2007). Molecular basis sets - a general similarity-based approach for representing chemical spaces. *J. Chem. Inf. Model.* 47, 1328–1240. doi: 10.1021/ci600552n

Reymond, J.-L. (2015). The chemical space project. *Acc. Chem. Res.* 48, 722–730. doi: 10.1021/ar500432k

Riniker, S., and Landrum, G. A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminf.* 5:26. doi: 10.1186/1758-2946-5-26

Rosén, J., Gottfries, J., Muresan, S., Backlund, A., and Oprea, T. I. (2009). Novel chemical space exploration via natural products. *J. Med. Chem.* 52, 1953–1962. doi: 10.1021/jm801514w

Ruddigkeit, L., Blum, L. C., and Reymond, J. L. (2013). Visualization and virtual screening of the chemical universe database GDB-17. *J. Chem. Inf. Model.* 53, 56–65. doi: 10.1021/ci300535x

Ruddigkeit, L.;, van Deursen, R., Blum, L. C., and Reymond, J. L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* 52, 2864–2875. doi: 10.1021/ci300415d

Sterling, T., and Irwin, J. J. (2015). ZINC 15 – ligand discovery for everyone. *J. Chem. Inform. Model.* 55, 2324–2337. doi: 10.1021/acs.jcim.5b00559

van Hilten, N., Chevillard, F., and Kolb, P. (2019). Virtual compound libraries in computer-assisted drug discovery. *J. Chem. Inf. Model.* 59, 644–651. doi: 10.1021/acs.jcim.8b00737

Visini, R., Arus-Pous, J., Awale, M., and Reymond, J. L. (2017a). Virtual exploration of the ring systems chemical universe. *J. Chem. Inf. Model.* 57, 2707–2718. doi: 10.1021/acs.jcim.7b00457

Visini, R., Awale, M., and Reymond, J. L. (2017b). Fragment database FDB-17. *J. Chem. Inf. Model.* 57, 700–709. doi: 10.1021/acs.jcim.7b00020

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Bühlmann and Reymond. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 12 February 2020
doi: 10.3389/fchem.2020.00081

A Pilot Study of All-Computational Drug Design Protocol–From Structure Prediction to Interaction Analysis

*Yifei Wu, Lei Lou and Zhong-Ru Xie**

Computational Drug Discovery Laboratory, School of Electrical and Computer Engineering, College of Engineering, University of Georgia, Athens, GA, United States

Speeding up the drug discovery process is of great significance. To achieve that, high-efficiency methods should be exploited. The conventional wet-bench methods hardly meet the high-speed demand due to time-consuming experiments. Conversely, *in silico* approaches are much more efficient for drug discovery and design. However, *in silico* approaches usually serve as a supportive role in research processes. To fully exert the strength of computational methods, we propose a protocol which integrates various *in silico* approaches, from *de novo* protein structure prediction to ligand-protein interaction simulation. As a proof of concept, human SK2/calmodulin complex was used as a target for validation. First, we obtained a predicted structure of SK2/calmodulin and predicted binding sites which were consistent with the literature data. Then we investigated the ligand-protein interaction via virtual mutagenesis, flexible docking, and binding affinity calculation. As a result, the binding energies of mutants have similar trends compared with the EC_{50} values ($R = 0.6$ for NS309 in V481 mutants). The results indicate that our protocol can be applied to the drug design of structure unknown proteins. Our study also demonstrates that the integration of *in silico* approaches is feasible and it facilitates the acceleration of new drug discovery.

Keywords: ligand-protein docking, molecular dynamics simulation, computational drug discovery, SK2, structural prediction, binding site prediction, virtual mutation, pharmaceutical industry

OPEN ACCESS

Edited by:
Teodorico Castro Ramalho, Universidade Federal de Lavras, Brazil

Reviewed by:
Daniel Henriques Soares Leal, Federal University of Itajubá, Brazil
Daiana Mancini, Universidade Federal de Lavras, Brazil

****Correspondence:***
Zhong-Ru Xie
paulxie@uga.edu

Specialty section:
This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

Received: *07 October 2019*
Accepted: *24 January 2020*
Published: *12 February 2020*

Citation:
Wu Y, Lou L and Xie Z-R (2020) A Pilot Study of All-Computational Drug Design Protocol–From Structure Prediction to Interaction Analysis. Front. Chem. 8:81. doi: 10.3389/fchem.2020.00081

1. INTRODUCTION

To discover a new drug is an urgent but time-consuming process (Zhou et al., 2016; Gómez-Bombarelli et al., 2018). In the process of new drug development, *in silico* approaches have been successfully exploited to perform multiple simulations, such as selecting drug candidates from a database via high throughput virtual screening (Aparna et al., 2014; de Ruyck et al., 2016; Vilar et al., 2016; Imam and Gilani, 2017). The application of *in silico* approaches not only speeds up new drug discovery, but also collects related information, reveals the mechanisms and proposes new hypotheses. Compared to bench research, computational experiments perform high-efficiency simulations which considerably reduce the research time and the cost of experiments (Abel et al., 2017). However, in most cases, *in silico* approaches played an assisting role in the process. In this study, we propose an all-computational protocol integrating multiple *in silico* approaches to simulate the entire drug discovery process from *de novo* protein structure prediction to drug-protein interaction disclosure.

Up to now, reliable *in silico* approaches, such as molecular dynamics simulation (MD simulation) and machine learning (Durrant and McCammon, 2011; Lavecchia, 2015; Mortier et al., 2015), have been increasingly developed and applied in finding new drugs and optimizing them for treatment of diseases. However, most research projects only use one single *in silico* approach. For instance, the homology models of CIB2, a calcium- and integrin- binding protein, were constructed based on CIB1 structures using SWISS-MODEL server (Waterhouse et al., 2018). Based on the models, the way how the point mutations affect the affinities of calcium- and integrin-binding were predicted and then validated by *in vitro* experiments (Riazuddin et al., 2012). Besides, the results of ligand-protein docking were used to test the substrate specificity of OCT-1 and OCT-2 (organic cation transporter) to guide the following *in vivo* experimental validation (Papaluca and Ramotar, 2016). These works simply employed the *in silico* approaches as supportive methods, which did not sufficiently leverage the high-speed advantages of computational methods. Integrating the established *in silico* researches into an all-computation pipeline and producing validated good results is a milestone in the "omics" era.

Human small conductance calcium-activated (SK2) ion channels, consisted of SK2 subunits and calmodulin molecule, have been proved to be therapeutic targets for treatment of neuronal diseases, such as Parkinson's and amyotrophic lateral sclerosis (ALS) (Bond et al., 2005; Lu et al., 2009). The crystalized structures of human SK2 bound with Riluzole, an approved drug for ALS (Romano et al., 2015), and an analog [pdb (Berman et al., 2003) ligand ID: 658] of the SK2 activator CyPPA, an anti-ataxic agent (Herrik et al., 2012), have been released recently. It was also reported that ligands of two different chemical classes, Riluzole and its analog NS309, and CyPPA and its analogs, all bind to the same binding site where the interface of SK2 and calmodulin is. In addition, two key residues in the binding site were mutated to investigate how different mutations affect the potency of two ligands (Cho et al., 2018). As a proof of concept, we chose SK2 as the target to examine the protocol we proposed. Hence, in this study, we basically repeated the entire procedure of the previous study of Cho et al. (2018) with consecutive *in silico* approaches and compared our results with those of the bench experiments. We first constructed the human SK2/calmodulin model (PSK2) using SWISS-MODEL server and docked the ligands Riluzole and CyPPA analog 1 (PDB ligand ID: 658, see section 2.1), onto the predicted binding site. The predicted site is consistent with the reported results (Cho et al., 2018). Then the residues V481 and A477 in the binding pocket were virtually mutated and the mutation effects were assessed via binding energies (MM-GBSA ΔG_{Bind}) calculation. The results show that the binding energies of mutants have similar trends compared with the EC_{50} (the concentration of a drug that gives half-maximal response) values ($R = 0.6$ for NS309 in V481 mutants). Overall, our results suggest that this protocol of *in silico* approach can provide a systematic prediction on the unknown structures of proteins and potential drugs, and they also demonstrate the ability of *in silico* approaches to speed up the new drug design process.

2. MATERIALS AND METHODS

2.1. Protein Structure Preparation

Having a determined or predicted structure of the drug target protein is the first prerequisite of structure-based drug design. To prove the all-computational protocol is valid, we started our process from structure prediction. To predict the structure of SK2 and calmodulin from *Homo sapiens*, its amino acid sequence was obtained from the UniProt website (uniprot ID: Q9H2S1). Then we uploaded the sequence onto the SWISS-MODEL server, the most widely-used and reliable structure prediction server, to build a structure model (Bienert et al., 2016; Waterhouse et al., 2018) and selected the 3-D complex structure of *Rattus norvegicus* SK2 ion channels with NS309 (PDB ID: 4J9Z) (Zhang et al., 2013) as the template of structure prediction. 4J9Z was downloaded from RCSB's Protein Data Bank (Berman et al., 2003), and the predicted structure of SK2 and calmodulin were combined in Maestro (11.5 version, Schrodinger). To test the accuracy of modeling, we uploaded the predicted protein structure and the crystalized complex structure of human SK2 and calmodulin with Riluzole (PDB ID: 5V02) onto Zhang's web server (Zhang and Skolnick, 2005) to calculate the TM-align score. 5V02 was downloaded from RCSB's Protein Data Bank. In addition, site-directed mutants were constructed using the Mutate Residue function of Maestro.

A PDB structure of the target protein cannot be directly used in molecular docking without preprocessing. In most of the cases, a PDB file does not include the information of hydrogens, the (potential) charges of atoms, or the bond orders between any two atoms. In addition, the protein structures may be determined with a missing fragment(s), a low resolution or alternate positions, or under an unnatural condition, for example, low or high pH values. To make sure molecular docking can simulate the binding between ligands and the target protein correctly and precisely, the protein and ligand preparation is necessary. The wild-type and mutant (predicted) protein structures to be used for docking were processed by protein preparation wizard in Maestro (Sastry et al., 2013). The workflow of protein preparation contains three steps as follows: (1) Preprocess: assigning bond orders, adding hydrogens, creating zero-order bonds to metals, creating disulfide, filling in missing side chains using Prime, deleting waters beyond 5.00 Å from het groups (to keep the water molecules which may form hydrogen-bond bridges between the protein and the ligand and remove those cannot form hydrogen-bond bridges) and generating het states using Epik ($pH = 7.0 \pm 2.0$) (Shelley et al., 2007); (2) Optimization: setting $pH = 7.0$ and performing optimization; (3) Minimization: this step was performed using the OPLS3 force field (Harder et al., 2015). The converge heavy atoms to root-mean-square deviation (RMSD) is 0.30 Å.

2.2. Ligand Preparation

The 3-D molecular structures of Riluzole and NS309 were obtained from the PubChem database (Kim et al., 2018). The 3-D molecular structures of CyPPA analog 1 ((4-chloro-phenyl)-[2-(3,5-dimethyl-pyrazol-1-yl)-pyrimidin-4-yl]-amine) and

analog 2 (4-chloro-phenyl)-[2-(3,5-dimethyl-pyrazol-1-yl)-6-methyl-pyrimidin-4-yl]-amine) were built in Maestro (11.5 version, Schrodinger) based on a previous study (Cho et al., 2018). All the compounds were prepared using OPLS3 force field in Ligprep panel in Maestro (Sastry et al., 2013; Harder et al., 2015). The preparation process included converting 2D structures to 3D ones, adding hydrogens, computing correct partial charges, and optimizing the structures.

2.3. Binding Site Prediction

Knowing the potential ligand binding site(s) is also an important prerequisite prior to molecular docking. There are many well-developed binding site prediction methods and servers (Xie and Hwang, 2015); however, the predictions produced by different methods may disagree with each other. Therefore, researchers usually compare the prediction results produced by different methods to find the consensus among all predictions. In our study, binding site prediction process was completed using the LISE web server (http://lise.ibms.sinica.edu.tw/applet/) (Xie and Hwang, 2012), which is reported to have the highest accuracy (80–90% for a soluble protein), and the binding site prediction tool, Sitemap, in Maestro, which we used in the docking procedure after this step. The SK2 predicted structure was uploaded onto the LISE web server for binding site prediction. The top three predictions from LISE were then downloaded and imported into Maestro. Meanwhile, in Maestro, SiteMap (Halgren, 2007, 2009) predicted five ligand binding sites. After comparing the results obtained using two prediction methods, we used the consensus region to define the docking grid box.

2.4. Ligand-Protein Docking

In order to predict the details of the interaction between ligands and the target protein and to estimate their binding affinities (see section 2.5), ligand docking was conducted the extra-precision (XP) mode in Maestro. Maestro has three precision options for docking including high throughput virtual screening (HTVS), standard precision (SP), and extra precision (XP). Users can choose an option based on their need or the computational load. After the ligands and the target proteins were processed using Ligprep and protein preparation, respectively, a receptor grid box was generated according to the results of binding site prediction. The size of the receptor grid box was set as default. To investigate the interaction of the protein and ligands, Induced Fit Docking (IFD) (Farid et al., 2006; Sherman et al., 2006a,b; Clark et al., 2016) was performed in Maestro. Using the IFD, the Ligprep outputs were imported and docked to the target protein. The standard protocol was applied to generate up to 20 poses. The force field was OPLS3. Under the prime refinement tab, the conformations of binding site residues within 5 Å (default value) of the ligand were refined. In the Glide redocking process, the energy of each protein/ligand complex structure and the number of top structures were set as the default settings. The XP mode was used for all IFD process.

2.5. MM-GBSA Calculation

The binding energy (ΔG_{Bind}) between a protein and a ligand reflects how stable they bind to each other and how a point mutation affects the ligand binding. Therefore, we examined if our model can correctly predict the trend of binding affinity changes of the mutations on the target protein. In this study, ΔG_{Bind} were estimated using the Prime MM-GBSA module in Maestro (Greenidge et al., 2012). In MM-GBSA panel, the pose viewer files of docked complex were uploaded. The solvation model was VSGB and the force field was OPLS3 (Li et al., 2011). Prime MM-GBSA ΔG_{Bind} was calculated using this equation:

$$\Delta G_{Bind} = E_{complex}(\text{minimized}) - [E_{ligand}(\text{minimized}) + E_{receptor}(\text{minimized})] \quad (1)$$

Where ΔG_{Bind} is binding free energy and $E_{complex}$(minimized), E_{ligand}(minimized), and $E_{receptor}$(minimized) are minimized energies of receptor-ligand complex, ligand and receptor, respectively.

2.6. Virtual Mutation

Based on the literature, valine 481 (V481) of SK2 is a crucial residue which forms the hydrophobic core between SK2 and calmodulin (Cho et al., 2018). To investigate the impacts of V481 mutations in the binding pocket using *in silico* approaches, we first implemented the site-directed mutagenesis in Maestro. The V481 was virtually mutated to alanine, serine, threonine, aspartate, or phenylalanine. Then NS309 and CyPPA analog 2 were docked onto the mutated binding site of PSK2 using IFD and the binding free energies were calculated using MM-GBSA to reveal the effects of mutated residues.

Alanine 477 (A477) is another vital residue in the binding pocket (Cho et al., 2018). We exploited the same method mentioned above to virtually mutate A477 to isoleucine, leucine, valine, serine, threonine, arginine, and aspartate, docked NS309 and CyPPA analog 2 onto the mutated binding site of PSK2 and calculated the binding free energies using MM-GBSA.

2.7. Molecular Dynamics Simulation

The molecular dynamics (MD) simulations were performed using GROMACS version 2018.1 and CHARMM36 all-atom force field (March 2019) (Vanommeslaeghe et al., 2010, 2012; Vanommeslaeghe and MacKerell, 2012; Yu et al., 2012; Gutiérrez et al., 2016). The starting coordinates of each protein-ligand complex were obtained from docking experiments. Then we defined a dodecahedral unit cell and filled it with water molecules. After adding ions, the complex was minimized for 50,000 steps of steepest descent minimization. Next, the complex was equilibrated using an NVT ensemble (constant Number of particles, Volume, and Temperature) and NPT ensemble (the Number of particles, Pressure, and Temperature). The target temperature for equilibration was 300 K. At last, the simulations were performed for 30 ns. After the MD simulations, we calculated the RMSD of the residues which were mentioned in the previous research in four trajectories (Cho et al., 2018). Then, we selected four time points of two residues: I100 on protein-NS309 complex (15,000, 18,000, 24,000, and 30,000 ps) and D64 on protein-CyPPA analog 1 (2,610, 6,000, 15,000, and 21,000 ps). Finally, four conformations of both residues were converted into PDB files and were superposed using PyMol.

3. RESULTS

3.1. Structure Prediction of Human SK2/Calmodulin Complex

The structure of human SK2/calmodulin was generated using template-based modeling on the SWISS-MODEL server (**Figure 1A**). Based on the structure of SK2/calmodulin from *Rattus norvegicus* (PDB ID: 4J9Z), the structure models were predicted using the amino acid sequences of human SK2 and calmodulin. Additionally, the loop (residue A403 to residue D413, the "intrinsically disordered fragment"—IDF), which had not been determined in structure of human SK2/calmodulin complex (PDB: 5V02), was obtained in the predicted model (**Figure 1A**). It suggests that the predicted structure can be used to supplement the crystallized structure.

To test the accuracy of the predicted model, we used TM-align server to calculate the TM-align score (Zhang and Skolnick, 2005). The predicted models were aligned with 5V02, and the TM-align scores of SK2 and calmodulin are 0.99124 and 0.89215. These TM-align scores show that the structure of human SK2/calmodulin complex has been accurately predicted.

3.2. Binding Site Prediction

To determine the binding pocket in SK2/calmodulin complex, the Sitemap and LISE were exploited to predict binding sites (Halgren, 2007, 2009; Xie and Hwang, 2012). The top three predicted results from LISE were overlapped with the results of Sitemap. After the comparison, we found that there was only one consensus. Then we selected this binding site to generate receptor grid for docking. To verify the accuracy of binding site prediction, we also overlapped the predicted binding site with that of 5V02 (**Figure 1B**). Notably, the predicted binding site is the same binding site of Riluzole in 5V02, which suggests that this binding site is the potential binding site for the ligands. Hence, the *in silico* approaches successfully predict the accurate binding site.

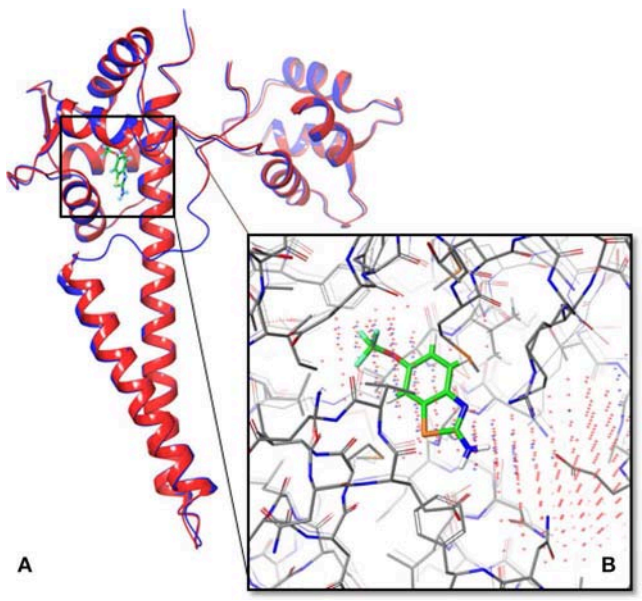


FIGURE 1 | Overlapped structures of the predicted model of human SK2 ion channels and 5V02. **(A)** The predicted structure of human SK2 ion channels (PSK2) in blue overlapped with its crystallized structure (5V02) in red. The structure of a missing loop (IDF) near the ligand binding pocket in 5V02/5V03 has also been predicted. The green compound was Riluzole, which was the ligand in 5V02. **(B)** The predicted binding sites (Blue dots represent the results from Sitemap; red dots represent the results from LISE) overlapped with the Riluzole binding site in 5V02.

3.3. Molecular Docking

Based on the previous study (Cho et al., 2018), the interface of SK2 and calmodulin can be bound by Riluzole, NS309, CyPPA analog 1, and analog 2. To investigate whether we can obtain the same results using *in silico* approaches, we first docked Riluzole and CyPPA analog 1 onto the predicted model (PSK2) via IFD (Induced Fit Docking) in Maestro, and redocked these ligands onto the determined structures (5V02 and 5V03) and calculated the binding energies as the control. Then we calculated the binding energies using MM-GBSA to estimate the binding affinities (Greenidge et al., 2012). As a result, the MM-GBSA ΔG_{Bind} of PSK2 with Riluzole and CyPPA analog 1 are −40.19 and −56.11 kcal/mol, respectively. As indicated in **Table 1**, those results of PSK2 are compatible to the results of 5V02 and 5V03, which demonstrate that accurate results can be obtained using *in silico* approaches. Additionally, the docking pose of PSK2 with Riluzole is almost identical with that in 5V02 (**Figure 2A**, the RMSD between the ligand of crystal structure and docking poses on 5V02 or PSK2 is 0.43 or 0.72 Å), which indicates that the simulated result from IFD can obtain accurate data in comparison with the results of crystallization. In **Figure 2B**, the coordinates of docked and native ligands are almost the same even though the poses of two ligands are not completely overlapped (the RMSD between the ligand of crystal structure and docking poses on 5V03 or PSK2 is 1.59 or 0.87 Å). Analyzing the docking results, Riluzole and CyPPA analog 1 interact with residues in both SK2 and calmodulin (**Figure 3**). The interacting residues in the binding site are mostly hydrophobic residues. As CyPPA analogs are larger molecules, they interact with more residues. Compared the list of interacting residues (**Table S1**), our results are almost identical to those of the previous study (Cho et al., 2018). The discrepancy may be because Maestro analyzes the interactions between ligands and the protein based on the interaction energy and the previous

TABLE 1 | MM-GBSA ΔG_{Bind} of ligands bound to crystallized structures and predicted structures.

Protein	Ligand	MM-GBSA ΔG_{Bind} (kcal/mol)
5V02	Riluzole	−40.83
5V03	CyPPA analog 1	−62.72
PSK2	Riluzole	−40.19
PSK2	NS309	−49.97
PSK2	CyPPA analog 1	−56.11
PSK2	CyPPA analog 2	−64.84

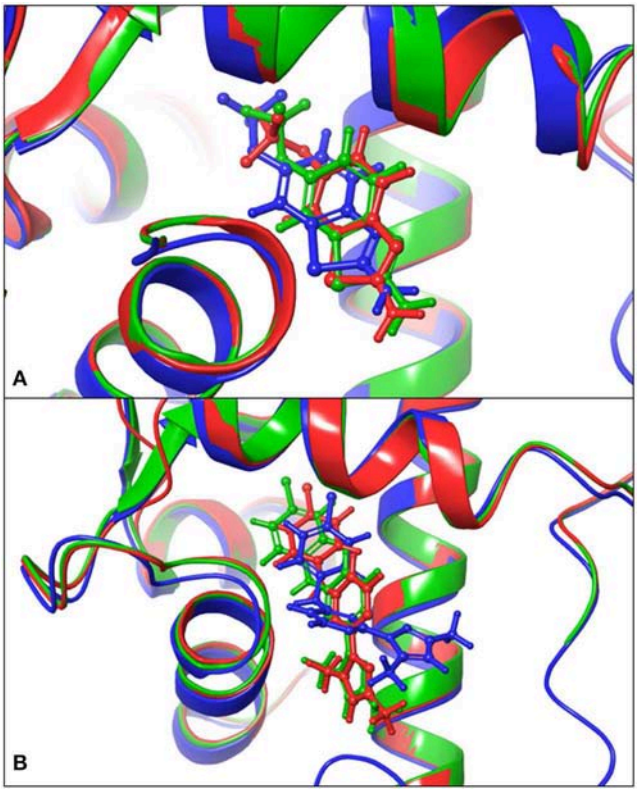


FIGURE 2 | Comparison of the docked poses and the native structures. **(A)** Comparison of the docking results on the PSK2 in blue color, the docking results on 5V02 in green color, and determined structures 5V02 in red color; The blue-colored Riluzole is the docked pose on the PSK2, the green-colored Riluzole is the docked pose on 5V02, and the red-colored Riluzole is the determined structure in 5V02; **(B)** Comparison of the docking results on the PSK2 in blue color, the docking results on 5V03 in green color, and determined structure 5V03 in red color; The blue-colored CyPPA analog 1 is the docked pose on the PSK2, the green-colored CyPPA analog 1 is the docked pose on 5V03 and the red-colored CyPPA analog 1 is the determined structure in 5v03.

study (Cho et al., 2018) simply lists the residues within 5 Å of either ligand. According to the docking results of PSK2 with Riluzole and CyPPA analog 1, we docked NS309 and CyPPA analog 2 on the same binding site using the same methods mentioned above (**Figure 4** and **Figure S1**). The binding affinities MM-GBSA ΔG_{Bind} are showed in **Table 1**. The MM-GBSA ΔG_{Bind} values of CyPPA analog 1 and CyPPA analog 2 on the PSK2 are more negative than those of the other ligands, which suggests that CyPPA analog 1 and CyPPA analog 2 are promising drug candidates.

To verify the ligand-induced perturbations, NMR spectrum was used to identify residues with significant chemical shifts in previous experiments (Cho et al., 2018). With computational approaches, we ran MD simulations for each protein-ligand complex to simulate the conformational changes after the binding of ligands. A previous research reports that the residues on EF hands of calmodulin had conformational changes due to the ligand binding (Cho et al., 2018). Therefore, we calculated

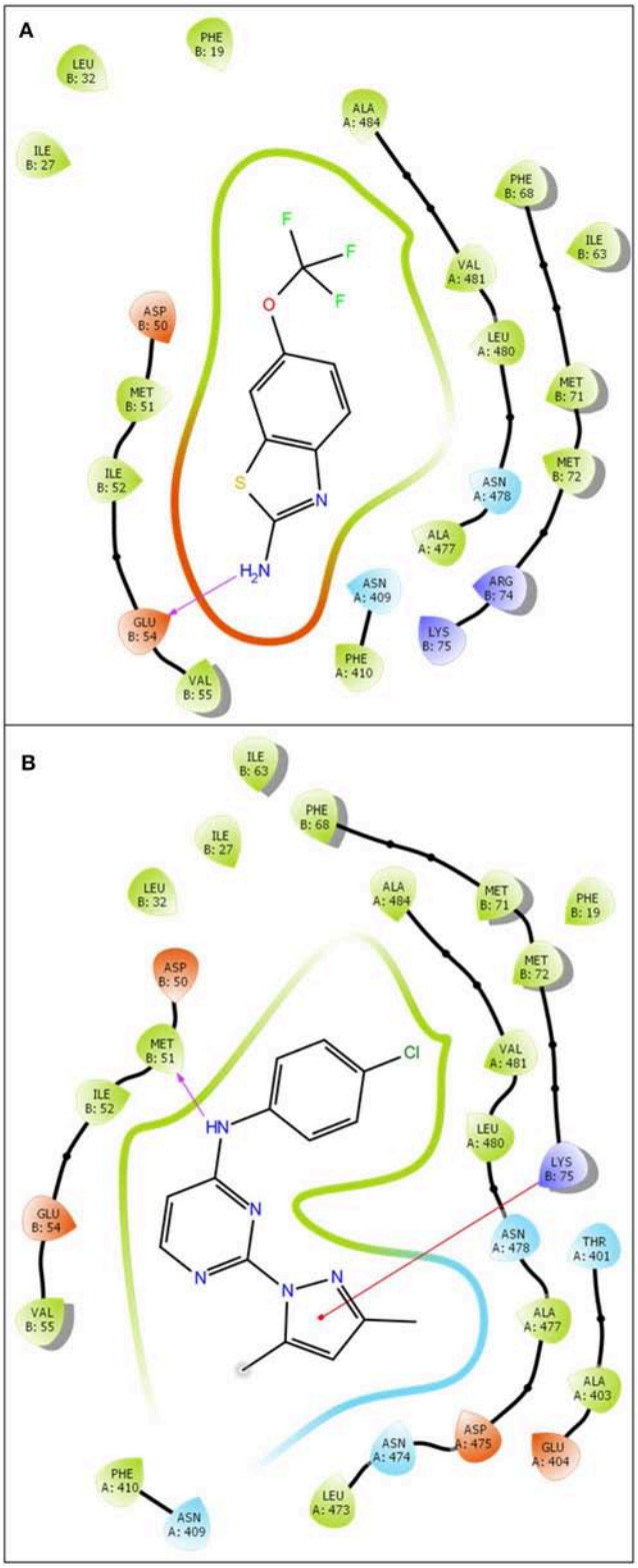


FIGURE 3 | 2D ligand-protein interaction of Riluzole **(A)** and CyPPA analog 1 **(B)** at PSK2 binding site. The pink arrow is referred to hydrogen bond. The red line is referred to Pi-cation interaction.

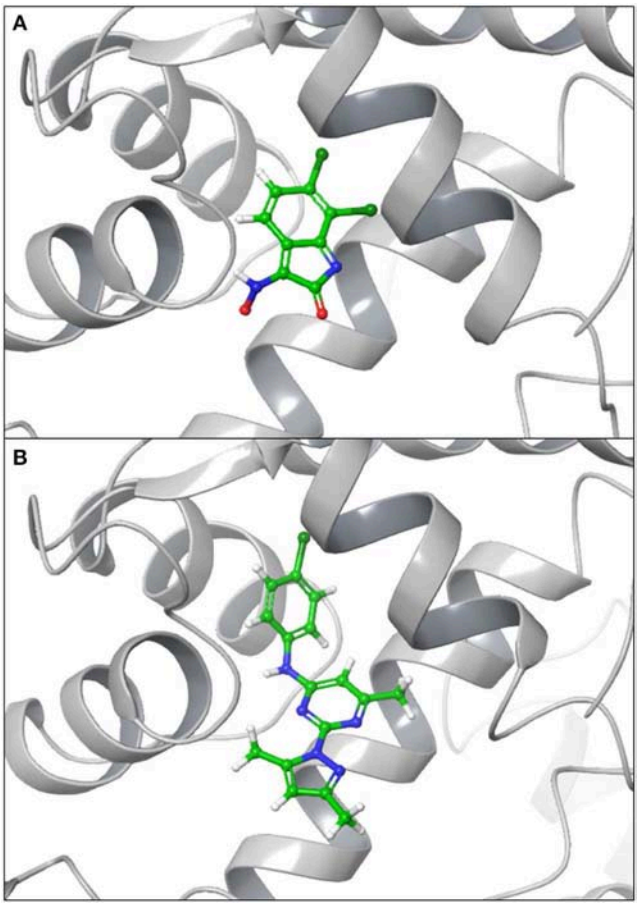


FIGURE 4 | Docking poses of SK2 predicted structure with NS309 **(A)** and CyPPA analog 2 **(B)**.

RMSD for those residues. According to the results of RMSD, we compared four trajectories of each residue and superposed their different poses at different time points. As a result, we found I100, on the complex of NS309, had obvious perturbations on four time points (**Figure 5** and **Figure S2**). In addition, D64 which was a critical residue for calcium sensing, also showed dramatically conformational changing on the complex structure of PSK2 and CyPPA analog 1 (**Figure 6** and **Figure S3**). Hence, the different poses of those residues demonstrated that the ligand could induce the perturbations of the calmodulin, which was consistent with the conclusion in the previous research (Cho et al., 2018).

3.4. Virtual Mutations at V481 and A477 in the Binding Pocket

The results of a previous study (Cho et al., 2018) show that the site-directed mutations on the key residues V481 and A477 in the binding site result in changes in the binding affinities. To validate that the *in silico* approach can simulate and predict the impacts of these mutations, we performed the virtual mutation experiments, docked the ligands on all mutants, and calculated the corresponding binding energies of each mutant. The MM-GBSA ΔG_{Bind} of V481 mutants are shown in **Table 2**. To verify the accuracy of *in silico* approaches, the Pearson's correlation coefficients between MM-GBSA ΔG_{Bind} and EC_{50} were calculated ($R = 0.6$ for NS309 in V481 mutants). **Table 2** indicates that substitutions with small side chains (V481S) or a charged amino acid (V481D) significantly decrease the binding affinities of NS309. Conversely, the binding affinity of NS309 in PSK2 V481F is close to that in wild-type. The results above are consistent with the conclusion in literature, that is, this position requires a non-charged residue with a bigger side chain (Cho et al., 2018).

Similarly, we find that CyPPA analog 2 in PSK2 V481F mutant with large aromatic side chains also shows a close binding affinity compared to that in PSK2 WT (**Table 2**), which demonstrates a good correlation between calculated binding affinities and EC_{50}. The mutants with small side chains (V481A and V481S) or a charged amino acid (V481D) also shows the relatively lower binding affinities of CyPPA analog 2. Those results demonstrate that the simulation results are compatible with the data from biological experiments. The consistent conclusion has successfully proved that the all-computational protocol can be widely applied in future biomedical research.

In **Table 3**, all PSK2 A477 mutants have slightly lower NS309 potency than that of PSK2 WT and the predicted binding affinities MM-GBSA ΔG_{Bind} have similar results. According to the results in literature (Cho et al., 2018), CyPPA analog 2 should exhibit shifted potencies at PSK2 A477L, PSK2 A477V, PSK2 A477S, PSK2 A477T, PSK2 A477R, and PSK2 A477D, but not at PSK2 A477I. However, in **Table 3**, the MM-GBSA ΔG_{Bind} of PSK2 A477I is not different from those of other mutants. As an isoleucine has a long side chain, different rotamers may largely change the estimated binding affinities. Performing an MD simulation may be a good solution to optimize the structures of mutants and improve the docked poses and estimated binding energies.

4. DISCUSSION

In the field of new drug discovery, research efficiency is particularly essential. On one hand, the speed of new drug development is of great importance to patients, especially the ones with fatal diseases such as cancers or acute infectious diseases (Shi et al., 2015). Statistics show that there will be around 1.7 million new cancer cases and 600 thousands cancer deaths in the United States in 2019 (Siegel et al., 2019). This race against time has always been a huge challenge for the researchers in this field. On the other hand, to bring a new drug to the market from compound identification to final FDA approval usually costs up to billions of dollars (Cleary et al., 2018). Therefore, the time and cost-efficient virtual process of drug development will benefit many pharmaceutical companies and our society. *In silico* approaches which can save considerable amount of research time and cost should be applied to drug design. With the rapid development of computer science and engineering, the availability and accuracy of *in silico* approaches have been constantly improving. Although many progresses have been made in utilizing *in silico* approaches to simulate certain biological experiments, the whole experimental processes completed in the virtual way from protein structure simulation

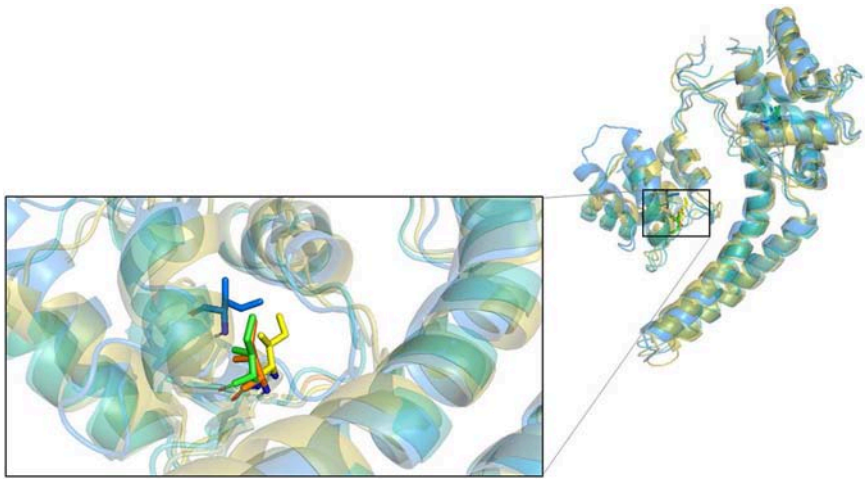


FIGURE 5 | Superposition of four conformations of I100 at four time points on the simulation trajectory of the complex of PSK2 and NS309.

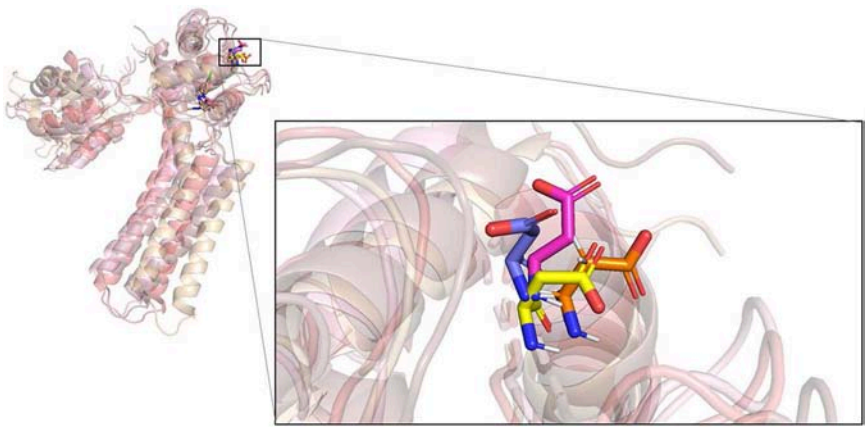


FIGURE 6 | Superposition of four conformations of D64 at four time points on the simulation trajectory of the complex of PSK2 and CyPPA analog 1.

TABLE 2 | EC_{50} and MM-GBSA ΔG_{Bind} of NS309 and CyPPA analog 2 bound to the V481 mutants.

Protein	NS309		CyPPA analog 2	
	EC_{50} (μM)	MM-GBSA ΔG_{Bind} (kcal/mol)	EC_{50} (μM)	MM-GBSA ΔG_{Bind} (kcal/mol)
WT	1.1	−49.97	2.5	−64.84
V481A	4.3	−54.28	>250	−29.05
V481S	5.1	−32.65	>250	−49.94
V481T	1.2	−32.77	46.6	−49.97
V481D	7.4	−25.97	>250	−46.75
V481F	0.8	−57.77	3.3	−64.09

TABLE 3 | EC_{50} and MM-GBSA ΔG_{Bind} of NS309 and CyPPA analog 2 bound to the A477 mutants.

Protein	NS309		CyPPA analog 2	
	EC_{50} (μM)	MM-GBSA ΔG_{Bind} (kcal/mol)	EC_{50} (μM)	MM-GBSA ΔG_{Bind} (kcal/mol)
WT	1.1	−49.97	2.5	−67.14
A477I	2.0	−37.26	2.0	−47.89
A477L	1.7	−24.71	99.0	−37.60
A477V	2.0	−39.67	>250	−45.63
A477S	3.0	−36.65	>250	−46.05
A477T	2.0	−23.36	>250	−51.04
A477R	2.1	−31.24	>250	−51.50
A477D	3.9	−30.94	>250	−49.71

to protein-drug interaction characterization has never been achieved before.

In this study, we proposed a protocol of integration of *in silico* approaches to simulate the process from protein structure determination to key residues mutagenesis and characterization. To validate this strategy, we selected the human SK2 ion channels as our target protein. With successful prediction, we obtained

accurate protein structures (TM-align score >0.5) and the same binding site as the crystallized structure. Furthermore, the docking poses of Riluzole and CyPPA analog 1 are consistent with the ligand bound conformations in the crystallized structures. We also successfully reproduced similar effects of site-directed mutagenesis on the ligand binding, which demonstrated great potential of the integration of *in silico* approaches. However, the purpose of integrating *in silico* approaches is not to completely replace biological experiments but to speed up the drug discovery process with the continuous and automatic computational process.

A possible reason why an all-computation protocol of drug design has not been proposed and implemented is the inaccuracy or uncertainty of prediction results might accumulate in the sequential computational pipeline. However, the state-of-the-art bioinformatics algorithms, software or servers have been highly accurate in many cases so that it is time to integrate them into a fully computational process or even a fully automatic process. This is the first study to demonstrate the feasibility of an all-computational protocol in drug design. To achieve the ultimate goal of "automatic" drug discovery, more online servers or computational algorithms like PROCHECK (Laskowski et al., 1993), which can be used to assess or estimate the reliability of prediction results generated by each *in silico* approach, need to be developed. Despite its innovative approach, there are a few limitations of this study. First, the accuracy of protein structure prediction relies on the methods and/or templates. In our research, we selected the SK2/calmodulin from *Rattus norvegicus* (PDB ID: 4J9Z) as the template to build protein structure on SWISS MODEL, whose results are more accurate than the results from other webservers (data not shown). Hence, a reliable tool or method is critical to the accuracy of the final simulated results. Second, the binding affinity changes are hard to be precisely reproduced, especially those of the mutants, because considering possible conformational changes on target proteins is still the biggest challenge of docking. This suggests that MD simulation which can simulate the conformational changes should be used to further improve the precision of the predictions.

In conclusion, this work established and demonstrated an integrated protocol of *in silico* approaches for the first time. Its applicability can be potentially extended beyond the characterizing of SK2 ion channels to investigating other proteins with unknown structures, such as the Alzheimer's disease related proteins (Fitzpatrick et al., 2017; Hatami et al., 2017), which are also treatment targets of neural degenerative diseases. Although there are challenges to the *in silico* approaches, our work still paves a new way toward an automatic procedure of drug design.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

YW carried out the research, analyzed the data, and drafted the manuscript. LL analyzed the data. Z-RX conceived the project, guided the research, and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work has been supported by a start-up grant from the College of Engineering, University of Georgia.

ACKNOWLEDGMENTS

The authors would like to acknowledge Dr. Gareth Young and Dr. Shenping Liu for kindly providing their experimental data (EC_{50}), and Dr. Wen-Ching Chan, Xingzi Yuan, Hsin-Tzu Wang, Jane Hua, and Crystal Zhu for the useful suggestions on editing this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2020.00081/full#supplementary-material

REFERENCES

Abel, R., Mondal, S., Masse, C., Greenwood, J., Harriman, G., Ashwell, M. A., et al. (2017). Accelerating drug discovery through tight integration of expert molecular design and predictive scoring. *Curr. Opin. Struct. Biol.* 43, 38–44. doi: 10.1016/j.sbi.2016.10.007

Aparna, V., Dineshkumar, K., Mohanalakshmi, N., Velmurugan, D., and Hopper, W. (2014). Identification of natural compound inhibitors for multidrug efflux pumps of *Escherichia coli* and pseudomonas aeruginosa using *in silico* high-throughput virtual screening and *in vitro* validation. *PLoS ONE* 9:e101840. doi: 10.1371/journal.pone.0101840

Berman, H. M., Bourne, P. E., Westbrook, J., and Zardecki, C. (2003). "The protein data bank," in *Protein Structure*, ed D. Chasman (Boca Raton, FL: CRC Press), 394–410. doi: 10.1201/9780203911327.ch14

Bienert, S., Waterhouse, A., de Beer, T. A., Tauriello, G., Studer, G., Bordoli, L., et al. (2016). The swiss-model repository—new features and functionality. *Nucleic Acids Res.* 45, D313–D319. doi: 10.1093/nar/gkw1132

Bond, C. T., Maylie, J., and Adelman, J. P. (2005). Sk channels in excitability, pacemaking and synaptic integration. *Curr. Opin. Neurol.* 15, 305–311. doi: 10.1016/j.conb.2005.05.001

Cho, L. T.-Y., Alexandrou, A. J., Torella, R., Knafels, J., Hobbs, J., Taylor, T., et al. (2018). An intracellular allosteric modulator binding pocket in SK2 ion channels is shared by multiple chemotypes. *Structure* 26, 533–544. doi: 10.1016/j.str.2018.02.017

Clark, A. J., Tiwary, P., Borrelli, K., Feng, S., Miller, E. B., Abel, R., et al. (2016). Prediction of protein–ligand binding poses via a combination of induced fit docking and metadynamics simulations. *J. Chem. Theory Comput.* 12, 2990–2998. doi: 10.1021/acs.jctc.6b00201

Cleary, E. G., Beierlein, J. M., Khanuja, N. S., McNamee, L. M., and Ledley, F. D. (2018). Contribution of NIH funding to new drug approvals 2010–2016. *Proc. Natl. Acad. Sci. U.S.A.* 115, 2329–2334. doi: 10.1073/pnas.1715368115

de Ruyck, J., Brysbaert, G., Blossey, R., and Lensink, M. F. (2016). Molecular docking as a popular tool in drug design, an *in silico* travel. *Adv. Appl. Bioinform. Chem.* 9:1. doi: 10.2147/AABC.S105289

Durrant, J. D., and McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biol.* 9:71. doi: 10.1186/1741-7007-9-71

Farid, R., Day, T., Friesner, R. A., and Pearlstein, R. A. (2006). New insights about herg blockade obtained from protein modeling, potential energy mapping, and docking studies. *Bioorg. Med. Chem.* 14, 3160–3173. doi: 10.1016/j.bmc.2005.12.032

Fitzpatrick, A. W., Falcon, B., He, S., Murzin, A. G., Murshudov, G., Garringer, H. J., Crowther, R. A., et al. (2017). Cryo-em structures of tau filaments from Alzheimer's disease. *Nature* 547:185. doi: 10.1038/nature23002

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., et al. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* 4, 268–276. doi: 10.1021/acscentsci.7b00572

Greenidge, P. A., Kramer, C., Mozziconacci, J.-C., and Wolf, R. M. (2012). MM/GBSA binding energy prediction on the PDBbind data set: successes, failures, and directions for further improvement. *J. Chem. Inf. Model.* 53, 201–209. doi: 10.1021/ci300425v

Gutiérrez, I. S., Lin, F.-Y., Vanommeslaeghe, K., Lemkul, J. A., Armacost, K. A., Brooks, C. L. III., et al. (2016). Parametrization of halogen bonds in the charmm general force field: improved treatment of ligand–protein interactions. *Bioorg. Med. Chem.* 24, 4812–4825. doi: 10.1016/j.bmc.2016.06.034

Halgren, T. (2007). New method for fast and accurate binding-site identification and analysis. *Chem. Biol. Drug Des.* 69, 146–148. doi: 10.1111/j.1747-0285.2007.00483.x

Halgren, T. A. (2009). Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* 49, 377–389. doi: 10.1021/ci800324m

Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J. Y., et al. (2015). OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* 12, 281–296. doi: 10.1021/acs.jctc.5b00864

Hatami, A., Monjazeb, S., Milton, S., and Glabe, C. G. (2017). Familial Alzheimer's disease mutations within the amyloid precursor protein alter the aggregation and conformation of the amyloid-β peptide. *J. Biol. Chem.* 292, 3172–3185. doi: 10.1074/jbc.M116.755264

Herrik, K. F., Redrobe, J. P., Holst, D., Hougaard, C., Nielsen, K. S., Nielsen, A. N., et al. (2012). CyPPA, a positive SK3/SK2 modulator, reduces activity of dopaminergic neurons, inhibits dopamine release, and counteracts hyperdopaminergic behaviors induced by methylphenidate1. *Front Pharmacol.* 3:11. doi: 10.3389/fphar.2012.00011

Imam, S. S., and Gilani, S. J. (2017). Computer aided drug design: a novel loom to drug discovery. *Org. Med. Chem.* 1, 1–6. doi: 10.19080/OMCIJ.2017.01.555567

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2018). Pubchem 2019 update: improved access to chemical data. *Nucleic Acids Res.* 47, D1102–D1109. doi: 10.1093/nar/gky1033

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). Procheck: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* 26, 283–291. doi: 10.1107/S0021889892009944

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today* 20, 318–331. doi: 10.1016/j.drudis.2014.10.012

Li, J., Abel, R., Zhu, K., Cao, Y., Zhao, S., and Friesner, R. A. (2011). The VSGB 2.0 model: a next generation energy model for high resolution protein structure modeling. *Proteins* 79, 2794–2812. doi: 10.1002/prot.23106

Lu, L., Timofeyev, V., Li, N., Rafizadeh, S., Singapuri, A., Harris, T. R., et al. (2009). α-actinin2 cytoskeletal protein is required for the functional membrane localization of a Ca2+-activated K+ channel (SK2 channel). *Proc. Natl. Acad. Sci. U.S.A.* 106, 18402–18407. doi: 10.1073/pnas.0908207106

Mortier, J., Rakers, C., Bermudez, M., Murgueitio, M. S., Riniker, S., and Wolber, G. (2015). The impact of molecular dynamics on drug design: applications for the characterization of ligand–macromolecule complexes. *Drug Discov. Today* 20, 686–702. doi: 10.1016/j.drudis.2015.01.003

Papaluca, A., and Ramotar, D. (2016). A novel approach using *C. elegans* DNA damage-induced apoptosis to characterize the dynamics of uptake transporters for therapeutic drug discoveries. *Sci. Rep.* 6:36026. doi: 10.1038/srep36026

Riazuddin, S., Belyantseva, I. A., Giese, A. P., Lee, K., Indzhykulian, A. A., Nandamuri, S. P., et al. (2012). Alterations of the CIB2 calcium-and integrin-binding protein cause usher syndrome type 1J and nonsyndromic deafness DFNB48. *Nat. Genet.* 44, 1265–1271. doi: 10.1038/ng.2426

Romano, S., Coarelli, G., Marcotulli, C., Leonardi, L., Piccolo, F., Spadaro, M., et al. (2015). Riluzole in patients with hereditary cerebellar ataxia: a randomised, double-blind, placebo-controlled trial. *Lancet Neurol.* 14, 985–991. doi: 10.1016/S1474-4422(15)00201-X

Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., and Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* 27, 221–234. doi: 10.1007/s10822-013-9644-8

Shelley, J. C., Cholleti, A., Frye, L. L., Greenwood, J. R., Timlin, M. R., and Uchimaya, M. (2007). Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. *J. Comput. Aided Mol. Des.* 21, 681–691. doi: 10.1007/s10822-007-9133-z

Sherman, W., Beard, H. S., and Farid, R. (2006a). Use of an induced fit receptor structure in virtual screening. *Chem. Biol. Drug Des.* 67, 83–84. doi: 10.1111/j.1747-0285.2005.00327.x

Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A., and Farid, R. (2006b). Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* 49, 534–553. doi: 10.1021/jm050540c

Shi, J., Wang, E., Milazzo, J. P., Wang, Z., Kinney, J. B., and Vakoc, C. R. (2015). Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat. Biotechnol.* 33, 661–667. doi: 10.1038/nbt.3235

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA Cancer J. Clin.* 69, 7–34. doi: 10.3322/caac.21551

Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., et al. (2010). Charmm general force field: a force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *J. Comput. Chem.* 31, 671–690. doi: 10.1002/jcc.21367

Vanommeslaeghe, K., and MacKerell, A. D. Jr. (2012). Automation of the CHARMM general force field (CGenFF) I: bond perception and atom typing. *J. Chem. Inf. Model.* 52, 3144–3154. doi: 10.1021/ci300363c

Vanommeslaeghe, K., Raman, E. P., and MacKerell, A. D. Jr. (2012). Automation of the charmm general force field (CGenFF) II: assignment of bonded parameters and partial atomic charges. *J. Chem. Inf. Model.* 52, 3155–3168. doi: 10.1021/ci3003649

Vilar, S., Quezada, E., Uriarte, E., Costanzi, S., Borges, F., Viña, D., et al. (2016). Computational drug target screening through protein interaction profiles. *Sci. Rep.* 6:36969. doi: 10.1038/srep36969

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). Swiss-model: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303. doi: 10.1093/nar/gky427

Xie, Z.-R., and Hwang, M.-J. (2012). Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. *Bioinformatics* 28, 1579–1585. doi: 10.1093/bioinformatics/bts182

Xie, Z.-R., and Hwang, M.-J. (2015). "Methods for predicting protein–ligand binding sites," in *Methods Mol. Biol.*, ed A. Kukol (New York, NY: Springer), 383–398.

Yu, W., He, X., Vanommeslaeghe, K., and MacKerell, A. D. Jr. (2012). Extension of the charmm general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *J. Comput. Chem.* 33, 2451–2468. doi: 10.1002/jcc.23067

Zhang, M., Pascal, J. M., and Zhang, J.-F. (2013). Unstructured to structured transition of an intrinsically disordered protein peptide in coupling Ca2+-sensing and SK channel activation. *Proc. Natl. Acad. Sci. U.S.A.* 110, 4828–4833. doi: 10.1073/pnas.1220253110

Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* 33, 2302–2309. doi: 10.1093/nar/gki524

Zhou, W., Wang, Y., Lu, A., and Zhang, G. (2016). Systems pharmacology in small molecular drug discovery. *Int. J. Mol. Sci.* 17:246. doi: 10.3390/ijms17020246

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Wu, Lou and Xie. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

REVIEW
published: 18 February 2020
doi: 10.3389/fchem.2020.00093

In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery

Lauro Ribeiro de Souza Neto[1†], José Teófilo Moreira-Filho[2†], Bruno Junior Neves[2,3†], Rocío Lucía Beatriz Riveros Maidana[1,4], Ana Carolina Ramos Guimarães[4], Nicholas Furnham[5], Carolina Horta Andrade[2*] and Floriano Paes Silva Jr.[1*]*

[1] LaBECFar – Laboratório de Bioquímica Experimental e Computacional de Fármacos, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil, [2] LabMol – Laboratory for Molecular Modeling and Drug Design, Faculdade de Farmácia, Universidade Federal de Goiás, Goiânia, Brazil, [3] Laboratory of Cheminformatics, Centro Universitário de Anápolis – UniEVANGÉLICA, Anápolis, Brazil, [4] Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil, [5] Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom

OPEN ACCESS

Edited by:
Teodorico Castro Ramalho, Universidade Federal de Lavras, Brazil

Reviewed by:
Hyun Lee, University of Illinois at Chicago, United States
Salvatore Guccione, University of Catania, Italy
Daiana Mancini, Universidade Federal de Lavras, Brazil

***Correspondence:**
Nicholas Furnham
nick.furnham@lshtm.ac.uk
Carolina Horta Andrade
carolina@ufg.br
Floriano Paes Silva Jr.
floriano@ioc.fiocruz.br

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: 25 October 2019
Accepted: 30 January 2020
Published: 18 February 2020

Citation:
de Souza Neto LR, Moreira-Filho JT, Neves BJ, Maidana RLBR, Guimarães ACR, Furnham N, Andrade CH and Silva FP Jr (2020) In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. Front. Chem. 8:93. doi: 10.3389/fchem.2020.00093

Fragment-based drug (or lead) discovery (FBDD or FBLD) has developed in the last two decades to become a successful key technology in the pharmaceutical industry for early stage drug discovery and development. The FBDD strategy consists of screening low molecular weight compounds against macromolecular targets (usually proteins) of clinical relevance. These small molecular fragments can bind at one or more sites on the target and act as starting points for the development of lead compounds. In developing the fragments attractive features that can translate into compounds with favorable physical, pharmacokinetics and toxicity (ADMET—absorption, distribution, metabolism, excretion, and toxicity) properties can be integrated. Structure-enabled fragment screening campaigns use a combination of screening by a range of biophysical techniques, such as differential scanning fluorimetry, surface plasmon resonance, and thermophoresis, followed by structural characterization of fragment binding using NMR or X-ray crystallography. Structural characterization is also used in subsequent analysis for growing fragments of selected screening hits. The latest iteration of the FBDD workflow employs a high-throughput methodology of massively parallel screening by X-ray crystallography of individually soaked fragments. In this review we will outline the FBDD strategies and explore a variety of *in silico* approaches to support the follow-up fragment-to-lead optimization of either: growing, linking, and merging. These fragment expansion strategies include hot spot analysis, druggability prediction, SAR (structure-activity relationships) by catalog methods, application of machine learning/deep learning models for virtual screening and several *de novo* design methods for proposing synthesizable new compounds. Finally, we will highlight recent case studies in fragment-based drug discovery where *in silico* methods have successfully contributed to the development of lead compounds.

Keywords: fragment-based, drug discovery, lead discovery, *in silico* methods, machine learning, *de novo* design, optimization, hot spot analysis

INTRODUCTION

Fragment-Based Drug Discovery

Since the inception of fragment-based drug discovery (FBDD) over 20 years ago it has become an established technology used in both industry and academia (Hubbard, 2015). FBDD offers an attractive approach for effectively exploring the chemical space for binding a target protein. In conventional high-throughput screening (HTS) campaigns, large libraries of often complex compounds are screened for activity against a target (Hall et al., 2014). In contrast, FBDD use relatively small libraries of low complexity compounds representing fragments of larger more drug-like compounds. By reducing the complexity of the chemicals screened more of the potential binding sites of a target protein can be explored through the binding promiscuity of the fragments (Thomas et al., 2017). Where fragments do bind, albeit with lower potency than the drug-like molecules of HTS, they offer good starting points to design larger higher affinity binders using knowledge of the protein structure as a template to generate compounds with greater ligand efficiency (improved per atom binding energy to the target). This bottom-up approach means that a greater range of chemical space can be explored, leading quickly to higher affinity lead compounds with greater specificity (Patel et al., 2014).

FBDD projects require relatively lower investments in research and development (R&D) than HTS (Davis and Roughley, 2017). An example is the discovery of vemurafenib (Zelboraf^{TM}), the first fragment-derived drug, which moved relatively very quickly (6 years) between the phases of R&D pipeline before reaching Food and Drug Association (FDA) approval (Erlanson et al., 2016). Thus, FBDD provides attractive opportunities for the drug discovery field.

Output of Structure-Enabled Fragment Screening Campaigns

FBDD workflows are multi-step starting with target selection and protein isolation and followed by an initial screen of the fragment library using biophysical techniques such as nuclear magnetic resonance (NMR), surface plasmon resonance (SPR), thermal-shift assay, microscale thermophoresis (MST), mass spectrometry, and others. For fragments which show evidence of binding, a further step of hit validation and characterization occurs principally using X-ray crystallography (Verdonk and Hartshorn, 2004). Using hit characterization, an iterative cycle of fragment development can occur employing a range of *in silico* and experimental techniques. Advances in this protocol try to compress the process by combining the initial fragment screen with the hit characterization. This has been implemented in a high throughput FBDD platform called XChem located at the United Kingdom's national synchrotron the Diamond Light Source (Cox et al., 2016). It uses the ability to produce and handle a large number of crystals of the target protein to screen the fragment library by soaking each individual crystal with a fragment and then using X-ray crystallography to determine which fragments have bound and where. Though this high throughput technique often provides multiple hits, care needs to be taken in interpreting the significance of the hit. Promiscuous fragments may bind parts of the protein which are not involved in the protein function and therefore are unlikely to yield a successful inhibitor. Additionally, as the fragments are by their very nature weak binders and X-ray crystallography being a sensitive technique, observed binding events might be transient and not easily reproducible. It is therefore important to confirm hits with orthogonal structural (e.g., NMR), biophysical techniques (SPR, MST, etc.) or *in vitro* biological assays.

Fragment Libraries

A crucial step in FBDD process is in the development and choice of the fragment library used in the screening campaign. Several fragment libraries have been developed that exploit certain properties or chemistries. An example of a fragment library is the Diamond-SGC Poised Library (DSPL) (Cox et al., 2016). This has been developed for use with high-throughput XChem platform and consists of around 760 fragments that have been selected to contain at least one functional group that is open to rapid, cheap follow-up synthesis using robust well-characterized reactions (poised) and maximizing chemical diversity. Other fragment libraries optimize other properties such as solubility, 3D traits or based on subsets of existing drugs and related molecules such as natural products (Schuffenhauer et al., 2005). The fragment libraries generally share similar properties of "Rule of 3" compliant i.e., less than 300 Da molecular weight, 3 or less hydrogen bond donors, 3 or less hydrogen bond acceptors and CLogP no more than 3. In addition, they are soluble in dimethyl sulfoxide (DMSO) or phosphate buffered saline. Fragment libraries generally tend to be $<$1,000 fragments, which is significantly less than the many millions of compounds screened in high-throughput and high content screening campaigns (Trevizani et al., 2017).

Fragment Expansion Strategies

Once the fragment screen has been completed and hits characterized, the next step is the challenge of expanding these fragments to generate larger molecular entities with high binding affinity and demonstrating inhibition activity. There are several strategies that can be followed (Lamoree and Hubbard, 2017) (**Figure 1**). One option is to use expert medicinal chemistry advice to design and synthesize larger molecules based on the protein and the fragment pose. Another approach is to define vectors along the fragment molecule based on the steric hindrance of the protein target in which the fragment can be expanded. The fragment is then searched for within a large library of synthesizable (or purchasable) molecules which are bigger by between one and three heavy atoms along the identified vectors. These expanded fragments can be synthesized and soaked/co-crystallized and re-screened by X-ray crystallography. Expanded fragments that show improved binding can be further extended or structurally modified using the same process, with this cycle continuing until a larger high-affinity binding entity is reached.

An alternative to this "small steps" approach is to try to get a larger higher affinity binding molecule in a single step. This can be achieved by having an *in silico* method using the fragment in a substructure search of a large purchasable compound library (e.g., Zinc15), and to virtually screen the results using the pose

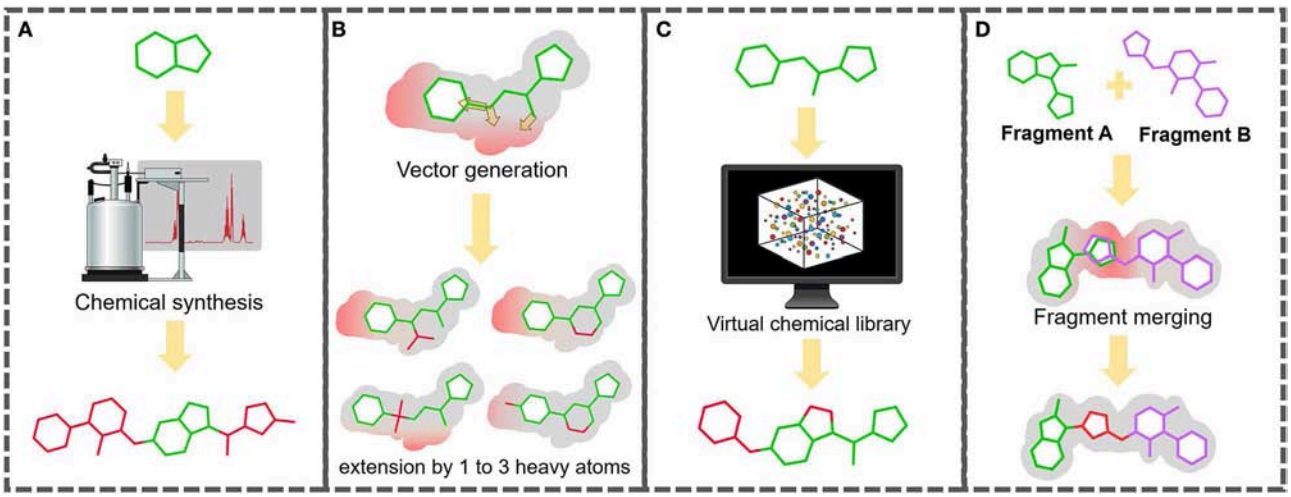


FIGURE 1 | Multiple routes to expanding fragment to more drug-like molecule with improved binding affinity. **(A)** Traditional medicinal chemistry route: knowledge-based design and synthesis. **(B)** "Small steps" route: successive cycles of extension of the fragment hit by 1–3 heavy atoms through vectors defined by high-resolution structural characterization methods, such as X-ray crystallography. **(C)** "Large leaps" or "SAR by catalog" route: from fragment to rule of 5 compliant molecules using virtual screening of commercial compound libraries. **(D)** Fragment merging route: bridging two overlapping fragments bound at neighbor sites. Regardless of the route, expanded fragments should be checked for biological activity using *in vitro*, *ex vivo,* or *in vivo* assays.

of the fragment to dock the molecules and rank them based on docking parameters (Trevizani et al., 2017). The top-ranking virtual screening hits can then be co-crystallized and well as evaluated *in vitro* and *in vivo*. A final expansion option is to link or merge fragments that hit near to each other or within the same site (Davis and Roughley, 2017). The combined fragments can then be further expanded using the approaches described previously. It is vital that as expansion progresses *in vitro* and *in vivo* assays are conducted to asses activity of the new molecules.

In the next section, we will discuss in depth the main optimization approaches used for a fragment structurally characterized in a binding site of its target. Further sections will describe existing software tools or modeling techniques (e.g., machine learning) employed for taking a fragment hit thorough the path for becoming a lead compound—a process known as fragment-to-lead (F2L)—for drug development and conclude by presenting case studies where *in silico* strategies have been successfully utilized to support the F2L optimization process.

FRAGMENT OPTIMIZATION APPROACHES

After the hit identification in a FBDD campaign, the fragment moves forward to the optimization phase. This optimization takes into account the structural characteristics of the ligand as well as its binding site. The principle in using fragments relies on the premise that these molecular entities are more efficient ligands compared to drug-like molecules, and their structures can be further optimized more efficiently. In fact, this constitutes one of its many advantages. As small entities, molecular fragments can be iteratively optimized to show a better pharmacokinetic profile in the later development stages. Drug-like molecules may contain functional groups that contribute poorly to protein binding or, in some cases, can even disrupt the protein-ligand interaction. On the other hand, fragments often form high-quality interactions able to more easily bind to the protein target, translating to a greater number of hits. **Figure 2** depicts schematically this concept.

Another advantage of FBDD is the potential for faster hit progression through the campaign, since the fragments are usually structurally simple and many follow-up compounds can be easily purchased from commercial databases (e.g., MolPort, ZINC15, and ChemBridge) instead of being synthesized. Another important characteristic often used to defend this approach is the high hit rates. In this sense, high hit rates means that the FBDD yields relatively more hits in comparison to the traditional methods such as HTS (Coutard et al., 2014; Mondal et al., 2015). This is due the inversely related nature between molecular complexity and the binding probability (Hann et al., 2001). Other advantages includes the more efficient chemical space sampling (Coutard et al., 2014; Mondal et al., 2015) and the relative low cost to implement the FBDD, as it can be seen from comparing the usual size of the HTS library (thousands of compounds) with fragment libraries (hundreds of compounds) (Macarron et al., 2011).

Assessment of the interactions between the fragment and its binding site should be carefully performed for further identification of synthetically accessible vectors on the ligand. Although x-ray crystallography data is a valuable technique in fragment optimization, is important to keep in mind that the observed structural data only represents a snapshot of the system under investigation. It's been known that the ligand affinity can be affected by the structural protein dynamics without changing the ligand-binding interface (Matias et al., 2000; Seo et al., 2014). This complex dynamic environment (Henzler-Wildman and Kern, 2007; Boehr et al., 2009) can affect small and weak ligands as

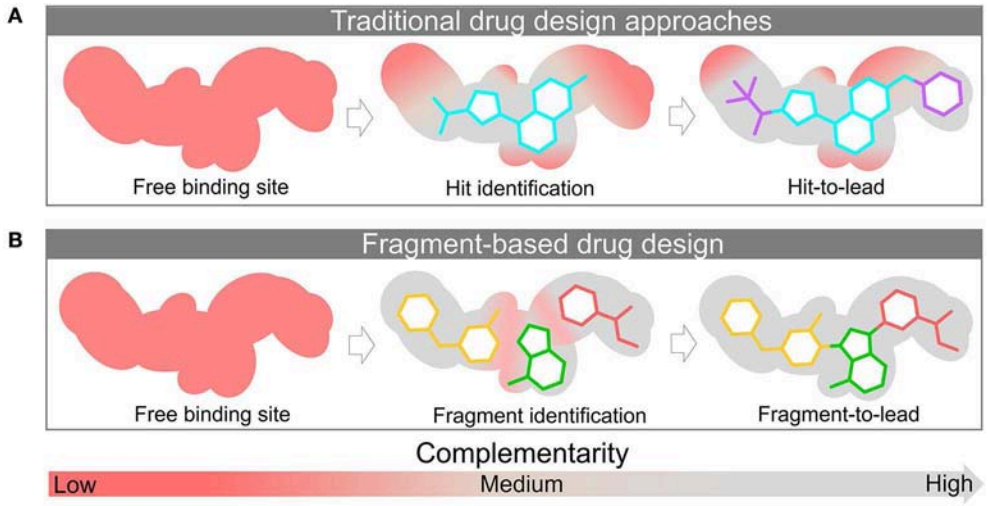


FIGURE 2 | Discovery and structural-optimization of drug-like molecules **(A)** and fragments **(B)** using protein target information. The surface represents the binding site. The red and gray colors represent the level of complementarity of ligand with the active site. Pockets with low complementarity with ligand are colored in red; pockets with high complementarity with ligand are highlighted in gray.

fragments. With this in mind, many methods can additionally be used to guide the fragment identification/optimization either providing complementary data (e.g., thermodynamic data) or acting as orthogonal approaches (Ciulli, 2013a). These methods are mostly biophysical (Shuker et al., 1996; Lo et al., 2004; Navratilova and Hopkins, 2010; Pedro and Quinn, 2016) and their use has some advantages such as, direct measurement of the binding, detection of small ligands with low affinity, and not needing any prior information about the protein function (Ciulli, 2013b). Despite the supremacy of biophysical methods, biochemical approaches are increasingly being used (Godemann et al., 2009; Boettcher et al., 2010; Mondal et al., 2015) in FBLD.

In addition to orthogonal and complementary methods, the ligand efficiency (LE) or one of its related metrics should ideally be used to keep track of the quality of follow-up ligands as they progress through the iterative optimization cycle. Some of these parameters are described below.

- **L**igand **E**fficiency (**LE**) (Hopkins et al., 2004; Nissink, 2009; Davis and Roughley, 2017) = ΔG/HAC [A];
- **B**inding **E**fficiency **I**ndex (**BEI**) (Abad-Zapatero, 2013) = pKi/MW [B];
- **P**ercentage **E**fficiency **I**ndex (**PEI**) (Abad-Zapatero, 2017; Davis and Roughley, 2017) = % inhibition/MW [B];
- **S**urface-binding **E**fficiency **I**ndex (**SEI**) (Abad-Zapatero, 2013) = pIC_{50}/(TPSA [C]);
- **Lip**ophilic **E**fficiency (**LipE/LLE**) (Shultz, 2013) = pIC_{50}–cLogP;
- **S**ize-**I**ndependent **L**igand **E**fficiency (**SILE**) (Nissink, 2009) = $\Delta G/HAC^{1-x}$;
- **L**igand **E**fficiency-Dependent **Lip**ophlicity (**LELP**) (Davis and Roughley, 2017) = logP/LE.

[A]Heavy Atom Count; [B]Molecular Weight; [C]Topological Polar Surface Area;

For the sake of brevity these metrics will not be further discussed and we recommend the references above for a deeper understanding. The structural complexity of the protein makes larger, more complex and less efficient molecules less likely to bind. This is one of the main reasons why fragment libraries often yield more hits when compared to a drug-like molecule commonly used in HTS (Hann et al., 2001). The use of fragments is a bottom-up approach, starting from less complex molecules with greater binding efficiency and ending up with a larger optimized molecule. As already highlighted, there are three main strategies that can be employed to optimize a ligand found bound in its target surface: linking, merging and growing (**Figure 3**). The next sections are dedicated to discussing in more depth each of them.

Growing

Fragment growing (**Figure 3A**) is the strategy most commonly employed during FBLD campaigns. As the name suggests, it consists of modifying the fragments to increase their size. Conceptually this approach is identical to the traditional compound modification methods employed in the optimization of hits from HTS campaigns. This modification occurs through the addition of groups.

A recent paper published by Strecker et al. (2019) is an example of how the growing strategy can be used to improve bind affinity. Using computer-aided drug design (CADD) and synthesis, the authors explored small structural modifications in a previously (PDB: 3U0X) identified compound (**1**) (K_i = 800 μM).

These studies showed that a modification of a fragment phenyl moiety to a naphthyl allowed two new simultaneous π-π interactions, a parallel-displaced with Trp300 and an edge-to-face with His233. This minor modification led to a compound (**2**) with a 3-fold improved binding affinity (K_i = 271 μM) (**Figure 4**).

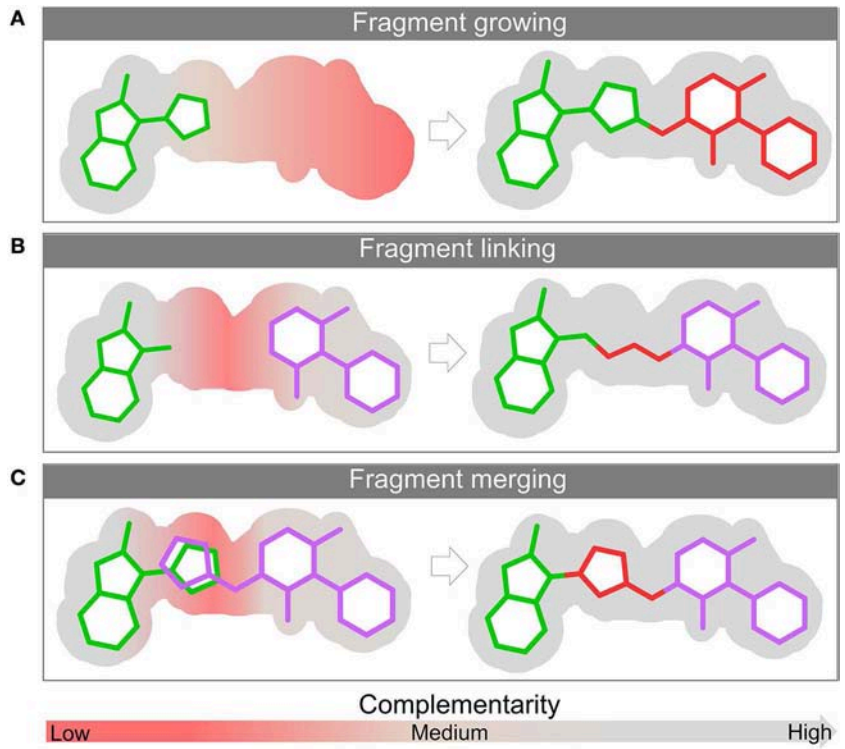


FIGURE 3 | Fragment optimization approaches: fragment growing **(A)**, fragment linking **(B)**, and fragment merging **(C)**. The surface of the binding site is depicted in gray. The red and gray colors represent the level of complementarity of ligand with the active site. Pockets with low complementarity with ligand are colored in red; pockets with high complementarity with ligand are highlighted in gray.

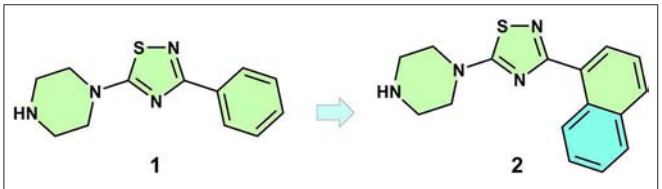


FIGURE 4 | Hit to lead progression of an initial fragment **(1)** to a compound **(2)** with improved affinity.

This example highlights the use of optimal growth vectors to introduce a rigid group, which led to an increased binding affinity. Alternatively, introduction of a moiety with increased number of rotatable bonds could impact negatively—due to the entropic penalty—in the affinity. Although this optimization approach can be computationally aided without further structural data, small modifications—as in the case of the hypothetical flexible moiety addition—can led to great changes in binding mode. When growing fragments is the chosen approach, structural data can be decisive to avoid misinterpretation.

Linking

Fragment linking (**Figure 3B**) describes the process of joining two non-competitive fragments (i.e., fragments that bind in two different sub-pockets of the binding site) with a chemical linker or spacer. Although conceptually simple, linking fragments is perhaps the most challenging strategy to implement. Although fragment linking is the most attractive approach in terms of rapid improvement of potency, the design of a linker with suitable flexibility while not disturbing the original binding modes of the fragments, makes it one of the most challenging optimization approaches.

As previously discussed, the introduction of flexible moieties affects these compounds properties and an optimal orientation should always be pursued. In fact, varying the degree of rigidity of a linker for the purpose of conformational restriction of the linked product can be used as a strategy for linker optimization, as it can be seen in Chung and colleagues work (Chung et al., 2009). This work shows how a conversion of oxime linkers into monoamine and diamines interferes with the rigidity and its impact on binding.

Although often neglected, the impact on the ADMET properties should also be taken in consideration. In the case of the linker, that usually adds rotatable bonds to the system (Ichihara et al., 2011; De Fusco et al., 2017), this modification can lead to poor PK features, like low permeability (Veber et al., 2002).

Merging

This strategy (**Figure 3C**) can be used in cases where two distinct fragments partially occupy the same region, or when two binding sites have regions in common and therefore their ligands are

partially competitive with respect to the site. In such cases the overlapping parts form a nucleus where dissimilar parts come together. In a recent example, a gain of 2 orders of magnitude in potency was achieved for an inhibitor of flavin-dependent monooxygenase (EthA) transcriptional repressor (EthR) (Nikiforov et al., 2016) where the existence of overlapping groups within fragments bound to EthR allowed the use of merging as an optimization strategy.

Although not always possible, merging is a simpler strategy than linking, as there is no need to design a spacer that joins fragments together (Xu et al., 2017; Miyake et al., 2019). As also seen in this example, like linking (Davis and Roughley, 2017), this approach has the drawback of relying on high-quality structural data to go further in the optimization process.

Therefore, merging is an approach related to the "molecular hybridization" strategy, a long-consolidated approach in medicinal chemistry for designing new compounds with improved potency through the fusion of other active compound structures.

IN SILICO STRATEGIES FOR F2L OPTIMIZATION

Hot Spots Analysis and Pocket Druggability Prediction

Hot spots analysis is an important tool for structure-based F2L that allows the prediction of the small regions of the binding sites containing residues mostly contributing to the binding free energy (Cukuroglu et al., 2014). Once a fragment hit is experimentally identified, the hot spots analysis can be used to map the subsites around the fragment hit using small organic probes, driving the optimization into higher-affinity ligands (Hall et al., 2012).

One of the most used methods of hot spot analysis is the FTMap web server (Kozakov et al., 2015). This algorithm places 16 small organic probe molecules of different shape, size, and polarity on the protein surface to find favorable positions for each probe. Then, each probe type is clustered and overlapping clusters of different probes, called consensus sites (CSs), represent the hot spots. The consensus sites are ranked by the number of probe clusters, and the main hot spot is, generally, where the fragment hit binds and secondary hot spots are used to extend the fragment in the best direction (Hall et al., 2012; Ngan et al., 2012; Kozakov et al., 2015).

As an example, we used the FTMap server for predicting the hot spots for the oncogenic B-RAF kinase, the target of the first marketed drug from fragment-based drug design, vemurafenib (Bollag et al., 2012). **Figure 5A** shows the fragment hit experimentally bound to B-RAF kinase (PDB ID: 2UVX) (Donald et al., 2007), and the predicted hot spots around this fragment (shown in yellow dots) using the FTMap server. In **Figures 5B–D**, the iterative process of growing the fragment hit led to the discovery of the drug vemurafenib (PDB ID: 3OG7) (Bollag et al., 2010) with the hot spots shown in yellow dots. Although hot spot analysis was not used in the F2L process of vemurafenib, the results here showed that the predicted hot spots overlap the grown portions of vemurafenib.

During fragment screening, the fragment hits can bind in different sites of the protein (Giordanetto et al., 2019). If the binding site is not well-defined, the researchers can use the pocket druggability prediction to move forward in F2L with the most druggable site, capable to accommodate ligands orally bioavailable (Schmidtke and Barril, 2010; Hussein et al., 2015). There are many available methods for predicting pocket druggability and these are well-described and reviewed elsewhere (Barril, 2013; Abi Hussein et al., 2017).

SAR by Catalog

One fast and cheap way in F2L optimization is the SAR by catalog approach (Hall et al., 2017). This approach relies on the search of analogs of *in-house* or commercial databases that can be purchased or rapidly accessed for testing (Schulz et al., 2011). This process can use the fragment hit features for similarity, ligand-based pharmacophores, shape-based, fingerprints (Rogers and Hahn, 2010; Riniker and Landrum, 2013; Alvarsson et al., 2014), and substructure searches to find suitable compounds (Hubbard and Murray, 2011; Andrade et al., 2018). Some databases often used for SAR by catalog are ZINC (Sterling and Irwin, 2015), MolPort (https://www.molport.com), Mcule (https://mcule.com/), and eMolecules (https://www.emolecules.com) that contains collections of commercially available compounds. The databases Enamine (https://enaminestore.com), ChemDiv (http://www.chemdiv.com/) and ChemBridge (https://www.chembridge.com) are direct suppliers.

SAR by catalog approach only retrieves similar compounds or superstructures of the fragment hit. Thus, other filters should be applied to filter compounds with more optimized properties. These filters are molecular docking, ADMET, machine learning models, aqueous solubility, among others, and will be discussed later in this review.

Molecular Docking

Molecular docking is a computational approach used to predict the position, orientation, and the binding scores of small molecules to proteins (Torres et al., 2019). Hence, as the F2L process is commonly addressed as a combinatorial problem, molecular docking is a method that can be used in combination with other approaches to enhance the F2L process, and to increase the chances to convert a fragment hit into higher affinity ligands. The SAR by catalog approach in combination with molecular docking, for example, can be used to select compounds that maintain the fragment hit binding mode while the binding energy is optimized. Moreover, the number of generated optimized fragments can exceed the number that can be tested experimentally. Thus, applying molecular docking, large compound datasets are efficiently assessed using SAR by catalog, and a small subset of most promising compounds can be selected by binding modes and scores for experimental testing (Grove et al., 2016).

To overcome the problem that SAR by catalog has the limitation to cover only the finite chemical space of commercially available compounds (Hoffer et al., 2018), it is possible to

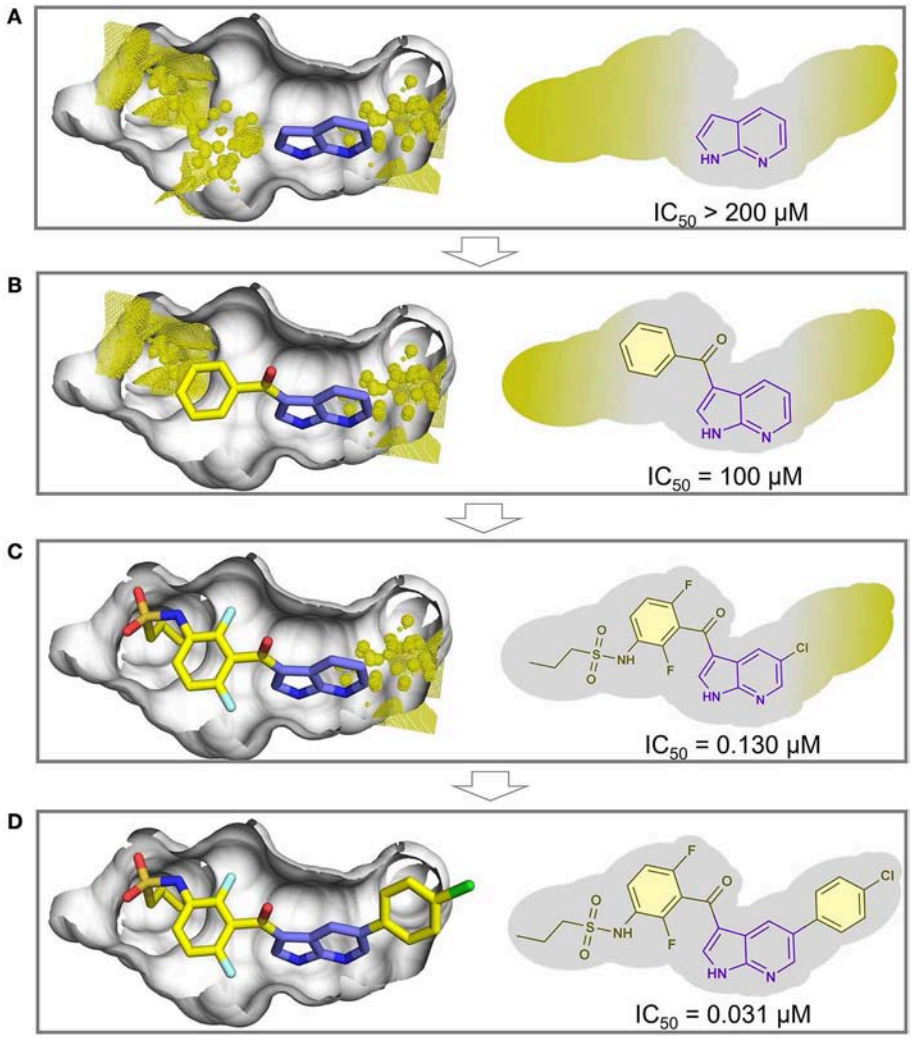


FIGURE 5 | Example of a hot spot analysis using FTMap web server of the oncogenic B-RAF kinase, the target of the first marketed drug from fragment-based drug design, vemurafenib. The surface of the binding site is depicted in gray. **(A)** (PDB ID: 2UVX) the fragment hit (carbon atoms in purple sticks) and the predicted hot spots (yellow dots and surface). **(B–D)** The iterative growing process of vemurafenib (PDB ID: 3OG7) overlapping the predicted hot spots (the carbon atoms of the fragment hit portion is shown in purple sticks and carbon atoms of the grown portions in yellow sticks).

generate virtual catalogs with analogs to hit fragments that can be easily synthesized, astronomically increasing the number of possible compounds. Then, a docking-based virtual screening can be applied to prioritize compounds for experimental evaluation (Rodríguez et al., 2016; Männel et al., 2017).

Another scenario in F2L is when the co-crystallization of a fragment hit commonly fails and no structural information about the binding mode is available. In these cases, alternative strategies for F2L process are required where the binding mode of a fragment can be predicted using molecular docking calculations (Kumar et al., 2012; Chevillard et al., 2018; Erlanson et al., 2019) on high-quality three-dimensional structures of the target in apo form or bound to other ligands. When neither of the latter are available, a theoretical model of the target protein can be obtained by homology modeling methods.

However, there are concerns about fragment docking in the scientific community. The assumption is that fragments, as low molecular weight compounds, are weak binders and promiscuous in binding modes, and consequently, the fragment docking implies in incorrect predictions of the binding modes. Also, there is a concern that scoring functions of the docking programs are parameterized to drug-like ligands, being inaccurate to differentiate native and other low-energy poses (Chen and Shoichet, 2009; Wang et al., 2015; Grove et al., 2016). To overcome these concerns, there are studies demonstrating no significant difference in docking performance between fragments and drug-like ligands (Verdonk et al., 2011; Joseph-mccarthy et al., 2013). They showed that molecular weight is not the principal parameter for docking performance, instead, for high LE compounds the docking performance fared better for both

fragments and drug-like ligands (Verdonk et al., 2011; Kumar et al., 2012).

When available, the use of experimental structural information data can be used to support and improve docking performance. These data are used in docking programs including distance constraints, pharmacophore constraints, shape-based constraints, similarity or substructure overlap, interaction fingerprints, hydrogen-bond constraints, and others (Verdonk et al., 2011; Erlanson et al., 2019; Jacquemard et al., 2019).

Similarly to hot spot analysis, molecular docking can also be used to discover secondary binding pockets and guide the F2L process (Männel et al., 2017).

Machine Learning (ML) and Deep Learning (DL) Models

A large variety of F2L approaches use structure-based methods to optimize fragments into high-affinity ligands taking into consideration the steric and electronic constraints within binding pockets of the target of interest (Schneider and Fechner, 2005). However, the optimized compounds generated constantly present drawbacks of poor synthetic feasibility and/or undesirable biological properties, including absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties (Yang et al., 2019b). In the last years, novel ligand-based methods, including machine learning (ML) models, have been used for F2L campaigns. ML models are statistical methods that present the capacity to learn from data without the explicit programming for this task, and then, make a prediction for new compounds (Mak and Pichika, 2019). The increase of storage capacity and the size of the datasets available, coupled to advances in computer hardware such as graphical processing units (GPUs) (Gawehn et al., 2018), provided means to move theoretical studies in ML to practical applications in drug discovery (Vamathevan et al., 2019).

The ML algorithms are widely used to construct quantitative structure-activity relationship (QSAR) models, able to find mathematical correlations between molecular features and compound activity/property, and this correlation can be categorical (active, inactive, toxic, nontoxic, etc.) or continuous (pIC_{50}, pEC_{50}, Ki, and others) by means of classification or regression techniques (Tropsha, 2010; Cherkasov et al., 2014). Thus, machine learning-based QSAR models can be constructed for biological activity, ADMET properties, solubility and synthetic feasibility, among other endpoints, and applied after fragment optimization, with aforementioned methods, in a cascade virtual screening for filtering compounds with the desired activities and properties (**Figure 6**) (Braga et al., 2014; Neves et al., 2018; Pérez-Sianes et al., 2018).

More recently, a subfield of ML called deep learning (DL) which utilizes artificial neural networks to learn from a large amount of data have been used to resolve complex problems (Mak and Pichika, 2019). DL models are not only able to learn from a dataset and to make predictions for new data but are also able to generate new data instances through a constructive process (Schneider, 2018). In this context, there has been a rising interest in using DL generative and predictive models for F2L optimization (Olivecrona et al., 2017; Gupta et al., 2018).

For this task, a combination of DL architectures is used and in many cases, generative DL models based on recurrent neural networks (RNNs) are trained on the simplified molecular input line entry system (SMILES) representation of compounds from large databases (DrugBank, ChEMBL, etc.) to learn the syntax of SMILES language and the chemical space distribution (Olivecrona et al., 2017). After training, the models are able to generate new strings that are new SMILES, corresponding to new compounds (Segler et al., 2018). Then, the transfer learning (TL) can be used to fine-tuning the model and generate compounds related to a fragment hit. As the name suggests, TL learns and transfers the information from an old source to a new application (Yang et al., 2019b). The aim of this integrative approach is to learn general features from a big dataset and, then, retrain the model focusing on a smaller dataset such as fragment hits, for F2L purposes (**Figure 7**) (Gupta et al., 2018; Segler et al., 2018). Gómez–Bombarelli et al. used variational autoencoder (VAE) to encode SMILES into a continuous latent-space, then a separate multilayer perceptron trained to predict several properties on the latent space was applied to generate new molecules with the desired properties. After this, a decoder was used to retrieve the molecules on the latent space into SMILES (Gómez-Bombarelli et al., 2018). Handling these DL methods in a multidimensional way, fragment hits can be optimized automatically taking into consideration several parameters such as bioactivity, solubility, synthetic feasibility, and ADMET properties, generating new compounds with optimized values for these parameters (**Figure 7**) (Olivecrona et al., 2017; Ramsundar et al., 2017; Gómez-Bombarelli et al., 2018; Harel and Radinsky, 2018; Li et al., 2018; Merk et al., 2018; Polykovskiy et al., 2018; Popova et al., 2018; Putin et al., 2018; Awale et al., 2019; Vamathevan et al., 2019).

De novo Design

The *de novo* approach looks for new chemical entities from scratch within a structurally defined binding site (Schneider and Clark, 2019). These entities are generated out of building blocks, either by growing from an initial fragment or by linking two or more non-overlapping fragments (Dey and Caflisch, 2008; Kumar et al., 2012). Since their arise, *in silico* methods have played an important role in FBDD (Kumar et al., 2012).

Software for Building New Compounds Within a Structurally-Defined Binding Site

De novo design software takes advantage of a known binding mode of a fragment, described experimentally or computationally, to propose modified analogs with improved binding affinities. The LUDI (Bohm, 1992) program was one of the first programs developed for *de novo* design. It calculates the interaction sites, maps the molecular fragments, and connects them using bridges, using an empirical scoring. Considering the vast chemical space, evolutionary algorithms are widely used (Srinivas Reddy et al., 2013). In this context, the program GANDI (Dey and Caflisch, 2008) connects pre-docked fragments with linker fragments using a genetic algorithm and a tabu search.

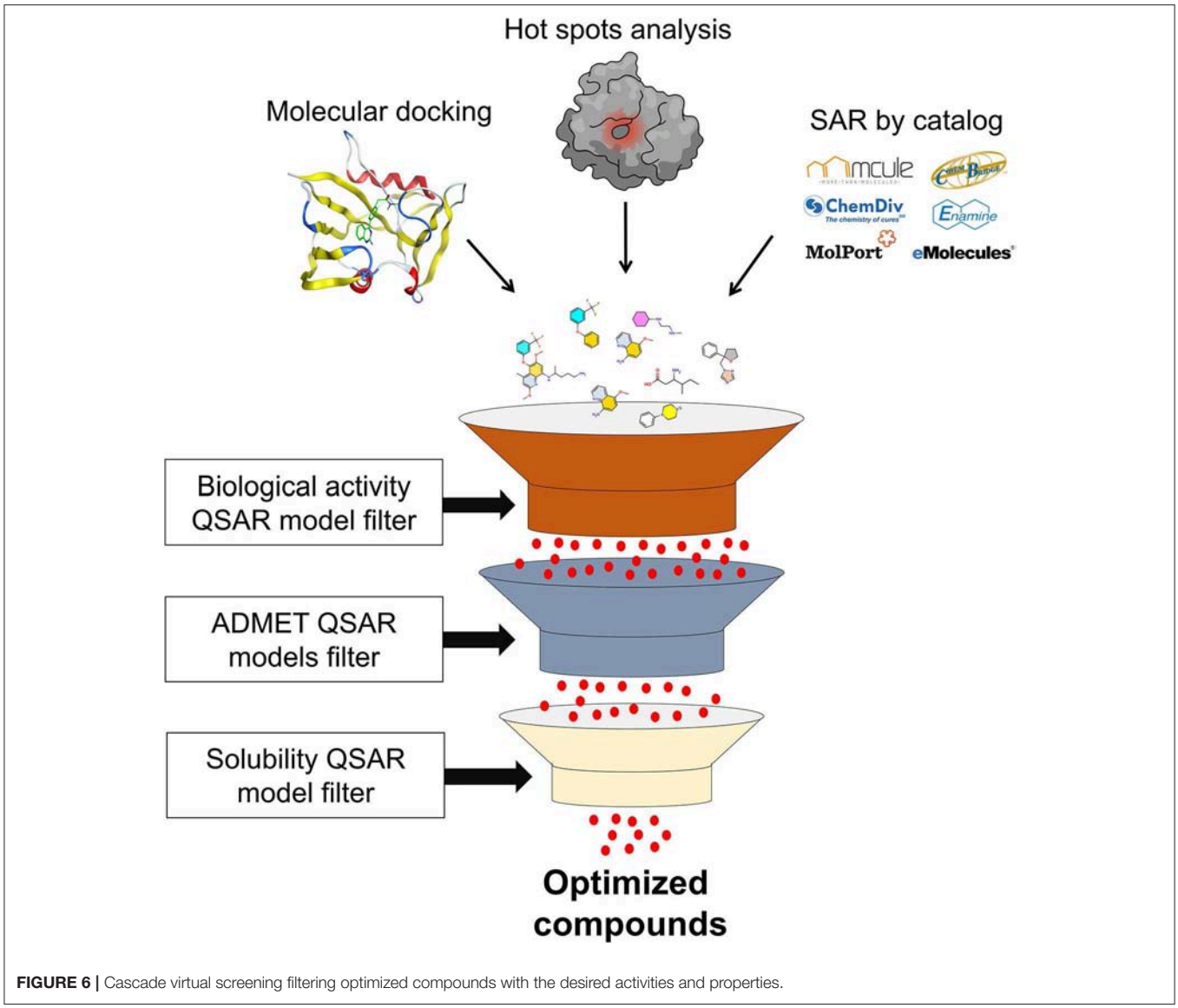


FIGURE 6 | Cascade virtual screening filtering optimized compounds with the desired activities and properties.

The scoring function is a linear combination of force-field binding energy and similarity measures. BREED (Pierce et al., 2004) is a computational method for merging fragments that is widely used. It aligns the 3D coordinates of two ligands and recombines the fragments or substructures into the overlapping bonds to generate new hybrid molecules in a strategy called fragment shuffling. LigBuilder (Wang et al., 2000; Yuan et al., 2011) is a program that uses a genetic algorithm to build up the ligands using a library of organic fragments. It contemplates the growing and linking approach. The 2.0 version includes the synthesis accessibility analysis through a chemical reaction database and retro-synthetic analysis. Autogrow (Durrant et al., 2009, 2013) is another growing approach algorithm that builds a fragment upon a "core" scaffold. The fragment is docked to the receptor. A genetic algorithm evaluates the docking score to select the best population which forms the subsequent generation. The last version considers the synthetic accessibility and the druggability. The program ADAPT (Pegg et al., 2001; Srinivas Reddy et al., 2013) applies a genetic algorithm which uses molecular interactions and docking calculations as a fitness function to reduce the search space. The initial sets of compounds are iteratively built until it reaches the predefined target value.

Prediction of ADMET Properties of New Compounds

The ADMET properties and synthetic accessibility (SA) constitutes the secondary constraints whereas primary constraints are geometric and chemical constraints derived from the receptor or target ligand(s) and internal constraints to the geometry and chemistry of the lead compound being constructed. Issues with these points result in the majority of clinical trial failures (Dong et al., 2018). Numerous software and web platforms were developed to predicted ADMET parameters but presented limitations due to narrow chemical

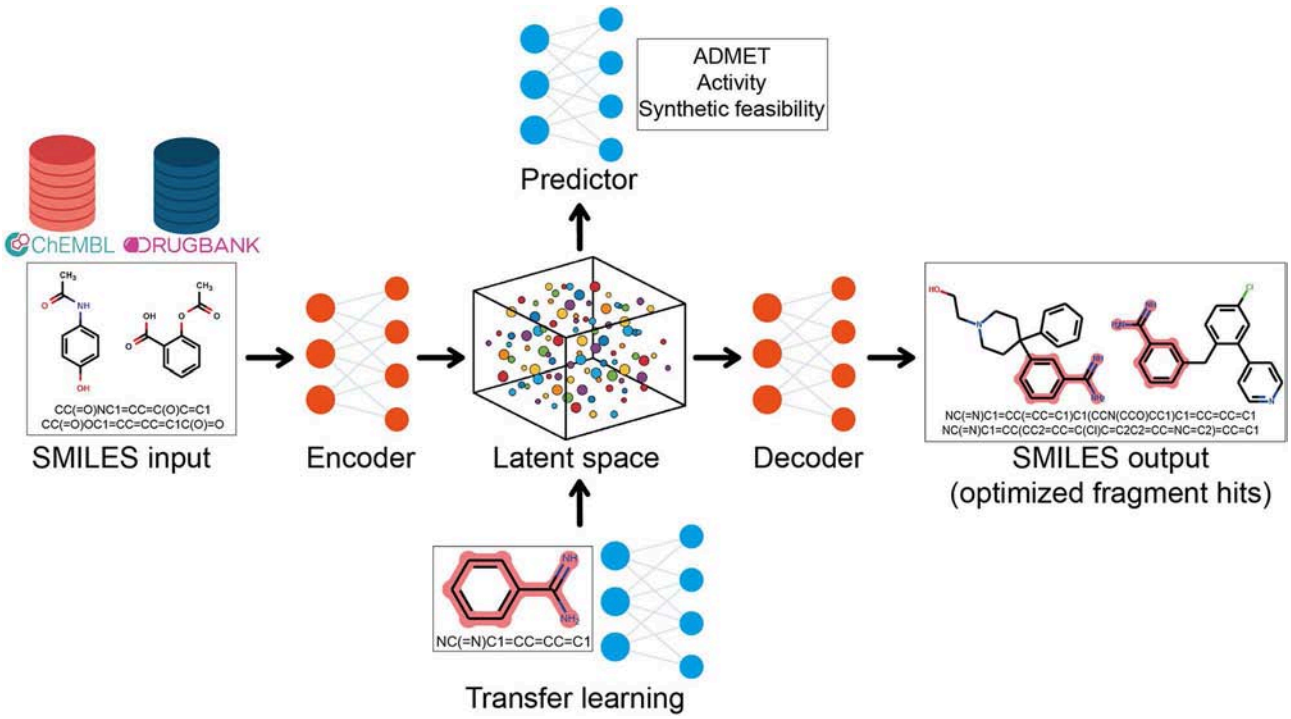


FIGURE 7 | Representation of the integrative approach of generative and predictive deep learning models and transfer learning for fragment-to-lead optimization.

space coverage or expensive prices (Cheng et al., 2012). Recent works predominantly rely on ML methods, like random forest (RF), support vector machine (SVM), and tree-based methods (Ferreira and Andricopulo, 2019). The vNN Web Server for ADMET predictions (Schyman et al., 2017) is a publicly available online platform to predict ADMET properties and to build new models based on the k-nearest neighbor (k-NN), which rest on the premise that compounds with similar structures have similar activities. vNN uses all nearest neighbors that are structurally similar to define the model's applicability domain. The similarity distance employed is Tanimoto's coefficient. The platform allows running pre-build ADMET models, and to build and run customized models. Those models assess cytotoxicity, mutagenicity, cardiotoxicity, drug-drug interactions, microsomal stability, and likelihood of causing drug-induced liver injury. Like all machine learning methods, the lack of training data is a limitation.

Pred-hERG (Braga et al., 2015; Alves et al., 2018) is a web app that allows users to predict blockers and non-blockers of the hERG channels, and important drug anti-target associated with lethal cardiac arrhythmia (Mitcheson et al., 2000). The current version of the app (v. 4.2) was developed using ChEMBL (Willighagen et al., 2013) version 23, containing 8,134 compounds with hERG blockage data after curation, using robust and predictive machine learning models based on RF. This app is publicly available at http://labmol.com.br/predherg/.

In admetSAR 2.0 (Cheng et al., 2012; Yang et al., 2019a) tool, the predictive models are built using RF, SVM and kNN algorithms. It presents 27 endpoints and also includes eco-toxicity models and an optimization module called ADMETopt that optimize the query molecule by scaffold hopping based on ADMET properties. The ADMETlab platform (Dong et al., 2018) performs its evaluations based on a database of collected entries and assess drug-likeness evaluation, ADMET prediction, systematic evaluation and database/similarity searching. It uses 31 endpoints applying RF, SVM, recursive partitioning regression (RP), naive Bayes (NB), and decision tree (DT).

SwissADME tool (Daina et al., 2017) uses predictive models for physicochemical properties, lipophilicity and water solubility. It also analyses pharmacokinetics models as BBB permeability, gastrointestinal absorption, P-gp binding, skin permeation (logKp), and CYP450 inhibition. Additionally, the tool presents five drug-likeness models (Lipinsky, Ghose, Veber, Egan, and Muegge) and medicinal chemistry alerts. It is integrated with the SwissDrugDesign workspace. The QikProp (Schrödinger, LLC, NY, 2019) provides rapid predictions of ADME properties for molecules with novel scaffolds as for analogs of well-known drugs and display information about octanol/water and water/gas logPs, logS, logBBB, overall CNS activity, Caco-2 and MDCK cell permeabilities, log Kd for human serum albumin binding, and log IC_{50} for HERG K+-channel blockage.

Prediction of Synthetic Tractability (Synthesizability) of New Compounds

Even though large numbers of molecules are generated by *de novo* design, many of them are synthetically infeasible

(Dey and Caflisch, 2008). To address this problem, methods to calculate the synthetic accessibility (SA) are being developed. SA can be addressed by estimating the complexity of the molecule or making a retrosynthetic approach, where the complete synthetic tree leading to the molecules needs to be processed (Ertl and Schuffenhauer, 2009). SYLVIA (Boda et al., 2007) is one of the programs that estimate the synthetic accessibility of an organic compound. It obtains the SA score by the addition of five variables as the molecular graph complexity, ring complexity, stereochemical complexity, starting material similarity and reaction center substructure, where the first three are structure-based and the other two utilize information from starting material catalogs and reaction databases. Ertl and Schuffenhauer (2009) developed another method that uses historical synthetic knowledge obtained by analyzing information from millions of already synthesized chemicals and also considers molecule complexity. The method is based on a combination of fragment contributions and a complexity penalty. Podolyan et al. (2010) presented two approaches to quickly predict the synthetic accessibility of chemical compounds by utilizing SVMs operating on molecular descriptors. The first approach (RSsvm) identifies compounds that can be synthesized using a specific set of reactions and starting materials and builds the model by training the compounds identified as synthetically or otherwise accessible by retrosynthetic analysis while the second approach (DRSVM) is constructed to generate a more general assessment. More recently, Fukunishi et al. (2014) designed a new method of predicting SA based on commercially available compound databases and molecular descriptors where the SA is estimated from the probability of the existence of substructures of the compound, the number of symmetry atoms, the graph complexity, and the number of the chiral center of the compound.

Synthesizability-Aware Methods

Given the difficulty of synthesis of most of the leads produced by *de novo* approaches, some programs added methods to score the SA. Lead+Op (Lin et al., 2018) is an example of these programs that takes an initial fragment, looks for associated reaction rules, virtually generate the reaction products and select the best binding conformation. Them it generates conformers and select one that becomes a reactant for another round. Also, programs mentioned above as LigBuilder and Autogrow include SA analysis on their current versions. In the medicinal chemistry component of **SwissADME,** a SA score is also included.

Different programs use distinct algorithms for *de novo* design compounds in CADD. **Table 1** summarizes some programs cited in this section.

TABLE 1 | FBDD programs with respective approaches.

Program	Algorithm	FBDD Approach	SA
AUTOGROW	Docking + Genetic Algorithm	Growing	YES (Latest version)
LUDI	Empirical scoring	Linking	NO
AUTO T and T	Transplants fragments into the lead	Merging	NO
LeadOp+R	Looks for associated reaction rules	Growing	YES
GANDI	Genetic Algorithm	Linking	NO
LigBuilder 2	Genetic Algorithm	Linking and Growing	YES
ADAPT	Genetic Algorithm	Growing	NO

CASE STUDIES IN THE LAST FIVE YEARS

Case 1: FBDD in the Development of New Anti-mycobacterium Drugs

A successful application of the FBDD techniques have been applied to early stage drug discovery of new therapeutics against *Mycobacterium sp.* and in particular *M. tuberculosis* (*Mtb*) and *M. abscessus (Mab)* (Thomas et al., 2017). *Mtb*, the causative agent of tuberculosis, has several therapeutic interventions developed to treat the disease. However, through their long-term use and misuse, the efficacy of these drugs is becoming reduced with strains currently circulating that are mono-resistant, multidrug-resistant, extensively drug-resistant and totally drug-resistant. Despite this little drug development activity has been undertaken since the 1960's. However, relatively in response to the growing drug-resistant threat many different approaches are being deployed to developing novel therapeutics, including FBDD. An example of such an effort is against the meta cleavage product hydrolase (HsaD) that is involved in cholesterol catabolism in *Mtb*. Initial screening was conducted on a library of 1,258 fragments using differential scanning fluorimetry, with hits confirmed by ligand-observed NMR spectroscopy and inhibition by enzymatic assay. The three confirmed fragment initial hits were structurally characterized by X-ray crystallography and fragment soaking. A small series of compounds based on these hits were further tested for activity both *in vitro* and *ex vivo* with promising results (Ryan et al., 2017).

Another target of *Mtb* where FBDD has been applied is the pantothenate synthetase (Pts) where a similar sized fragment library of 1,250 rule-of-three compliant fragments was investigated. An initial screen was performed using a thermal shift assay, followed by a secondary screen using 1-D NMR spectroscopy with ultimate hit validation by isothermal titration calorimetry and characterization by X-ray crystallography. Three distinct fragment binding sites were identified (Silvestre et al., 2013). Follow-up expansion of one of the fragment sites using a combination of fragment linking and fragment growing generated a new series of inhibitors. Though fragment linking seemed to be an attractive approach, the limitation in the repertoire of linkers compromised the binding mode. Greater success came from fragment growth using expert knowledge and the protein target as a template (Hung et al., 2009).

Targets in other pathogenic *Mycobacterium* sp. have also been subject to successful FBDD campaigns. Most notably the recent development of inhibitors against tRNA methyltransferase (TrmD) of *M. abscessus* (Mab). This multi-drug resistant pathogen is increasingly problematic in individuals with cystic

fibrosis and other chronic lung conditions. A library of 960 fragments was screened biophysically using differential scanning fluorimetry in a similar fashion used for HsaD, with 53 hits taken to validation and structural characterization using X-ray crystallography (no NMR based validation was undertaken). Only 27 fragments could be validated all of which bound within the substrate binding pocket. A strategy of fragment-merging centered around the overlap of a 4-methoxyphenyl ring system with the indole ring system of two fragments that spanned the adenine and ribose binding pockets. This was explored successfully with a new combined compound providing a new aminopyrazole-indole scaffold with both improved affinity (K_d = 110 μM, LE of 0.36) and prospects for further elaboration relative to the parent fragments. It also exhibited inhibition activity *in vitro* and *ex vivo* with promising *in vivo* activity also against *M. leprae*, the causative agent of leprosy (Whitehouse et al., 2019).

These successful FBDD campaigns against a range of targets in pathogenic *Mycobacterium* have yielded promising leads with indications of efficacy in *ex vivo* and *in vivo* demonstrating both the power and efficacy of the approach. The ability of these leads to work across a range of pathogens is also highly encouraging. However, work still needs to be done to improve these leads to progress them into early clinical evaluations and into clinical use.

Case 2: Inhibitors of Dengue Virus Enzymes

A 2014 paper (Coutard et al., 2014) describes the use of FBLD in the discovery of inhibitors for an important subunit of dengue virus (DENV) viral replication complex. In this work, 500 fragments were screened against two subunits of the viral replication complex: NS3 helicase (Hel) and the NS5 mRNA methyltransferase (MTase) subunits. DENV Hel, located in the *C*-terminal region of the NS3 subunit of the replication complex, is involved in viral genome replication and RNA capping. The role of DENV NS5 MTase is related with a double methylation (N-7 and 2'-O) during the cap formation process in flavivirus (Dong et al., 2008).

The authors used a combination of Thermal Shift Assay (TSA), X-ray diffraction crystallography (XRD) and enzymatic assays in order to screen compounds against NS3 DENV Hel and NS5 DENV MTase subunits. The TSA was used as the primary screening technique. During the TSA screening, not surprisingly part of the fragments—used at high concentrations and with poorly optimized physicochemical properties—presented solubility problems. This was the reason for the exclusion of ∼4.8% of the screened compounds during this phase. This initial screening yielded 68 hits, from those, 7 were found bound to the DENV MTase subunit by XRD.

Using a direct colorimetric ATPase-based assay to identify inhibitors, from those previous 7 crystallographic hits, 5 fragments (**Table 2**) were classified as hits with their potency varying between 180 μM and 9 mM.

In the most recent work, the fragments **3** and **4** (**Figure 8**) were found bound at the DENV MTase S-Adenosyl-L-methionine (AdoMet) binding site using XRD. A computer-aided fragment optimization gave rise to a new series of compounds using these two fragments. The urea was used as a linker to connect the fragments. Further modifications yielded compounds **5** and **6** (**Figure 8**).

TABLE 2 | Inhibition and potency data from the final hits (Coutard et al., 2014).

	DENV 2'O-Mtase activity inhibition (%)	DENV 2'O-MTase activity IC_{50} (mM)
81	4	3.90 ± 0.16
91	11	2.83 ± 0.18
95	85	0.18 ± 0.01
157	9	9.39 ± 0.90
217	11	3.12 ± 0.27

During the optimization process, the authors had good insights about the important features to the molecule binding on this target. One of these features is the presence of phenyl rings substituted in meta position and is crucial for favoring binding.

This work yielded two inhibitors (**5** and **6**) with potency around 100 μM, even though no effect was observed on a cell assay. Despite this negative result, this work showed the feasibility of the FBDD approach in getting micromolar inhibitors from structurally simple fragments.

Case 3: MTH1 Inhibitors for Anticancer Drug Discovery

The mutT homolog 1 (MTH1) is an enzyme involved in the prevention of incorporation of deoxynucleoside triphosphates (dNTPs) oxidized by reactive oxygen species (ROS), e.g., 8-oxodGTP or 2-OH-dATP, into DNA, which prevents the killing of the cell. MTH1 is frequently overexpressed in cancer cells and is non-essential in normal cells, proving to be a druggable target for cancer treatment (Smits and Gillespie, 2014; Berglund et al., 2016).

Rudling et al. applied a combination of molecular docking, SAR by catalog, and experimental testing for discovering and optimizing MTH1 inhibitors (Rudling et al., 2017). Initially, a molecular docking-based virtual screening using a crystal structure of MTH1 was performed using 0.3 million fragments from the ZINC fragment-like database, all commercially available. Subsequently, for the 5,000 top-ranked fragments, allowed the search of analogs representing superstructures of the fragment or containing similar substructures in the ZINC database using the chemical structures encoded as circular fingerprints and the Tversky similarity index (Tversky, 1977). The criteria used to select analogs from 4.4 million commercially available compounds in the ZINC database was the following: (i) Tversky similarity >0.8; (ii) up to six additional heavy atoms (HAs) compared to the parent fragment; (iii) improved docking score 80% lower compared to the parent fragment; (iv) visual inspection of the binding modes. After these analyses, a set of 22 commercially available fragments with at least five analogs comprising the above-mentioned criteria were selected for experimental evaluation. Five of these 22 fragments showing IC_{50} values ranging from 5.6 to 79 μM were considered hits and were used for F2L (**Table 3**). The fragment **7** presented an

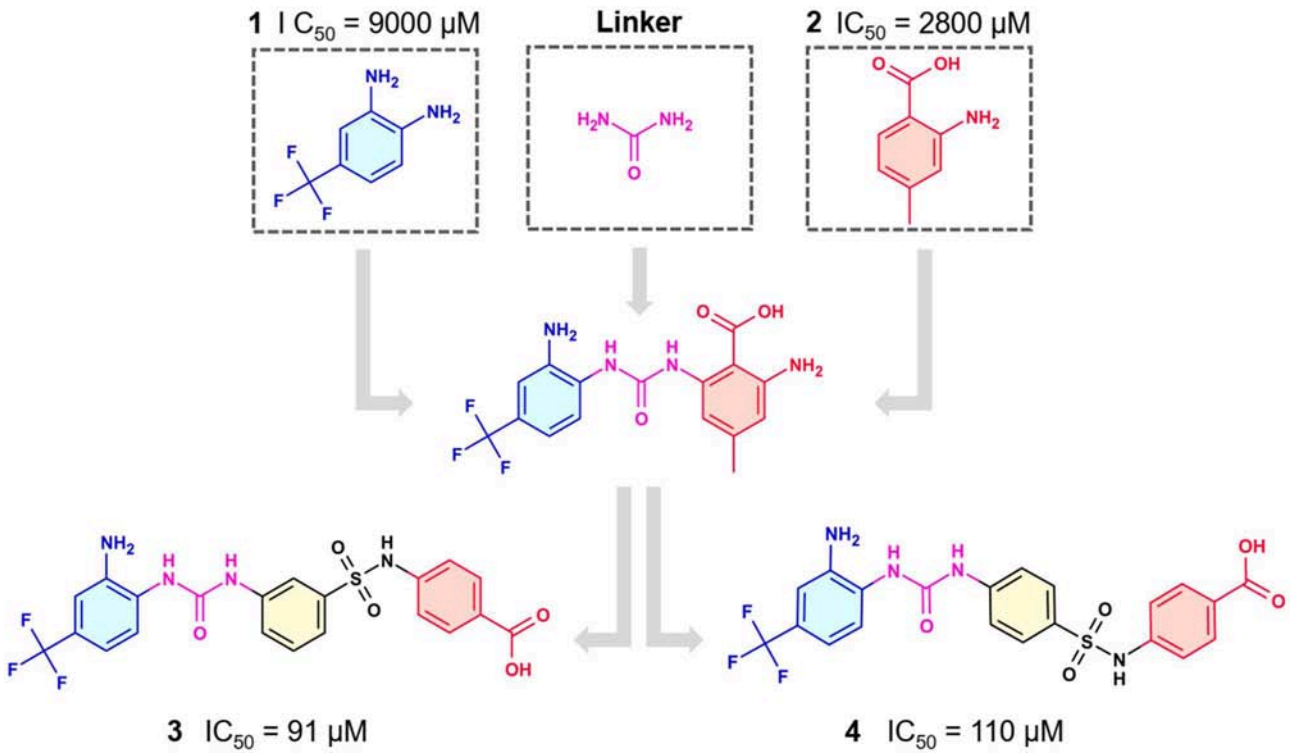


FIGURE 8 | Fragment optimization predecessors and products (Coutard et al., 2014; Benmansour et al., 2017).

IC_{50} value of 79 µM and its most potent analog presented an IC_{50} of 170 nM, representing a 470-fold improvement (**Table 3**). Although the crystal structure of the fragment **7** in complex with MTH1 was not obtained, the crystallization of its most potent analog and MTH1 was solved at 1.85 Å resolution, demonstrating an RMSD of 0.6 Å between the common atoms of the fragment **7** (binding mode predicted by molecular docking) and the analog (crystal structure). Because of the closely related structures and binding modes of fragments **8** and **9**, its analogs were analyzed together, but the most potent analog presented an IC_{50} of only 3.5 µM. The most potent analog of fragment **10** presented a 190-fold increase of the activity, with an IC_{50} of 120 nM (**Table 3**). The crystallization of fragment **10** with MTH1 was also unsuccessful, but the molecular docking was able to predict the binding mode, showing an RMSD of 0.9 comparing the overlapping atoms with the crystal of the most potent analog obtained at 1.50 Å resolution. The analogs of the fragment **11** were not available during the study. This work demonstrated that virtual screening and SAR by catalog can be used to rapidly identify and optimize fragments into nanomolar inhibitors (Rudling et al., 2017).

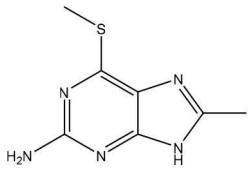
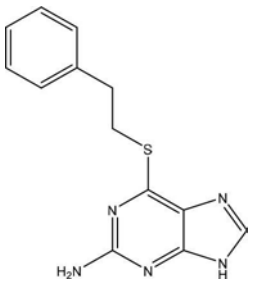
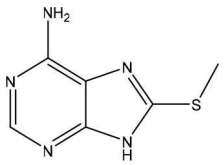
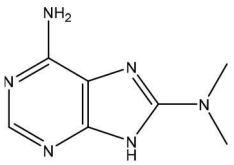
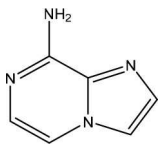
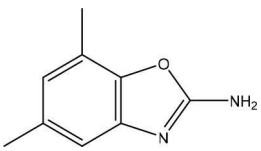
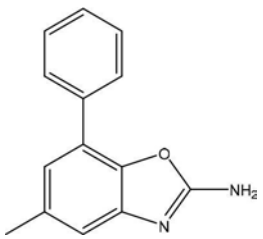
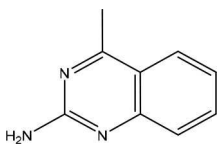
Case 4: New Acetylcholinesterase Inhibitors Against Alzheimer's Disease

Alzheimer's disease is a neurodegenerative disorder and has no cure. The actual treatments are based on drugs that leverage the transmission of electrical impulses. Pascoini et al. (2019), computationally developed new acetylcholinesterase (AChE) inhibitors. AChE is responsible for decreasing levels of acetylcholine in the synaptic cleft. Their inhibition enhances the transmission of the electric impulse (Polinsky, 1998; Talesa, 2001).

For the *in silico* inhibitor development, they divided the process into four steps. First, a *de novo* design was applied to generate an initial library of compounds. The first library was then filtered according to ADME properties at the second step. In the third step, the filtered library was filtered again using a similarity criterion. Finally, the resulting library was used in docking studies. The best three complexes were used for molecular dynamic studies. In this work, they used three reference drugs for AD treatment: donepezil, galantamine, and rivastigmine.

For the *de novo* design, the LigBuilder software was used. The CAVITY procedure was employed to detect and analyze ligand-binding sites of the target. It classified the cavities' druggability, who would be used for docking studies. The BUILD procedure was used in the exploring and growing/linking modes. In the explore mode, fragments from the program's database were added in the protein site and their interaction was scored. Then, the fragments with the best scores were linked. In the growing mode, seeds molecules were put at the binding site and fragments were added to the seeds. At the linking mode, the seed was divided into fragments and other fragments were added to them. After the BUILD procedure, they got a library of 2.5 million compounds. The resulting library was filtered according to ADME properties with the software QUIKPROP where molecules that infringed more than five properties (physicochemical properties, lipophilicity, water solubility, pharmacokinetics models) were discarded. A library

TABLE 3 | Experimental data for the five most potent MTH1 inhibitors (data taken from Rudling et al., 2017).

Fragment ID	Fragment 2D structure	Fragment IC_{50}	Most potent analog 2D structure	Analog IC_{50}
7		79 μM		0.17 μM
8		24 μM		3.5 μM
9		26 μM	n/a	n/a
10		23 μM		0.12 μM
11		5.6 μM	n/a	n/a

n/a, not available.

of 6,000 compounds results from this process. After this, the Tanimoto's coefficient was applied to measure the similarity among the molecules. Molecules below 0.85 were excluded and 1,500 molecules were considered for the next step. The final step consisted of docking studies, carried out with the GLIDE software and the Induced Fit Docking protocol. Afterward, they selected the three best complexes from the docking and used them as input structures for molecular dynamic studies. Finally, they obtained three compounds with high stability and good binding energies, some of them even better than the reference drugs.

CONCLUDING REMARKS

FBDD has matured to become a key strategy in modern pharmaceutical research. With less requirement for large chemical libraries and the possibility of using a range of biophysical methods for screening, the easier and scalable implementation of this strategy has facilitated its popularization, especially among academic institutions and smaller pharmaceutical companies.

The main reason for the success of the FBDD strategy is because it presents a more efficient and consistent route for optimization of initial screening hits into lead compounds. As reviewed here, many routes are available for expansion of fragment hits and *in silico* methods are key to support or guide the majority of them.

A variety of *in silico* methods have been used in F2L optimization in FBDD, from binding site analysis to *de novo* design of new fragment-derived ligands with synthesizability-aware methods. The case studies highlighted here clearly

demonstrate how the different *in silico* methods can be used in integrated form and combined with experimental approaches to successfully develop higher affinity ligands from fragments.

Advances in artificial intelligence methods, such as deep learning, hold a great potential to accelerate the optimization of fragment hits in lead compounds. Recent examples show that these hits can be already optimized automatically taking into consideration several parameters such as bioactivity, solubility, synthetic feasibility, and ADMET properties.

AUTHOR CONTRIBUTIONS

FS, CA, and NF drafted the review topics with the abstract and revised the manuscript. FS drafted the conclusion and contributed to several parts of the manuscript. NF drafted the introduction and one of the case studies. LS drafted the fragment optimization section and one of the case studies. JM-F drafted most of the *in silico* strategies section and one of the case studies. BN designed all figures and revised the manuscript. AG and RM drafted the *de novo* design section and one of the case studies.

ACKNOWLEDGMENTS

The authors would like to thank Brazilian funding agencies CNPq, CAPES, FAPERJ, FAPEG, and FIOCRUZ for financial support and fellowships. This study was also supported by funding from the United Kingdom's Academy of Medical Sciences Newton Fund (NAF\R2\180641).

REFERENCES

Abad-Zapatero, C. (2013). *Ligand Efficiency Indices for Drug Discovery*. Oxford, UK: Elsevier.

Abad-Zapatero, C. (2017). Are SAR tables obsolete? *Drug Discov. Today* 22, 195–198. doi: 10.1016/j.drudis.2016.12.002

Abi Hussein, H., Geneix, C., Petitjean, M., Borrel, A., Flatters, D., and Camproux, A.-C. (2017). Global vision of druggability issues: applications and perspectives. *Drug Discov. Today* 22, 404–415. doi: 10.1016/j.drudis.2016.11.021

Alvarsson, J., Eklund, M., Engkvist, O., Spjuth, O., Carlsson, L., Wikberg, J. E. S., et al. (2014). Ligand-based target prediction with signature fingerprints. *J. Chem. Inf. Model.* 54, 2647–2653. doi: 10.1021/ci500361u

Alves, V., Braga, R., Muratov, E., and Andrade, C. (2018). Development of web and mobile applications for chemical toxicity prediction. *J. Braz. Chem. Soc.* 29, 982–988. doi: 10.21577/0103-5053.20180013

Andrade, C. H., Neves, B. J., Melo-Filho, C. C., Rodrigues, J., Silva, D. C., Braga, R. C., et al. (2018). *In silico* chemogenomics drug repositioning strategies for neglected tropical diseases. *Curr. Med. Chem.* 25, 1–22. doi: 10.2174/0929867325666180309114824

Awale, M., Sirockin, F., Stiefl, N., and Reymond, J.-L. (2019). Drug analogs from fragment-based long short-term memory generative neural networks. *J. Chem. Inf. Model.* 59, 1347–1356. doi: 10.1021/acs.jcim.8b00902

Barril, X. (2013). Druggability predictions: methods, limitations, and applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 3, 327–338. doi: 10.1002/wcms.1134

Benmansour, F., Trist, I., Coutard, B., Decroly, E., Querat, G., Brancale, A., et al. (2017). Discovery of novel dengue virus NS5 methyltransferase non-nucleoside inhibitors by fragment-based drug design. *Eur. J. Med. Chem.* 125, 865–880. doi: 10.1016/j.ejmech.2016.10.007

Berglund, U. W., Sanjiv, K., Gad, H., Kalderén, C., Koolmeister, T., Pham, T., et al. (2016). Validation and development of MTH1 inhibitors for treatment of cancer. *Ann. Oncol.* 27, 2275–2283. doi: 10.1093/annonc/mdw429

Boda, K., Seidel, T., and Gasteiger, J. (2007). Structure and reaction based evaluation of synthetic accessibility. *J. Comput. Aided. Mol. Des.* 21, 311–325. doi: 10.1007/s10822-006-9099-2

Boehr, D. D., Nussinov, R., and Wright, P. E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* 5, 789–796. doi: 10.1038/nchembio.232

Boettcher, A., Ruedisser, S., Erbel, P., Vinzenz, D., Schiering, N., Hassiepen, U., et al. (2010). Fragment-based screening by biochemical assays. *J. Biomol. Screen.* 15, 1029–1041. doi: 10.1177/1087057110380455

Bohm, H.-J. (1992). The computer program LUDI: a new method for the *de novo* design of enzyme inhibitors. *J. Comput. Aided. Mol. Des.* 6, 61–78. doi: 10.1007/BF00124387

Bollag, G., Hirth, P., Tsai, J., Zhang, J., Ibrahim, P. N., Cho, H., et al. (2010). Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma. *Nature* 467, 596–599. doi: 10.1038/nature09454

Bollag, G., Tsai, J., Zhang, J., Zhang, C., Ibrahim, P., Nolop, K., et al. (2012). Vemurafenib: the first drug approved for BRAF-mutant cancer. *Nat. Rev. Drug Discov.* 11, 873–886. doi: 10.1038/nrd3847

Braga, R., Alves, V., Silva, A., Nascimento, M., Silva, F., Liao, L., et al. (2014). Virtual screening strategies in medicinal chemistry: the state of the art and current challenges. *Curr. Top. Med. Chem.* 14, 1899–1912. doi: 10.2174/1568026614666140929120749

Braga, R. C., Alves, V. M., Silva, M. F. B., Muratov, E., Fourches, D., Lião, L. M., et al. (2015). Pred-hERG: a novel web-accessible computational tool for predicting cardiac toxicity. *Mol. Inform.* 34, 698–701. doi: 10.1002/minf.201500040

Chen, Y., and Shoichet, B. K. (2009). Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat. Chem. Biol.* 5, 358–364. doi: 10.1038/nchembio.155

Cheng, F., Li, W., Zhou, Y., Shen, J., Wu, Z., Liu, G., et al. (2012). admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. *J. Chem. Inf. Model.* 52, 3099–3105. doi: 10.1021/ci300367a

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., et al. (2014). QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* 57, 4977–5010. doi: 10.1021/jm4004285

Chevillard, F., Rimmer, H., Betti, C., Pardon, E., Ballet, S., van Hilten, N., et al. (2018). Binding-site compatible fragment growing applied to the design of β2 -adrenergic receptor ligands. *J. Med. Chem.* 61, 1118–1129. doi: 10.1021/acs.jmedchem.7b01558

Chung, S., Parker, J. B., Bianchet, M., Amzel, L. M., and Stivers, J. T. (2009). Impact of linker strain and flexibility in the design of a fragment-based inhibitor. *Nat. Chem. Biol.* 5, 407–413. doi: 10.1038/nchembio.163

Ciulli, A. (2013a). Biophysical screening for the discovery of small-molecule ligands. *Methods Mol. Biol.* 1008, 357–388. doi: 10.1007/978-1-62703-398-5_13

Ciulli, A. (2013b). "Biophysical screening for the discovery of small-molecule ligands," in *Methods in Molecular Biology*, eds M. A. Williams and T. Daviter (Totowa, NJ: Humana Press), 357–388.

Coutard, B., Decroly, E., Li, C., Sharff, A., Lescar, J., Bricogne, G., et al. (2014). Assessment of Dengue virus helicase and methyltransferase as targets for fragment-based drug discovery. *Antiviral Res.* 106, 61–70. doi: 10.1016/j.antiviral.2014.03.013

Cox, O. B., Krojer, T., Collins, P., Monteiro, O., Talon, R., Bradley, A., et al. (2016). A poised fragment library enables rapid synthetic expansion yielding the first reported inhibitors of PHIP(2), an atypical bromodomain. *Chem. Sci.* 7, 2322–2330. doi: 10.1039/C5SC03115J

Cukuroglu, E., Engin, H. B., Gursoy, A., and Keskin, O. (2014). Hot spots in protein-protein interfaces: towards drug discovery. *Prog. Biophys. Mol. Biol.* 116, 165–173. doi: 10.1016/j.pbiomolbio.2014.06.003

Daina, A., Michielin, O., and Zoete, V. (2017). SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* 7:42717. doi: 10.1038/srep42717

Davis, B. J., and Roughley, S. D. (2017). "Fragment-based lead discovery," in *Annual Reports in Medicinal Chemistry*, ed R. A. Goodnow (London: Elsevier Inc.), 371–439. doi: 10.1016/bs.armc.2017.07.002

De Fusco, C., Brear, P., Iegre, J., Georgiou, K. H., Sore, H. F., Hyvönen, M., et al. (2017). A fragment-based approach leading to the discovery of a novel binding site and the selective CK2 inhibitor CAM4066. *Bioorg. Med. Chem.* 25, 3471–3482. doi: 10.1016/j.bmc.2017.04.037

Dey, F., and Caflisch, A. (2008). Fragment-based *de novo* ligand design by multiobjective evolutionary optimization. *J. Chem. Inf. Model.* 48, 679–690. doi: 10.1021/ci700424b

Donald, A., McHardy, T., Rowlands, M. G., Hunter, L.-J. K., Davies, T. G., Berdini, V., et al. (2007). Rapid evolution of 6-phenylpurine inhibitors of protein kinase B through structure-based design. *J. Med. Chem.* 50, 2289–2292. doi: 10.1021/jm0700924

Dong, H., Zhang, B., and Shi, P.-Y. (2008). Flavivirus methyltransferase: a novel antiviral target. *Antiviral Res.* 80, 1–10. doi: 10.1016/j.antiviral.2008.05.003

Dong, J., Wang, N.-N., Yao, Z.-J., Zhang, L., Cheng, Y., Ouyang, D., et al. (2018). ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. *J. Cheminform.* 10:29. doi: 10.1186/s13321-018-0283-x

Durrant, J. D., Amaro, R. E., and McCammon, J. A. (2009). AutoGrow: a novel algorithm for protein inhibitor design. *Chem. Biol. Drug Des.* 73, 168–178. doi: 10.1111/j.1747-0285.2008.00761.x

Durrant, J. D., Lindert, S., and McCammon, J. A. (2013). AutoGrow 3.0: an improved algorithm for chemically tractable, semi-automated protein inhibitor design. *J. Mol. Graph. Model.* 44, 104–112. doi: 10.1016/j.jmgm.2013.05.006

Erlanson, D. A., Davis, B. J., and Jahnke, W. (2019). Fragment-based drug discovery: advancing fragments in the absence of crystal structures. *Cell Chem. Biol.* 26, 9–15. doi: 10.1016/j.chembiol.2018.10.001

Erlanson, D. A., Fesik, S. W., Hubbard, R. E., Jahnke, W., and Jhoti, H. (2016). Twenty years on: the impact of fragments on drug discovery. *Nat. Rev. Drug Discov.* 15, 605–619. doi: 10.1038/nrd.2016.109

Ertl, P., and Schuffenhauer, A. (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* 1:8. doi: 10.1186/1758-2946-1-8

Ferreira, L. L. G., and Andricopulo, A. D. (2019). ADMET modeling approaches in drug discovery. *Drug Discov. Today* 24, 1157–1165. doi: 10.1016/j.drudis.2019.03.015

Fukunishi, Y., Kurosawa, T., Mikami, Y., and Nakamura, H. (2014). Prediction of synthetic accessibility based on commercially available compound databases. *J. Chem. Inf. Model.* 54, 3259–3267. doi: 10.1021/ci500568d

Gawehn, E., Hiss, J. A., Brown, J. B., and Schneider, G. (2018). Advancing drug discovery via GPU-based deep learning. *Expert Opin. Drug Discov.* 13, 579–582. doi: 10.1080/17460441.2018.1465407

Giordanetto, F., Jin, C., Willmore, L., Feher, M., and Shaw, D. E. (2019). Fragment hits: what do they look like and how do they bind? *J. Med. Chem.* 62, 3381–3394. doi: 10.1021/acs.jmedchem.8b01855

Godemann, R., Madden, J., Krämer, J., Smith, M., Fritz, U., Hesterkamp, T., et al. (2009). Fragment-based discovery of BACE1 inhibitors using functional assays. *Biochemistry* 48, 10743–10751. doi: 10.1021/bi901061a

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., et al. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* 4, 268–276. doi: 10.1021/acscentsci.7b00572

Grove, L. E., Vajda, S., and Kozakov, D. (2016). "Computational methods to support fragment-based drug discovery," in *Fragment-Based Drug Discovery: Lessons and Outlook*, eds D. A. Erlanson and W. Jahnke (Weinheim: Wiley-VCH Verlag GmbH and Co. KGaA), 197–222.

Gupta, A., Müller, A. T., Huisman, B. J. H., Fuchs, J. A., Schneider, P., and Schneider, G. (2018). Generative recurrent networks for *de novo* drug design. *Mol. Inform.* 37:1700111. doi: 10.1002/minf.201700111

Hall, D. R., Ngan, C. H., Zerbe, B. S., Kozakov, D., and Vajda, S. (2012). Hot spot analysis for driving the development of hits into leads in fragment-based drug discovery. *J. Chem. Inf. Model.* 52, 199–209. doi: 10.1021/ci200468p

Hall, R. J., Mortenson, P. N., and Murray, C. W. (2014). Efficient exploration of chemical space by fragment-based screening. *Prog. Biophys. Mol. Biol.* 116, 82–91. doi: 10.1016/j.pbiomolbio.2014.09.007

Hall, R. J., Murray, C. W., and Verdonk, M. L. (2017). The fragment network: a chemistry recommendation engine built using a graph database. *J. Med. Chem.* 60, 6440–6450. doi: 10.1021/acs.jmedchem.7b00809

Hann, M. M., Leach, A. R., and Harper, G. (2001). Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* 41, 856–864. doi: 10.1021/ci000403i

Harel, S., and Radinsky, K. (2018). Prototype-based compound discovery using deep generative models. *Mol. Pharm.* 15, 4406–4416. doi: 10.1021/acs.molpharmaceut.8b00474

Henzler-Wildman, K., and Kern, D. (2007). Dynamic personalities of proteins. *Nature* 450, 964–972. doi: 10.1038/nature06522

Hoffer, L., Voitovich, Y. V., Raux, B., Carrasco, K., Muller, C., Fedorov, A. Y., et al. (2018). Integrated strategy for lead optimization based on fragment growing: the diversity-oriented-target-focused-synthesis approach. *J. Med. Chem.* 61, 5719–5732. doi: 10.1021/acs.jmedchem.8b00653

Hopkins, A. L., Groom, C. R., and Alex, A. (2004). Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* 9, 430–431. doi: 10.1016/S1359-6446(04)03069-7

Hubbard, R. (2015). Current perspectives in fragment-based ligand discovery. *Acta Crystallogr. Sect. A Found. Adv.* 71:s37. doi: 10.1107/S2053273315099416

Hubbard, R. E., and Murray, J. B. (2011). *Experiences in Fragment-Based Lead Discovery*, 1st Edn. Oxford, UK: Elsevier Inc.

Hung, A. W., Silvestre, H. L., Wen, S., Ciulli, A., Blundell, T. L., and Abell, C. (2009). Application of fragment growing and fragment linking to the discovery of inhibitors of *Mycobacterium tuberculosis* pantothenate synthetase. *Angew. Chemie Int. Ed.* 48, 8452–8456. doi: 10.1002/anie.200903821

Hussein, H. A., Borrel, A., Geneix, C., Petitjean, M., Regad, L., and Camproux, A.-C. (2015). PockDrug-server: a new web server for predicting pocket druggability on holo and apo proteins. *Nucleic Acids Res.* 43, W436–W442. doi: 10.1093/nar/gkv462

Ichihara, O., Barker, J., Law, R. J., and Whittaker, M. (2011). Compound design by fragment-linking. *Mol. Inform.* 30, 298–306. doi: 10.1002/minf.201000174

Jacquemard, C., Drwal, M. N., Desaphy, J., and Kellenberger, E. (2019). Binding mode information improves fragment docking. *J Cheminform* 11:24. doi: 10.1186/s13321-019-0346-7

Joseph-mccarthy, D., Campbell, A. J., Kern, G., and Moustakas, D. (2013). Fragment-based lead discovery and design. *J. Chem. Inf. Model.* 54, 693–704. doi: 10.1021/ci400731w

Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., et al. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat Protoc* 10, 733–755. doi: 10.1038/nprot.2015.043

Kumar, A., Voet, A., and Zhang, K. Y. J. (2012). Fragment based drug design: from experimental to computational approaches. *Curr. Med. Chem.* 19, 5128–5147. doi: 10.2174/092986712803530467

Lamoree, B., and Hubbard, R. E. (2017). Current perspectives in fragment-based lead discovery (FBLD). *Essays Biochem.* 61, 453–464. doi: 10.1042/EBC20170028

Li, Y., Zhang, L., and Liu, Z. (2018). Multi-objective *de novo* drug design with conditional graph generative model. *J. Cheminform.* 10, 1–24. doi: 10.1186/s13321-018-0287-6

Lin, F.-Y., Esposito, E. X., and Tseng, Y. J. (2018). LeadOp+R: structure-based lead optimization with synthetic accessibility. *Front. Pharmacol.* 9:96. doi: 10.3389/fphar.2018.00096

Lo, M.-C., Aulabaugh, A., Jin, G., Cowling, R., Bard, J., Malamas, M., et al. (2004). Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. *Anal. Biochem.* 332, 153–159. doi: 10.1016/j.ab.2004.04.031

Macarron, R., Banks, M. N., Bojanic, D., Burns, D. J., Cirovic, D. A., Garyantes, T., et al. (2011). Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* 10, 188–195. doi: 10.1038/nrd3368

Mak, K. K., and Pichika, M. R. (2019). Artificial intelligence in drug development: present status and future prospects. *Drug Discov. Today* 24, 773–780. doi: 10.1016/j.drudis.2018.11.014

Männel, B., Jaiteh, M., Zeifman, A., Randakova, A., Möller, D., Hübner, H., et al. (2017). Structure-guided screening for functionally selective D 2 dopamine receptor ligands from a virtual chemical library. *ACS Chem. Biol.* 12, 2652–2661. doi: 10.1021/acschembio.7b00493

Matias, P. M., Donner, P., Coelho, R., Thomaz, M., Peixoto, C., Macedo, S., et al. (2000). Structural evidence for ligand specificity in the binding domain of the human androgen receptor: implications for pathogenic gene mutations. *J. Biol. Chem.* 275, 26164–26171. doi: 10.1074/jbc.M004571200

Merk, D., Friedrich, L., Grisoni, F., and Schneider, G. (2018). *De novo* design of bioactive small molecules by artificial intelligence. *Mol. Inform.* 37, 3–6. doi: 10.1002/minf.201700153

Mitcheson, J. S., Chen, J., Lin, M., Culberson, C., and Sanguinetti, M. C. (2000). A structural basis for drug-induced long QT syndrome. *Proc. Natl. Acad. Sci. U.S.A.* 97, 12329–12333. doi: 10.1073/pnas.210244497

Miyake, Y., Itoh, Y., Hatanaka, A., Suzuma, Y., Suzuki, M., Kodama, H., et al. (2019). Identification of novel lysine demethylase 5-selective inhibitors by inhibitor-based fragment merging strategy. *Bioorg. Med. Chem.* 27, 1119–1129. doi: 10.1016/j.bmc.2019.02.006

Mondal, M., Groothuis, D. E., and Hirsch, A. K. H. (2015). Fragment growing exploiting dynamic combinatorial chemistry of inhibitors of the aspartic protease endothiapepsin. *Medchemcomm* 6, 1267–1271. doi: 10.1039/C5MD00157A

Navratilova, I., and Hopkins, A. L. (2010). Fragment screening by surface plasmon resonance. *ACS Med. Chem. Lett.* 1, 44–48. doi: 10.1021/ml900002k

Neves, B. J., Braga, R. C., Melo-Filho, C. C., Moreira-Filho, J. T., Muratov, E. N., and Andrade, C. H. (2018). QSAR-based virtual screening: advances and applications in drug discovery. *Front. Pharmacol.* 9:1275. doi: 10.3389/fphar.2018.01275

Ngan, C. H., Bohnuud, T., Mottarella, S. E., Beglov, D., Villar, E. A., Hall, D. R., et al. (2012). FTMAP: extended protein mapping with user-selected probe molecules. *Nucleic Acids Res.* 40, W271–W275. doi: 10.1093/nar/gks441

Nikiforov, P. O., Surade, S., Blaszczyk, M., Delorme, V., Brodin, P., Baulard, A. R., et al. (2016). A fragment merging approach towards the development of small molecule inhibitors of Mycobacterium tuberculosis EthR for use as ethionamide boosters. *Org. Biomol. Chem.* 14, 2318–2326. doi: 10.1039/C5OB02630J

Nissink, J. W. M. (2009). Simple size-independent measure of ligand efficiency. *J. Chem. Inf. Model.* 49, 1617–1622. doi: 10.1021/ci900094m

Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. (2017). Molecular *de-novo* design through deep reinforcement learning. *J. Cheminform.* 9, 1–14. doi: 10.1186/s13321-017-0235-x

Pascoini, A. L., Federico, L. B., Arêas, A. L. F., Verde, B. A., Freitas, P. G., and Camps, I. (2019). *In silico* development of new acetylcholinesterase inhibitors. *J. Biomol. Struct. Dyn.* 37, 1007–1021. doi: 10.1080/07391102.2018.1447513

Patel, D., Bauman, J. D., and Arnold, E. (2014). Advantages of crystallographic fragment screening: functional and mechanistic insights from a powerful platform for efficient drug discovery. *Prog. Biophys. Mol. Biol.* 116, 92–100. doi: 10.1016/j.pbiomolbio.2014.08.004

Pedro, L., and Quinn, R. (2016). Native mass spectrometry in fragment-based drug discovery. *Molecules* 21:E984. doi: 10.3390/molecules21080984

Pegg, S. C. H., Haresco, J. J., and Kuntz, I. D. (2001). A genetic algorithm for structure-based *de novo* design. *J. Comput. Aided. Mol. Des.* 15, 911–933. doi: 10.1023/A:1014389729000

Pérez-Sianes, J., Pérez-Sánchez, H., and Díaz, F. (2018). Virtual screening meets deep learning. *Curr. Comput. Aided. Drug Des.* 15, 6–28. doi: 10.2174/1573409914666181018141602

Pierce, A. C., Rao, G., and Bemis, G. W. (2004). BREED: generating novel inhibitors through hybridization of known ligands. Application to CDK2, P38, and HIV protease. *J. Med. Chem.* 47, 2768–2775. doi: 10.1021/jm030543u

Podolyan, Y., Walters, M. A., and Karypis, G. (2010). Assessing synthetic accessibility of chemical compounds using machine learning methods. *J. Chem. Inf. Model.* 50, 979–991. doi: 10.1021/ci900301v

Polinsky, R. J. (1998). Clinical pharmacology of rivastigmine: a new-generation acetylcholinesterase inhibitor for the treatment of alzheimer's disease. *Clin. Ther.* 20, 634–647. doi: 10.1016/S0149-2918(98)80127-6

Polykovskiy, D., Zhebrak, A., Vetrov, D., Ivanenkov, Y., Aladinskiy, V., Mamoshina, P., et al. (2018). Entangled conditional adversarial autoencoder for *de novo* drug discovery. *Mol. Pharm.* 15, 4398–4405. doi: 10.1021/acs.molpharmaceut.8b00839

Popova, M., Isayev, O., and Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. *Sci. Adv.* 4:eaap7885. doi: 10.1126/sciadv.aap7885

Putin, E., Asadulaev, A., Ivanenkov, Y., Aladinskiy, V., Sanchez-Lengeling, B., Aspuru-Guzik, A., et al. (2018). Reinforced adversarial neural computer for *de novo* molecular design. *J. Chem. Inf. Model.* 58, 1194–1204. doi: 10.1021/acs.jcim.7b00690

Ramsundar, B., Liu, B., Wu, Z., Verras, A., Tudor, M., Sheridan, R. P., et al. (2017). Is multitask deep learning practical for pharma? *J. Chem. Inf. Model.* 57, 2068–2076. doi: 10.1021/acs.jcim.7b00146

Riniker, S., and Landrum, G. A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Riniker Landrum J. Cheminformatics* 5:26. doi: 10.1186/1758-2946-5-26

Rodríguez, D., Chakraborty, S., Warnick, E., Crane, S., Gao, Z.-G., O'Connor, R., et al. (2016). Structure-based screening of uncharted chemical space for atypical adenosine receptor agonists. *ACS Chem. Biol.* 11, 2763–2772. doi: 10.1021/acschembio.6b00357

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi: 10.1021/ci100050t

Rudling, A., Gustafsson, R., Almlöf, I., Homan, E., Scobie, M., Warpman Berglund, U., et al. (2017). Fragment-based discovery and optimization of enzyme inhibitors by docking of commercial chemical space. *J. Med. Chem.* 60, 8160–8169. doi: 10.1021/acs.jmedchem.7b01006

Ryan, A., Polycarpou, E., Lack, N. A., Evangelopoulos, D., Sieg, C., Halman, A., et al. (2017). Investigation of the mycobacterial enzyme HsaD as a potential novel target for anti-tubercular agents using a fragment-based drug design approach. *Br. J. Pharmacol.* 174, 2209–2224. doi: 10.1111/bph.13810

Schmidtke, P., and Barril, X. (2010). Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* 53, 5858–5867. doi: 10.1021/jm100574m

Schneider, G. (2018). Generative models for artificially-intelligent molecular design. *Mol. Inform.* 37:1880131. doi: 10.1002/minf.201880131

Schneider, G., and Clark, D. E. (2019). Automated *de novo* drug design: are we nearly there yet? *Angew. Chemie Int. Ed.* 58, 10792–10803. doi: 10.1002/anie.201814681

Schneider, G., and Fechner, U. (2005). Computer-based *de novo* design of drug-like molecules. *Nat. Rev. Drug Discov.* 4, 649–663. doi: 10.1038/nrd1799

Schuffenhauer, A., Ruedisser, S., Marzinzik, A., Jahnke, W., Selzer, P., and Jacoby, E. (2005). Library design for fragment based screening. *Curr. Top. Med. Chem.* 5, 751–762. doi: 10.2174/1568026054637700

Schulz, M. N., Landström, J., Bright, K., and Hubbard, R. E. (2011). Design of a fragment library that maximally represents available chemical space. *J. Comput. Aided. Mol. Des.* 25, 611–620. doi: 10.1007/s10822-011-9461-x

Schyman, P., Liu, R., Desai, V., and Wallqvist, A. (2017). vNN web server for ADMET predictions. *Front. Pharmacol.* 8:889. doi: 10.3389/fphar.2017.00889

Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4, 120–131. doi: 10.1021/acscentsci.7b00512

Seo, M.-H., Park, J., Kim, E., Hohng, S., and Kim, H.-S. (2014). Protein conformational dynamics dictate the binding affinity for a ligand. *Nat. Commun.* 5:3724. doi: 10.1038/ncomms4724

Shuker, S. B., Hajduk, P. J., Meadows, R. P., and Fesik, S. W. (1996). Discovering high-affinity ligands for proteins: SAR by NMR. *Science* 274, 1531–1534. doi: 10.1126/science.274.5292.1531

Shultz, M. D. (2013). Setting expectations in molecular optimizations: strengths and limitations of commonly used composite parameters. *Bioorg. Med. Chem. Lett.* 23, 5980–5991. doi: 10.1016/j.bmcl.2013.08.029

Silvestre, H. L., Blundell, T. L., Abell, C., and Ciulli, A. (2013). Integrated biophysical approach to fragment screening and validation for fragment-based lead discovery. *Proc. Natl. Acad. Sci. U.S.A.* 110, 12984–12989. doi: 10.1073/pnas.1304045110

Smits, V. A. J., and Gillespie, D. A. (2014). Cancer therapy: targeting the poison within. *Cell Cycle* 13, 2330–2333. doi: 10.4161/cc.29756

Srinivas Reddy, A., Chen, L., and Zhang, S. (2013). "Structure-based *de novo* drug design," in *De novo Molecular Design*, ed. G. Schneider (Weinheim: Wiley-VCH Verlag GmbH and Co. KGaA), 97–124.

Sterling, T., and Irwin, J. J. (2015). ZINC 15 – ligand discovery for everyone. *J. Chem. Inf. Model.* 55, 2324–2337. doi: 10.1021/acs.jcim.5b00559

Strecker, C., Peters, H., Hackl, T., Peters, T., and Meyer, B. (2019). Fragment growing to design optimized inhibitors for human blood

group B Galactosyltransferase (GTB). *ChemMedChem* 1, 1336–1342. doi: 10.1002/cmdc.201900296

Talesa, V. N. (2001). Acetylcholinesterase in Alzheimer's disease. *Mech. Ageing Dev.* 122, 1961–1969. doi: 10.1016/S0047-6374(01)00309-8

Thomas, S. E., Mendes, V., Kim, S. Y., Malhotra, S., Ochoa-Montaño, B., Blaszczyk, M., et al. (2017). Structural biology and the design of new therapeutics: from HIV and cancer to mycobacterial infections. *J. Mol. Biol.* 429, 2677–2693. doi: 10.1016/j.jmb.2017.06.014

Torres, P. H. M., Sodero, A. C. R., Jofily, P., and Silva, F. P. Jr. (2019). Key topics in molecular docking for drug design. *Int. J. Mol. Sci.* 20:4574. doi: 10.3390/ijms20184574

Trevizani, R., Custódio, F. L., dos Santos, K. B., and Dardenne, L. E. (2017). Critical features of fragment libraries for protein structure prediction. *PLoS ONE* 12:e0170131. doi: 10.1371/journal.pone.0170131

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Mol. Inform.* 29, 476–488. doi: 10.1002/minf.201000061

Tversky, A. (1977). Features of similarity. *Psychol. Rev.* 84, 327–352. doi: 10.1037/0033-295X.84.4.327

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 18, 463–477. doi: 10.1038/s41573-019-0024-5

Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., and Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* 45, 2615–2623. doi: 10.1021/jm020017n

Verdonk, M. L., Giangreco, I., Hall, R. J., Korb, O., Mortenson, P. N., and Murray, C. W. (2011). Docking performance of fragments and druglike compounds. *J. Med. Chem.* 54, 5422–5431. doi: 10.1021/jm200558u

Verdonk, M. L., and Hartshorn, M. J. (2004). Structure-guided fragment screening for lead discovery. *Curr. Opin. Drug Discov. Devel.* 7, 404–410.

Wang, R., Gao, Y., and Lai, L. (2000). LigBuilder: a multi-purpose program for structure-based drug design. *J. Mol. Model.* 6, 498–516. doi: 10.1007/s0089400060498

Wang, T., Wu, M.-B., Chen, Z.-J., Chen, H., Lin, J.-P., and Yang, L.-R. (2015). Fragment-based drug discovery and molecular docking in drug design. *Curr. Pharm. Biotechnol.* 16, 11–25. doi: 10.2174/1389201015666141122204532

Whitehouse, A. J., Thomas, S. E., Brown, K. P., Fanourakis, A., Chan, D. S.-H., Libardo, M. D. J., et al. (2019). Development of inhibitors against *Mycobacterium abscessus* tRNA (m 1 G37) methyltransferase (TrmD) using fragment-based approaches. *J. Med. Chem.* 62, 7210–7232. doi: 10.1021/acs.jmedchem.9b00809

Willighagen, E. L., Waagmeester, A., Spjuth, O., Ansell, P., Williams, A. J., Tkachenko, V., et al. (2013). The ChEMBL database as linked open data. *J. Cheminform.* 5:23. doi: 10.1186/1758-2946-5-23

Xu, X., Fang, X., Wang, J., and Zhu, H. (2017). Identification of novel ROS inducer by merging the fragments of piperlongumine and dicoumarol. *Bioorg. Med. Chem. Lett.* 27, 1325–1328. doi: 10.1016/j.bmcl.2016.08.016

Yang, H., Lou, C., Sun, L., Li, J., Cai, Y., Wang, Z., et al. (2019a). admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties. *Bioinformatics* 35, 1067–1069. doi: 10.1093/bioinformatics/bty707

Yang, X., Wang, Y., Byrne, R., Schneider, G., and Yang, S. (2019b). Concepts of artificial intelligence for computer-assisted drug discovery. *Chem. Rev.* 119, 10520–10594. doi: 10.1021/acs.chemrev.8b00728

Yuan, Y., Pei, J., and Lai, L. (2011). LigBuilder 2: a practical *de novo* drug design approach. *J. Chem. Inf. Model.* 51, 1083–1091. doi: 10.1021/ci100350u

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 de Souza Neto, Moreira-Filho, Neves, Maidana, Guimarães, Furnham, Andrade and Silva. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 21 February 2020
doi: 10.3389/fchem.2020.00108

In silico Investigations of the Mode of Action of Novel Colchicine Derivatives Targeting β-Tubulin Isotypes: A Search for a Selective and Specific β-III Tubulin Ligand

Lorenzo Pallante[1], *Antonio Rocca*[1], *Greta Klejborowska*[2], *Adam Huczynski*[2], *Gianvito Grasso*[3], *Jack A. Tuszynski*[1,4] *and Marco A. Deriu*[1]*

[1] PolitoBIOMed Lab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Turin, Italy, [2] Department of Chemistry, Adam Mickiewicz University, Poznań, Poland, [3] Dalle Molle Institute for Artificial Intelligence (IDSIA), University of Applied Sciences of Southern Switzerland (SUPSI), University of Italian Switzerland (USI), Manno, Switzerland, [4] Department of Oncology, University of Alberta, Edmonton, AB, Canada

OPEN ACCESS

Edited by:
Kamil Kuca,
University of Hradec Králové, Czechia

Reviewed by:
John Holmes Miller,
Victoria University of Wellington, New Zealand
Ambarish Kunwar,
Indian Institute of Technology Bombay, India

***Correspondence:**
Marco A. Deriu
marco.deriu@polito.it

Specialty section:
This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

Received: 29 October 2019
Accepted: 04 February 2020
Published: 21 February 2020

Citation:
Pallante L, Rocca A, Klejborowska G, Huczynski A, Grasso G, Tuszynski JA and Deriu MA (2020) In silico Investigations of the Mode of Action of Novel Colchicine Derivatives Targeting β-Tubulin Isotypes: A Search for a Selective and Specific β-III Tubulin Ligand. Front. Chem. 8:108. doi: 10.3389/fchem.2020.00108

The cardinal role of microtubules in cell mitosis makes them interesting drug targets for many pharmacological treatments, including those against cancer. Moreover, different expression patterns between cell types for several tubulin isotypes represent a great opportunity to improve the selectivity and specificity of the employed drugs and to design novel compounds with higher activity only on cells of interest. In this context, tubulin isotype βIII represents an excellent target for anti-tumoral therapies since it is overexpressed in most cancer cells and correlated with drug resistance. Colchicine is a well-known antimitotic agent, which is able to bind the tubulin dimer and to halt the mitotic process. However, it shows high toxicity also on normal cells and it is not specific for isotype βIII. In this context, the search for colchicine derivatives is a matter of great importance in cancer research. In this study, homology modeling techniques, molecular docking, and molecular dynamics simulations have been employed to characterize the interaction between 55 new promising colchicine derivatives and tubulin isotype βIII. These compounds were screened and ranked based on their binding affinity and conformational stability in the colchicine binding site of tubulin βIII. Results from this study point the attention on an amide of 4-chlorine thiocolchicine. This colchicine-derivative is characterized by a unique mode of interaction with tubulin, compared to all other compounds considered, which is primarily characterized by the involvement of the α-T5 loop, a key player in the colchicine binding site. Information provided by the present study may be particularly important in the rational design of colchicine-derivatives targeting drug resistant cancer phenotypes.

Keywords: molecular modeling, drug discovery, microtubule, cancer, drug resistance, tubulin, colchicine, colchicine derivatives

INTRODUCTION

The pivotal role of microtubules (MTs) in the mitotic process make them important targets for anticancer therapies since cancerous cells proliferate by unregulated cell division (Gajewski et al., 2013). By either stabilizing MTs or enhancing their depolymerization, it is possible to halt the mitotic process and eventually lead cells to apoptosis (Nettles et al., 2002). Among antimitotic agents, colchicine is able to block cell division (Bhattacharyya et al., 2008) by destabilizing MT assembly kinetics and dynamics. In particular, when colchicine binds in its specific binding site (located at the interface between tubulin α and β monomers) the structural conformation of the tubulin dimer is affected in such a way that tubulin integration into the MT lattice is hampered.

However, one of the main drawbacks of colchicine is its general toxicity (Wallace, 1974; Finkelstein et al., 2010). Several studies in the past have proposed less toxic colchicine derivatives as an alternative to colchicine (Lu et al., 2012; Wang et al., 2016; Johnson et al., 2017; Majcher et al., 2018a,b; Klejborowska et al., 2019). Moreover, these novel colchicine derivatives may be designed to show high specificity only for tubulin isotypes, which are over-expressed in cancer, in order to maximize their effect only on tumor cells and reduce side effects of the drug due to its toxicity on normal cells (Lu and Luduena, 1994; Luduena et al., 1995).

Differing in point or restricted sequence variations, several tubulin isotypes (Leandro-García et al., 2010) are differently expressed by cells under both physiological and pathological conditions. For example, tubulin isotype αβIII is considered as an excellent target for anti-tumoral therapies because it is over-expressed in tumoral cells and it is less widespread than other isotypes, such as αβI, αβII and αβIV, in normal cells (Ferlini et al., 2007; Tseng et al., 2010). Moreover, an over-expression of tubulin isotype αβIII by cancer cells is considered as one among several known drug resistance mechanisms (Derry et al., 1997; Ludueña, 1998; Katsetos et al., 2003; Kamath et al., 2005; Seve, 2005; Ferlini et al., 2007; Sève and Dumontet, 2008; Tseng et al., 2010). Thereby, it is of primary importance to identify specific compounds, which selectively target isotype αβIII.

In this context, computational molecular modeling techniques, such as molecular dynamics (MD) and molecular docking, represent powerful tools to shed light on the molecular mechanisms concerning protein functions and their interaction between different ligands and a specific receptor (Lepre et al., 2017; Omar et al., 2018; Brogi, 2019; Sirous et al., 2019). These computational methods can be applied to investigate the action of different ligands on tubulin dimers (Mitra and Sept, 2008; Natarajan and Senapati, 2012; Gajewski et al., 2013; Kumbhar et al., 2016). Computational drug discovery may help to accelerate and economize the drug discovery process as a complementary tool for experimental research of novel inhibitors.

In this work, ensemble molecular docking, molecular dynamics simulations, and binding energy estimation methods have been employed to characterize the binding of 55 novel colchicine derivatives to the βIII tubulin isotype. We have identified an interesting 4-chlorine thiocolchicine derivative characterized by similar affinity but a different mode of binding to tubulin with respect to its parent compound, colchicine. The main findings of our study indicate this ligand as a promising candidate to overcome colchicine drawbacks and provide information for further developments in designing more selected and specific colchicine derivatives with an intended use as cancer chemotherapy agents.

MATERIALS AND METHODS

Atomic Models of Investigated Compounds

Several series of novel colchicine derivatives (Majcher et al., 2018a,b; Klejborowska et al., 2019) were considered in this work. All 55 compounds have shown *in vitro* anti-proliferative effects on normal and cancer cells. In particular, they were tested on human lung adenocarcinoma, human breast adenocarcinoma, human colon adenocarcinoma cell lines and a doxorubicin-resistant subline (Majcher et al., 2018a,b; Klejborowska et al., 2019).

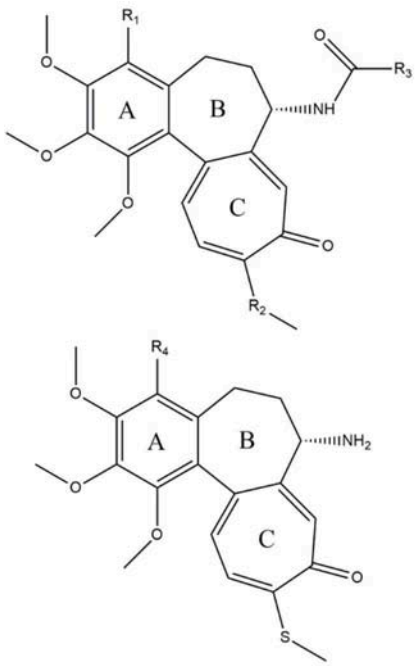
These compounds can be divided into five classes: 4-Br-Amides (10 compounds), 4-Cl-Amides (10 compounds), DT-and-4I-Amides (19 compounds), 4-Cl-Carbamates (8 compounds) and 4-I-Carbamates (8 compounds). The chemical structures of colchicine (C01) and its derivatives (C02-C56) are summarized in **Figure 1**.

The 2D structures of the colchicine derivatives have been drawn using ChemDraw 12.0, whereas their 3D structure was designed by AVOGADRO (Hanwell et al., 2012).

Human αβIII Tubulin Modeling and Conformational Dynamics

The atomic structures of human βIII tubulin isotype were obtained by homology modeling, starting from the Protein Data Bank (PDB) entry 4O2B model (Prota et al., 2014) as a template. This structure was chosen due to its high resolution (2.3 Å) and a low number of missing residues (Aryapour et al., 2017). First, from the starting template the information concerning tubulin αβ, GTP, GDP, Mg^{2+} ion and colchicine was extracted. Missing residues in β tubulin (from 276 to 281) were added by MODELER 9.20 (Šali and Blundell, 1993) where the best model was selected on the basis of the obtained DOPE (Discrete optimized protein energy) score. Then, the Fasta sequences Q71U36 and Q13509 were selected from the Uniprot website, respectively, for the α and β subunits. The above-mentioned amino acid sequences pertain to the isotype αβIII (Gajewski et al., 2013; Kumbhar et al., 2016). Homology modeling was then employed by MODELER 9.20 to generate a 3D structure of the αβIII sequence using the 4O2B model. The quality and the reliability of the generated model were evaluated using PROCHECK (Laskowski et al., 1993), VERIFY3D (Colovos and Yeates, 1993) and ERRAT (Bowie et al., 1991), as reported in previous literature in this area (Huzil et al., 2006; Deriu et al., 2007; Mane et al., 2008; Kumbhar et al., 2016).

Two systems were subsequently considered: (I) tubulin, GTP, GDP, and Mg^{2+} ion and (II) tubulin, GTP, GDP, Mg^{2+} ion and colchicine bound to tubulin. Information on colchicine binding



C01 $R_1 = H$ $R_2 = O$ $R_3 = CH_3$
C02 $R_1 = Br$ $R_2 = O$ $R_3 = CH_3$
C03 $R_1 = Br$ $R_2 = S$ $R_3 = CH_3$
C04 $R_4 = Br$
C05 $R_1 = Br$ $R_2 = S$ $R_3 = CH_2\text{-}O\text{-}CH_3$
C06 $R_1 = Br$ $R_2 = S$ $R_3 = CH_2\text{-}CH_2\text{-}CH_2\text{-}Cl$
C07 $R_1 = Br$ $R_2 = S$ $R_3 = Ph$
C08 $R_1 = Br$ $R_2 = S$ $R_3 = CH_2\text{-}CH_3$
C09 $R_1 = Br$ $R_2 = S$ $R_3 = CH(\text{-}CH_3)_2$
C10 $R_1 = Br$ $R_2 = S$ $R_3 = (CH_2)_8\text{-}CH_3$
C11 $R_1 = Br$ $R_2 = S$ $R_3 = N(\text{-}CH_2\text{-}CH_3)_2$
C12 $R_1 = Cl$ $R_2 = O$ $R_3 = CH_3$
C13 $R_1 = Cl$ $R_2 = S$ $R_3 = CH_3$
C14 $R_4 = Cl$
C15 $R_1 = Cl$ $R_2 = S$ $R_3 = CH_2\text{-}O\text{-}CH_3$
C16 $R_1 = Cl$ $R_2 = S$ $R_3 = CH_2\text{-}CH_2\text{-}CH_2\text{-}Cl$
C17 $R_1 = Cl$ $R_2 = S$ $R_3 = Ph$
C18 $R_1 = Cl$ $R_2 = S$ $R_3 = CH_2\text{-}CH_3$
C19 $R_1 = Cl$ $R_2 = S$ $R_3 = CH(\text{-}CH_3)_2$
C20 $R_1 = Cl$ $R_2 = S$ $R_3 = (CH_2)_8\text{-}CH_3$
C21 $R_1 = Cl$ $R_2 = S$ $R_3 = N(\text{-}CH_2\text{-}CH_3)_2$
C22 $R_1 = H$ $R_2 = S$ $R_3 = CH_3$
C23 $R_4 = H$
C24 $R_1 = H$ $R_2 = S$ $R_3 = CH_2\text{-}O\text{-}CH_3$
C25 $R_1 = H$ $R_2 = S$ $R_3 = CH_2\text{-}CH_2\text{-}CH_2\text{-}Cl$
C26 $R_1 = H$ $R_2 = S$ $R_3 = Ph$
C27 $R_1 = H$ $R_2 = S$ $R_3 = CH_2\text{-}CH_3$
C28 $R_1 = H$ $R_2 = S$ $R_3 = CH(\text{-}CH_3)_2$
C29 $R_1 = H$ $R_2 = S$ $R_3 = (CH_2)_8\text{-}CH_3$
C30 $R_1 = H$ $R_2 = S$ $R_3 = N(\text{-}CH_2\text{-}CH_3)_2$
C31 $R_1 = I$ $R_2 = O$ $R_3 = CH_3$
C32 $R_1 = I$ $R_2 = S$ $R_3 = CH_3$
C33 $R_4 = I$
C34 $R_1 = I$ $R_2 = S$ $R_3 = CH_2\text{-}O\text{-}CH_3$
C35 $R_1 = I$ $R_2 = S$ $R_3 = CH_2\text{-}CH_2\text{-}CH_2\text{-}Cl$
C36 $R_1 = I$ $R_2 = S$ $R_3 = Ph$
C37 $R_1 = I$ $R_2 = S$ $R_3 = CH_2\text{-}CH_3$
C38 $R_1 = I$ $R_2 = S$ $R_3 = CH(\text{-}CH_3)_2$
C39 $R_1 = I$ $R_2 = S$ $R_3 = (CH_2)_8\text{-}CH_3$
C40 $R_1 = I$ $R_2 = S$ $R_3 = N(\text{-}CH_2\text{-}CH_3)_2$
C41 $R_1 = Cl$ $R_2 = S$ $R_3 = O\text{-}CH_2\text{-}CH_2\text{-}OH$
C42 $R_1 = Cl$ $R_2 = S$ $R_3 = O\text{-}CH_3$
C43 $R_1 = Cl$ $R_2 = S$ $R_3 = O\text{-}CH_2\text{-}CH_3$
C44 $R_1 = Cl$ $R_2 = S$ $R_3 = O\text{-}CH_2\text{-}CF_3$
C45 $R_1 = Cl$ $R_2 = S$ $R_3 = O\text{-}CH_2\text{-}CH_2\text{-}CH_3$
C46 $R_1 = Cl$ $R_2 = S$ $R_3 = O\text{-}CH(\text{-}CH_3)_2$
C47 $R_1 = Cl$ $R_2 = S$ $R_3 = O\text{-}CH_2\text{-}CH_2\text{-}CH_2\text{-}CH_3$
C48 $R_1 = Cl$ $R_2 = S$ $R_3 = O\text{-}CH_2\text{-}CH_2\text{-}O\text{-}CH_2\text{-}CH_2\text{-}OH$
C49 $R_1 = I$ $R_2 = S$ $R_3 = O\text{-}CH_2\text{-}CH_2\text{-}OH$
C50 $R_1 = I$ $R_2 = S$ $R_3 = O\text{-}CH_3$
C51 $R_1 = I$ $R_2 = S$ $R_3 = O\text{-}CH_2\text{-}CH_3$
C52 $R_1 = I$ $R_2 = S$ $R_3 = O\text{-}CH_2\text{-}CF_3$
C53 $R_1 = I$ $R_2 = S$ $R_3 = O\text{-}CH_2\text{-}CH_2\text{-}CH_3$
C54 $R_1 = I$ $R_2 = S$ $R_3 = O\text{-}CH(\text{-}CH_3)_2$
C55 $R_1 = I$ $R_2 = S$ $R_3 = O\text{-}CH_2\text{-}CH_2\text{-}CH_2\text{-}CH_3$
C56 $R_1 = I$ $R_2 = S$ $R_3 = O\text{-}CH_2\text{-}CH_2\text{-}O\text{-}CH_2\text{-}CH_2\text{-}CH_3$

FIGURE 1 | Colchicine and its derivatives considered in this work.

was taken from the 4O2B model. The AMBER ff99SB-ILDN forcefield (Lindorff-Larsen et al., 2010) was used to describe protein, water and ion topology. GTP, GDP, and ligands were described by the General Amber Force Field (GAFF) (Wang et al., 2004) and AM1-BCC charge method (Jakalian et al., 2002), as applied in many previous studies (Gajewski et al., 2013; Kumbhar et al., 2016; Klejborowska et al., 2019; Sahakyan et al., 2019).

Each of the above mentioned protein systems (I and II) was then inserted in a dodecahedron box filled with TIP3P explicit water molecules (Mark and Nilsson, 2001), sodium and chlorine ions (150 mM). Particle-mesh Ewald (PME) method (Darden et al., 1993) was used to treat electrostatics (cut-off = 1.0 nm) whereas Van der Waals (VdW) interactions were treated by a plain cut-off at 1 nm (Natarajan and Senapati, 2012; Natarajan et al., 2013; Bueno et al., 2018). Each system was then energy minimized by the steepest descent algorithm for 1,000 steps with a maximum force of 100 $kJmol^{-1}nm^{-1}$. All systems were simulated in an NVT and NPT ensemble with position restraints applied on protein and ligand atoms. In detail, a 100 ps position restrained MD simulation in the NVT ensemble (Bussi et al., 2007), was followed by a 300 ps position restrained MD in the NPT ($T = 300$ K, $P = 1$ bar) ensemble (Berendsen et al., 1984; Bussi et al., 2007).

Temperature and pressure were controlled by weak coupling algorithms (Berendsen et al., 1984; Bussi et al., 2007).

Finally, production MD simulations (without restraints) were carried out for 100 ns in presence and in absence of colchicine, respectively. Ten configurations of each system were extracted as representative of structural equilibrium. The above-mentioned system configurations were then used for ensemble docking procedure.

Ensemble Docking and Binding Energy Refinement

Ensemble docking was performed using AUTODOCK VINA 1.1.2 (Trott and Olson, 2010). The center of the search space was defined by taking, from the 4O2B model, the relative position of the colchicine in its binding site. The docking was performed using a grid space 2 × 2 × 2 nm around the center of the search space and an exhaustiveness equal to 64 was set. Each compound was docked to the ten different isotype configurations extracted from the production MD, as explained above.

Then, for each VINA pose, the binding energy refinement was performed by running short 1 ns MD simulations on the ligand-protein complex starting from the VINA best pose for each considered ligand. Each complex was followed by solvation, neutralization, energy minimization, position restrained MD, and short production MD. Simulation set up was the same as described in the previous section. On the last 100 ps of MD production the ligand-protein binding was evaluated using two criteria. Firstly, the binding energy was quantified by the Molecular Mechanics Generalized Born Surface Area (MMGBSA) method (Genheden and Ryde, 2015). The parameters were set according to the previous literature (Nguyen et al., 2013, 2015; Su et al., 2015). Secondly, the ligand conformational displacement in the binding site was quantified by calculating the Root Mean Square Deviation (RMSD) of ligand carbon rings (a common feature of all considered compounds with colchicine). In particular, for each ligand, the MD protein trajectory was fitted on a reference structure (the starting configuration of the colchicine-protein complex). In this way, the RMSD quantifies the relative deviation of each ligand with respect to the colchicine starting position throughout the overall

MD trajectory. Based on the above-mentioned criteria, the best colchicine derivative and colchicine, both bound to the βIII isotype, were simulated for 100 ns in order to highlight binding conformational differences at equilibrium.

All MD simulations were carried out using GROMACS 2018.3 (Abraham et al., 2015). The Visual Molecular Dynamics (VMD) package was employed for the visual inspection of the simulated systems (Humphrey et al., 1996). Dedicated GROMACS tools were used for a quantitative analysis in terms of Root-Mean-Square Deviation (RMSD), Root-Mean-Square Fluctuation (RMSF), and clustering, while analysis of the secondary structure was performed by applying the STRuctural IDEntification (STRIDE) algorithm (Heinig and Frishman, 2004).

RESULTS

Human αβIII Tubulin Model Development and Conformational Dynamics

The Ramachandran plot (see also **Figure S1**) obtained by PROCHECK highlighted the 95.6% of residues in most favored regions, 4.2% in additional allowed regions, and 0.1% in generously allowed regions. No residues were found in disallowed regions. Since a good quality model is expected to have at least 90% of the residues in the most favored regions (Santoshi and Naik, 2014), the built model was considered reliable. Moreover, the Overall Quality Factor obtained by the ERRAT tool for the isotype αβIII was 80.29 for the α and 84.73 for the β tubulin monomer model. It is worth mentioning that the generally accepted range is higher than 50 for a high quality model (Messaoudi et al., 2013). Finally, the VERIFY3D confirmed that 98.15% of residues showed an averaged 3D-1D score higher than 0.2 (Messaoudi et al., 2013).

First, the backbone RMSD was calculated for isotype βIII both in presence and in absence of colchicine during the overall MD simulation (100 ns): all the simulated structures reached structural equilibrium, with values under 0.3 nm (see also **Figure S2**). Moreover, the cluster analysis on the last 50 ns of the simulations highlighted only one cluster using an RMSD cut-off of 0.15 nm, indicating a strong stability of the simulated systems. Moreover, the cluster analysis indicated that the colchicine presence did not modify significantly the conformation of the interaction site.

Ensemble Docking and Binding Energy Calculation

The 55 colchicine derivatives were docked to ten different configurations of βIII tubulin, extracted from the last 50 ns of the MD simulation described above. Only the best ligand pose in terms of binding affinity was considered (see also **Figure S3**). In order to take into account also the dynamic nature of the binding process, we have performed a MD simulation of 1 ns for each ligand-receptor complex. Throughout the quick MD run, the binding energy was quantified by means of the MM-GBSA method (Huzil et al., 2010; Gajewski et al., 2013; Kumbhar et al., 2016). Moreover, the ligand displacement in the binding site was quantified by the RMSD calculated as described in Materials and Methods. It is worth mentioning that low RMSD values indicate a compound which is stable in a spot close to the starting colchicine position, whereas high RMSD values identify a compound moving further apart (**Figure 2**). Most compounds showed RMSD lower than 0.2 nm, suggesting that the derivatives investigated here behaves similarly to colchicine (highly stable in its binding site during the short MD run). The only exception found is represented by compound C19 which displays high variation from the colchicine starting position (RMSD = 0.47 nm).

Our binding energy analysis highlights four specific compounds (C19, C20, C29, and C48) as possible hits. In fact, they exhibit similar values of their binding energy for βIII tubulin compared to colchicine. All binding energy values are reported in **Supporting Information** text (see also **Figure S4**). In order to better describe differences between investigated compounds, we have merged RMSD and binding energy information in a single plot (**Figure 3**).

Interestingly, compound C19 features a peculiar behavior, i.e., it exhibits a large deviation from the colchicine starting position in the binding site (high RMSD value), with a significant difference to all the other compounds. It is also characterized by binding energy values comparable to colchicine and higher than most other derivatives. This result points the attention on the compound C19 as a promising candidate able to bind strongly to βIII human tubulin with a different mode of action with respect to colchicine.

Conformational Dynamics of Colchicine and C19 Bound to βIII Human Tubulin

Conformational dynamics of the C19-tubulin complex has been investigated by a 100 ns long MD simulation. For comparison, a 100 ns-long MD was also carried out on the colchicine-tubulin complex. Systems were replicated to confirm the consistency of the data (**Figures S5**, **S7**).

Structural modifications of the colchicine binding site were first analyzed by computing the RMSD of the tubulin binding cleft, i.e., residues within 1 nm from the ligand, from its starting

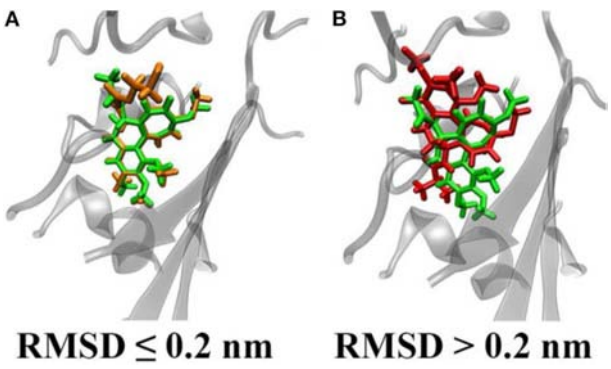


FIGURE 2 | Representative snapshots of ligand conformational displacements in the colchicine binding site. Colchicine is represented in green, whereas two different derivatives with low **(A)** and high **(B)** RMSD with respect to the colchicine starting pose are depicted orange and red, respectively.

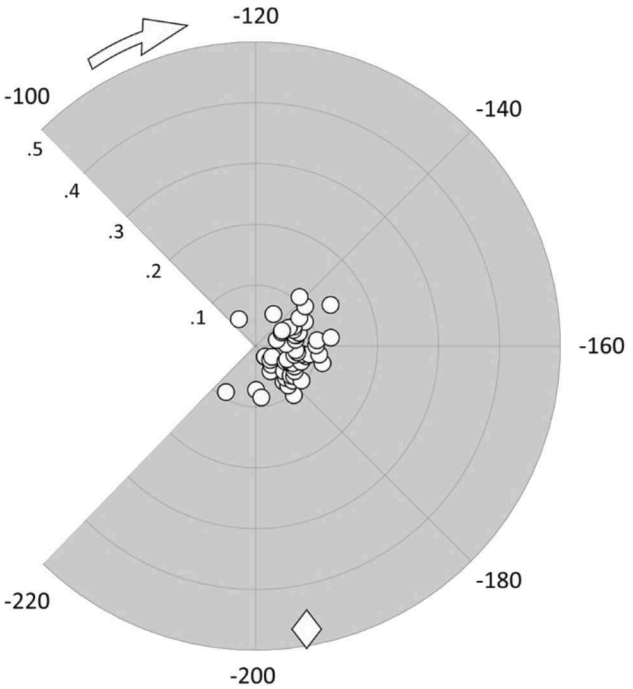


FIGURE 3 | Polar scatter plot representing ligands' RMSD from the colchicine starting pose (radial coordinate) in nm and their binding energy (angular coordinate) in kJ/mol; the diamond marker represents the C19 compound.

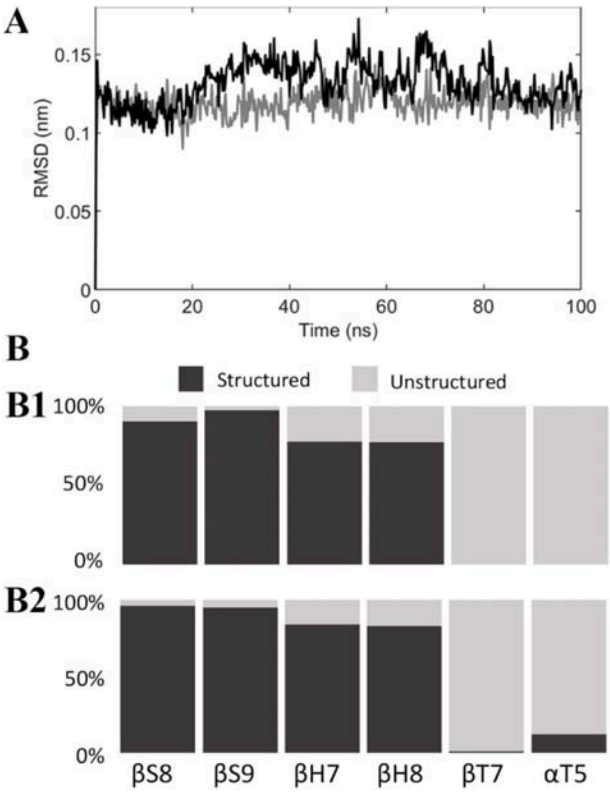


FIGURE 4 | (A) RMSD of the colchicine binding site from its starting position when colchicine (black) or compound C19 (gray) are bound to the tubulin dimer. **(B)** Secondary structure probability of residues in the colchicine binding site when colchicine **(B1)** or compound C19 **(B2)** are bound to the tubulin dimer.

position and the secondary structure probability during the last 50 ns of the simulation (**Figure 4**). A second replica was performed to ensure the repeatability of the results (see also **Figure S5**). The binding site was characterized by a structural stability throughout the overall MD, exhibiting low RMSD values (lower than 0.18 nm) and highly conserved secondary structures. The only noteworthy difference is represented by the α-T5 loop, which exhibits tendency to rearrange in a more structured shape only in the presence of C19.

The binding energy was estimated by the MM-GBSA approach in order to compare the binding affinities of the analyzed compounds at the structural equilibrium: again, compound C19 and colchicine showed similar binding energy for isotype αβIII, respectively, −229.98 ± 22.26 kJ/mol and −223.70 ± 22.31 kJ/mol (see also **Figure S6**). Nevertheless, the energy decomposition over the tubulin binding cleft residues reveals that the compound C19 shows a higher binding energy compared to the colchicine for residues 178–180 of the α tubulin, which belong to the αT5 loop (**Figure 5**).

In light of these results, the ligands' behavior in the binding site and their interaction with the αT5 loop were investigated in more depth (**Figure 6**). First, ligand RMSD (used to quantify the ligand movement in the binding site throughout the simulation) showed that compound C19 has a more marked tendency than colchicine to move apart, reaching a more favorable pose for the interaction with the αT5 loop (**Figure 6A** and see also **Figure S7**). Second, the interaction surface between each ligand and the αT5 loop, which quantifies the available area for their binding, is higher for C19 than colchicine (**Figure 6B**). **Figures 6C1,C2** and **6D** represent ligand structures and their relative position in the tubulin binding cleft (see also **Movies S1**, **S2**).

In conclusion, compound C19 was shown to be stable in the tubulin binding site with a relative position differing from the colchicine site. Specifically, C19 is predicted to be mostly stabilized by its interaction with the αT5 loop.

DISCUSSION

In this study, 55 colchicine derivatives were screened for their binding properties to tubulin isotype βIII. The research work was aimed at identifying alternative compounds able to overcome colchicine's well-known limitations. After the docking of all compounds to the target isotype of tubulin, a molecular dynamics simulation of 1 ns was performed on each generated receptor-ligand complex. The obtained trajectories were analyzed considering the deviations of the compounds from the colchicine's starting pose, using the RMSD, and the binding energy evaluated with the MM-GBSA method. All compounds were characterized by low RMSD values, except for compound C19, which showed high deviations (RMSD = 0.47 nm). This evidence suggests a different particular pose for this derivative. From the affinity analysis we found out that the binding energies for compounds C19, C20, C29, and C48 are similar to that for

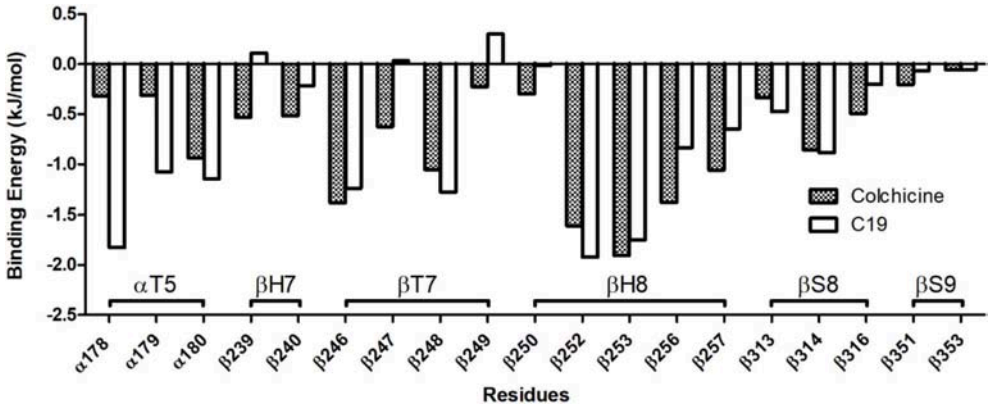


FIGURE 5 | Binding Energy decomposition over the residues of the colchicine binding site (only residues with the highest energies are reported). Compound C19 has a significantly higher affinity than colchicine for the αT5 loop.

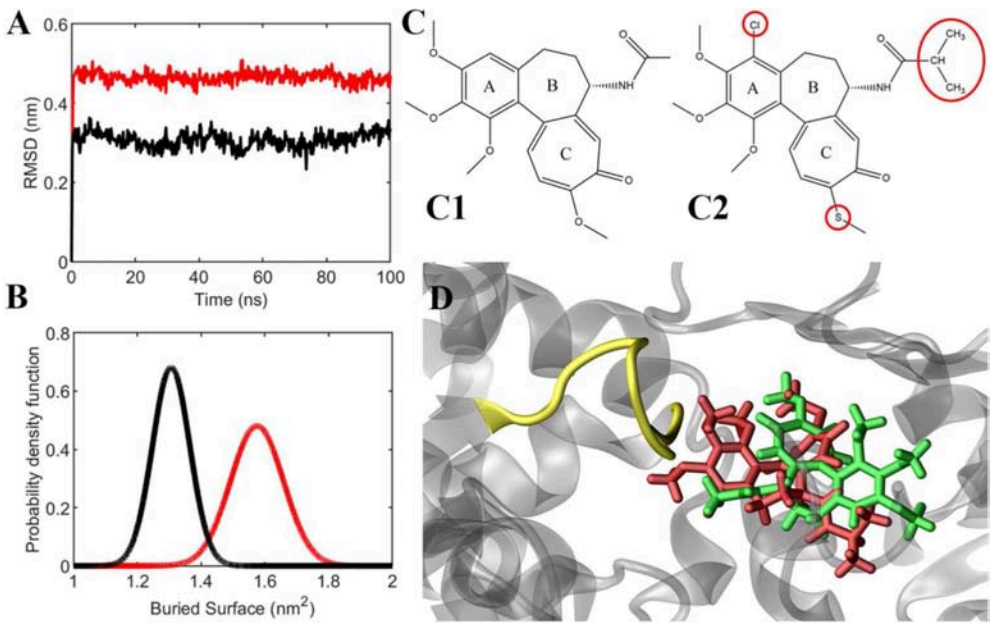


FIGURE 6 | (A) Ligands' RMSD from their starting position (colchicine in black, C19 in red). **(B)** Probability density function of the buried surface between the ligands and the αT5 loop (colchicine in black, C19 in red), averaged between two replicas during the last 20 ns of simulation. **(C)** Chemical structures of colchicine **(C1)** and compound C19 **(C2)**. **(D)** representative snapshot of the simulation, which shows that compound C19 (red) is closer to αT5 loop (yellow) than colchicine (green).

colchicine and higher than those found for most other ligands. These results indicate that C19 is a promising compound to be further investigated and experimentally validated. Its specific binding to tubulin is characterized by a different conformational organization and dynamics in the tubulin binding site with high affinity. RMSD analysis indicates that C19 is able to be accommodated in the binding site by moving toward more favorable poses for interaction with the αT5 loop. This feature is less pronounced by colchicine. Moreover, the buried surface between C19 and the tubulin isotype βIII, which measures the available area for the binding, is greater than the one exhibited with colchicine, confirming a higher stability of C19 in the binding site. Finally, the ligand binding to the αT5 loop may affect its secondary structure toward a more structured arrangement. Therefore, a compound able to influence the αT5 loop structure could affect the dynamics of the entire microtubule.

The above mentioned evidences might be of a significant interest given that the αT5 loop is a key player region in the colchicine binding site and for intra-dimer contacts (Ravelli et al., 2004). Nonetheless, previous literature (Bueno et al., 2018) already highlighted the importance of the αT5 loop, identified as relevant for the binding of a promising anti-proliferative compound (Bueno et al., 2018).

In conclusion, our study clarifies some features characterizing the βIII tubulin binding mode of a promising novel 4-chlorine

thiocolchicine derivative, which differs profoundly from that known for colchicine. The specific interaction of compound C19 with the αT5 loop is a promising feature that could be related to an increased destabilizing activity of the ligand with respect to microtubule dynamics. Moreover, this unique behavior exhibited in complex with the βIII tubulin isotype is of primary importance since this isotype is overexpressed in cancer cells, while very insignificantly represented in most normal cells and also implicated in drug resistance (Katsetos et al., 2003; Kamath et al., 2005; Seve, 2005; Sève and Dumontet, 2008; Leandro-García et al., 2010). In light of these results, C19 or similar compounds, as promising candidates able to possibly overcome some colchicine's drawbacks, deserve further investigations, including biological toxicity assessment and cancer cell cytotoxicity experiments to prove its specificity and selectivity for βIII isotype of tubulin.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

MD, JT, and AH conceived the research. LP, AR, and GG did the molecular dynamics simulations. LP, AR, GG, and GK analyzed and rationalized the data. All authors wrote the paper and critically commented to the manuscript, read, and approved the final manuscript.

ACKNOWLEDGMENTS

Financial support by grant of the Polish National Science Centre (NCN)–No. 2016/21/B/ST5/00111 is gratefully acknowledged. GK gratefully acknowledges the doctoral scholarship ETIUDA–No. 2018/28/T/ST5/00041 financed by the Polish National Science Centre (NCN).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2020.00108/full#supplementary-material

REFERENCES

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2, 19–25. doi: 10.1016/j.softx.2015.06.001

Aryapour, H., Dehdab, M., Sohraby, F., and Bargahi, A. (2017). Prediction of new chromene-based inhibitors of tubulin using structure-based virtual screening and molecular dynamics simulation methods. *Comput. Biol. Chem.* 71, 89–97. doi: 10.1016/j.compbiolchem.2017.09.007

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81, 3684–3690. doi: 10.1063/1.448118

Bhattacharyya, B., Panda, D., Gupta, S., and Banerjee, M. (2008). Anti-mitotic activity of colchicine and the structural basis for its interaction with tubulin. *Med. Res. Rev.* 28, 155–183. doi: 10.1002/med.20097

Bowie, J., Luthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253, 164–170. doi: 10.1126/science.1853201

Brogi, S. (2019). Computational approaches for drug discovery. *Molecules* 24:3061. doi: 10.3390/molecules24173061

Bueno, O., Estévez Gallego, J., Martins, S., Prota, A. E., Gago, F., Gómez-SanJuan, A., et al. (2018). High-affinity ligands of the colchicine domain in tubulin based on a structure-guided design. *Sci. Rep.* 8:4242. doi: 10.1038/s41598-018-22382-x

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126:014101. doi: 10.1063/1.2408420

Colovos, C., and Yeates, T. O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.* 2, 1511–1519. doi: 10.1002/pro.5560020916

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh ewald: an *N* log(*N*) method for ewald sums in large systems. *J. Chem. Phys.* 98:10089. doi: 10.1063/1.464397

Deriu, M. A., Enemark, S., Soncini, M., Montevecchi, F. M., and Redaelli, A. (2007). Tubulin: from atomistic structure to supramolecular mechanical properties. *J. Mater. Sci.* 42, 8864–8872. doi: 10.1007/s10853-007-1784-6

Derry, W. B., Wilson, L., Khan, I. A., Ludueña, R. F., and Jordan, M. A. (1997). Taxol differentially modulates the dynamics of microtubules assembled from unfractionated and purified β-tubulin isotypes. *Biochemistry* 36, 3554–3562. doi: 10.1021/bi962724m

Ferlini, C., Raspaglio, G., Cicchillitti, L., Mozzetti, S., Prislei, S., Bartollino, S., et al. (2007). Looking at drug resistance mechanisms for microtubule interacting drugs: does TUBB3 work? *Curr. Cancer Drug Targets* 7, 704–712. doi: 10.2174/156800907783220453

Finkelstein, Y., Aks, S. E., Hutson, J. R., Juurlink, D. N., Nguyen, P., Dubnov-Raz, G., et al. (2010). Colchicine poisoning: the dark side of an ancient drug. *Clin. Toxicol.* 48, 407–414. doi: 10.3109/15563650.2010.495348

Gajewski, M. M., Tuszynski, J. A., Barakat, K., Torin Huzil, J., and Klobukowski, M. (2013). Interactions of laulimalide, peloruside, and their derivatives with the isoforms of β-tubulin. *Can. J. Chem.* 91, 511–517. doi: 10.1139/cjc-2012-0360

Genheden, S., and Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Exp. Opin. Drug Discov.* 10, 449–461. doi: 10.1517/17460441.2015.1032936

Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E., and Hutchison, G. R. (2012). Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminform.* 4:17. doi: 10.1186/1758-2946-4-17

Heinig, M., and Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* 32, W500–W502. doi: 10.1093/nar/gkh429

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graph.* 14, 33–38. doi: 10.1016/0263-7855(96)00018-5

Huzil, J. T., Ludueña, R. F., and Tuszynski, J. (2006). Comparative modelling of human β tubulin isotypes and implications for drug binding. *Nanotechnology* 17, S90–S100. doi: 10.1088/0957-4484/17/4/014

Huzil, J. T., Mane, J., and Tuszynski, J. A. (2010). Computer assisted design of second-generation colchicine derivatives. *Interdiscip. Sci. Comput. Life Sci.* 2, 169–174. doi: 10.1007/s12539-010-0076-z

Jakalian, A., Jack, D. B., and Bayly, C. I. (2002). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. parameterization and validation. *J. Comput. Chem.* 23, 1623–1641. doi: 10.1002/jcc.10128

Johnson, L., Goping, I. S., Rieger, A., Mane, J. Y., Huzil, T., Banerjee, A., et al. (2017). Novel colchicine derivatives and their anti-cancer activity. *Curr. Top. Med. Chem.* 17, 2538–2558. doi: 10.2174/1568026617666170104143618

Kamath, K., Wilson, L., Cabral, F., and Jordan, M. A. (2005). βIII-tubulin induces paclitaxel resistance in association with reduced effects on microtubule dynamic instability. *J. Biol. Chem.* 280, 12902–12907. doi: 10.1074/jbc.M414477200

Katsetos, C. D., Herman, M. M., and Mörk, S. J. (2003). Class III β-tubulin in human development and cancer. *Cell Motil. Cytoskelet.* 55, 77–96. doi: 10.1002/cm.10116

Klejborowska, G., Urbaniak, A., Preto, J., Maj, E., Moshari, M., Wietrzyk, J., et al. (2019). Synthesis, biological evaluation and molecular docking studies of new amides of 4-bromothiocolchicine as anticancer agents. *Bioorg. Med. Chem.* 76:115144. doi: 10.1016/j.bmc.2019.115144

Kumbhar, B. V., Borogaon, A., Panda, D., and Kunwar, A. (2016). Exploring the origin of differential binding affinities of human tubulin isotypes αβII, αβIII and αβIV for DAMA-colchicine using homology modelling, molecular docking and molecular dynamics simulations. *PLoS ONE* 11:e0156048. doi: 10.1371/journal.pone.0156048

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* 26, 283–291. doi: 10.1107/S0021889892009944

Leandro-García, L. J., Leskelä, S., Landa, I., Montero-Conde, C., López-Jiménez, E., Letón, R., et al. (2010). Tumoral and tissue-specific expression of the major human β-tubulin isotypes. *Cytoskeleton* 67, 214–223. doi: 10.1002/cm.20436

Lepre, M., Omar, S., Grasso, G., Morbiducci, U., Deriu, M., and Tuszynski, J. (2017). Insights into the effect of the G245S single point mutation on the structure of p53 and the binding of the protein to DNA. *Molecules* 22:E1358. doi: 10.3390/molecules22081358

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., et al. (2010). Improved side-chain torsion potentials for the amber ff99SB protein force field. *Proteins* 78, 1950–1958. doi: 10.1002/prot.22711

Lu, Q., and Luduena, R. F. (1994). *In vitro* analysis of microtubule assembly of isotypically pure tubulin dimers: intrinsic differences in the assembly properties of alpha beta II, alpha beta III, and alpha beta IV tubulin dimers in the absence of microtubule-associated proteins. *J. Biol. Chem.* 269, 2041–2047.

Lu, Y., Chen, J., Xiao, M., Li, W., and Miller, D. D. (2012). An overview of tubulin inhibitors that interact with the colchicine binding site. *Pharm. Res.* 29, 2943–2971. doi: 10.1007/s11095-012-0828-z

Ludueña, R. F. (1998). Multiple forms of tubulin: different gene products and covalent modifications. *Int. Rev. Cytol.*178, 207–275. doi: 10.1016/S0074-7696(08)62138-5

Luduena, R. F., Roach, M. C., Prasad, V., Chaudhuri, A. R., Tomita, I., Mizuhashi, F., et al. (1995). Interaction of bovine brain tubulin with the 4(1H)-pyrizinone derivative IKP104, an antimitotic drug with a complex set of effects on the conformational stability of the tubulin molecule. *Biochemistry* 34, 15751–15759. doi: 10.1021/bi00048a020

Majcher, U., Klejborowska, G., Moshari, M., Maj, E., Wietrzyk, J., Bartl, F., et al. (2018a). Antiproliferative activity and molecular docking of novel double-modified colchicine derivatives. *Cells* 7:192. doi: 10.3390/cells7110192

Majcher, U., Urbaniak, A., Maj, E., Moshari, M., Delgado, M., Wietrzyk, J., et al. (2018b). Synthesis, antiproliferative activity and molecular docking of thiocolchicine urethanes. *Bioorg. Chem.* 81, 553–566. doi: 10.1016/j.bioorg.2018.09.004

Mane, J. Y., Klobukowski, M., Huzil, J. T., and Tuszynski, J. (2008). Free energy calculations on the binding of colchicine and its derivatives with the α/β-tubulin isoforms. *J. Chem. Inf. Model.* 48, 1824–1832. doi: 10.1021/ci800054n

Mark, P., and Nilsson, L. (2001). Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *J. Phys. Chem. A* 105, 9954–9960. doi: 10.1021/jp003020w

Messaoudi, A., Belguith, H., and Ben Hamida, J. (2013). Homology modeling and virtual screening approaches to identify potent inhibitors of VEB-1 β-lactamase. *Theor. Biol. Med. Model.* 10, 1–10. doi: 10.1186/1742-4682-10-22

Mitra, A., and Sept, D. (2008). Taxol allosterically alters the dynamics of the tubulin dimer and increases the flexibility of microtubules. *Biophys. J.* 95, 3252–3258. doi: 10.1529/biophysj.108.133884

Natarajan, K., Mohan, J., and Senapati, S. (2013). Relating nucleotide-dependent conformational changes in free tubulin dimers to tubulin assembly. *Biopolymers* 99, 282–291. doi: 10.1002/bip.22153

Natarajan, K., and Senapati, S. (2012). Understanding the basis of drug resistance of the mutants of αβ-tubulin dimer via molecular dynamics simulations. *PLoS ONE* 7:e42351. doi: 10.1371/journal.pone.0042351

Nettles, J. H., Snyder, J. P., Nogales, E., Cornett, B., and Downing, K. H. (2002). The binding conformation of taxol in -tubulin: a model based on electron crystallographic density. *Proc. Natl. Acad. Sci. U.S.A.* 98, 5312–5316. doi: 10.1073/pnas.051309398

Nguyen, H., Pérez, A., Bermeo, S., and Simmerling, C. (2015). Refinement of generalized born implicit solvation parameters for nucleic acids and their complexes with proteins. *J. Chem. Theory Comput.* 11, 3714–3728. doi: 10.1021/acs.jctc.5b00271

Nguyen, H., Roe, D. R., and Simmerling, C. (2013). Improved generalized born solvent model parameters for protein simulations. *J. Chem. Theory Comput.* 9, 2020–2034. doi: 10.1021/ct3010485

Omar, S. I., Lepre, M. G., Morbiducci, U., Deriu, M. A., and Tuszynski, J. A. (2018). Virtual screening using covalent docking to find activators for G245S mutant p53. *PLoS ONE* 13:e0200769. doi: 10.1371/journal.pone.0200769

Prota, A. E., Danel, F., Bachmann, F., Bargsten, K., Buey, R. M., Pohlmann, J., et al. (2014). The novel microtubule-destabilizing drug BAL27862 binds to the colchicine site of tubulin with distinct effects on microtubule organization. *J. Mol. Biol.* 426, 1848–1860. doi: 10.1016/j.jmb.2014.02.005

Ravelli, R. B. G., Gigant, B., Curmi, P. A., Jourdain, I., Lachkar, S., Sobel, A., et al. (2004). Insight into tubulin regulation from a complex with colchicine and a stathmin-like domain. *Nature* 428, 198–202. doi: 10.1038/nature02393

Sahakyan, H., Abelyan, N., Arakelov, V., Arakelov, G., and Nazaryan, K. (2019). *In silico* study of colchicine resistance molecular mechanisms caused by tubulin structural polymorphism. *PLoS ONE* 14:e0221532. doi: 10.1371/journal.pone.0221532

Šali, A., and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of Spatial restraints. *J. Mol. Biol.* 234, 779–815. doi: 10.1006/jmbi.1993.1626

Santoshi, S., and Naik, P. K. (2014). Molecular insight of isotypes specific β-tubulin interaction of tubulin heterodimer with noscapinoids. *J. Comput. Aided Mol. Des.* 28, 751–763. doi: 10.1007/s10822-014-9756-9

Seve, P. (2005). Class III -tubulin expression in tumor cells predicts response and outcome in patients with non-small cell lung cancer receiving paclitaxel. *Mol. Cancer Ther.* 4, 2001–2007. doi: 10.1158/1535-7163.MCT-05-0244

Sève, P., and Dumontet, C. (2008). Is class III β-tubulin a predictive factor in patients receiving tubulin-binding agents? *Lancet Oncol.* 9, 168–175. doi: 10.1016/S1470-2045(08)70029-9

Sirous, H., Chemi, G., Campiani, G., and Brogi, S. (2019). An integrated *in silico* screening strategy for identifying promising disruptors of p53-MDM2 interaction. *Comput. Biol. Chem.* 83:107105. doi: 10.1016/j.compbiolchem.2019.107105

Su, P. C., Tsai, C. C., Mehboob, S., Hevener, K. E., and Johnson, M. E. (2015). Comparison of radii sets, entropy, QM methods, and sampling on MM-PBSA, MM-GBSA, and QM/MM-GBSA ligand binding energies of *F. tularensis* enoyl-ACP reductase (FabI). *J. Comput. Chem.* 36, 1859–1873. doi: 10.1002/jcc.24011

Trott, O., and Olson, A. J. (2010). AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455–461. doi: 10.1002/jcc.21334

Tseng, C.-Y., Mane, J. Y., Winter, P., Johnson, L., Huzil, T., Izbicka, E., et al. (2010). Quantitative analysis of the effect of tubulin isotype expression on sensitivity of cancer cell lines to a set of novel colchicine derivatives. *Mol. Cancer* 9:131. doi: 10.1186/1476-4598-9-131

Wallace, S. L. (1974). Colchicine. *Semin. Arthritis Rheum.* 3, 369–381. doi: 10.1016/0049-0172(74)90006-7

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *J. Comput. Chem.* 25, 1157–1174. doi: 10.1002/jcc.20035

Wang, Y., Zhang, H., Gigant, B., Yu, Y., Wu, Y., Chen, X., et al. (2016). Structures of a diverse set of colchicine binding site inhibitors in complex with tubulin provide a rationale for drug discovery. *FEBS J.* 283, 102–111. doi: 10.1111/febs.13555

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Pallante, Rocca, Klejborowska, Huczynski, Grasso, Tuszynski and Deriu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 27 February 2020
doi: 10.3389/fchem.2020.00107

Exploring the RNA-Recognition Mechanism Using Supervised Molecular Dynamics (SuMD) Simulations: Toward a Rational Design for Ribonucleic-Targeting Molecules?

*Maicol Bissaro, Mattia Sturlese and Stefano Moro**

Molecular Modeling Section, Department of Pharmaceutical and Pharmacological Sciences, University of Padova, Padua, Italy

OPEN ACCESS

Edited by:
Kamil Kuca,
University of Hradec Králové, Czechia

Reviewed by:
Marco Tutone,
University of Palermo, Italy
Elzbieta Malinowska,
Warsaw University of Technology, Poland

****Correspondence:***
Stefano Moro
stefano.moro@unipd.it

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *29 November 2019*
Accepted: *04 February 2020*
Published: *27 February 2020*

Citation:
Bissaro M, Sturlese M and Moro S (2020) Exploring the RNA-Recognition Mechanism Using Supervised Molecular Dynamics (SuMD) Simulations: Toward a Rational Design for Ribonucleic-Targeting Molecules? Front. Chem. 8:107. doi: 10.3389/fchem.2020.00107

Although proteins have represented the molecular target of choice in the development of new drug candidates, the pharmaceutical importance of ribonucleic acids has gradually been growing. The increasing availability of structural information has brought to light the existence of peculiar three-dimensional RNA arrangements, which can, contrary to initial expectations, be recognized and selectively modulated through small chemical entities or peptides. The application of classical computational methodologies, such as molecular docking, for the rational development of RNA-binding candidates is, however, complicated by the peculiarities characterizing these macromolecules, such as the marked conformational flexibility, the singular charges distribution, and the relevant role of solvent molecules. In this work, we have thus validated and extended the applicability domain of SuMD, an all-atoms molecular dynamics protocol that allows to accelerate the sampling of molecular recognition events on a nanosecond timescale, to ribonucleotide targets of pharmaceutical interest. In particular, we have proven the methodological ability by reproducing the binding mode of viral or prokaryotic ribonucleic complexes, as well as that of artificially engineered aptamers, with an impressive degree of accuracy.

Keywords: nucleic acids, RNA, SMIRNA, molecular recognition, molecular dynamics (MD), supervised molecular dynamics (SuMD), structure-based drug design (SBDD)

INTRODUCTION

Ribonucleic acid (RNA) is a polymer whose biological importance has increased progressively over the last 50 years. Despite the central dogma of molecular biology considering this nucleic acid simply as a functional messenger between DNA genetic information storage and protein biosynthesis, RNA has recently been reappraised as an ancestral molecule of primary importance in the abiogenesis process. At the origin of life, RNA probably encompassed both an informational role, which progressively evolved toward involving the more stable and easily replicable DNA polymer, and a catalytic function, which was gradually flanked by more versatile proteins

(Morris and Mattick, 2014). The complexity hiding behind RNA's biological functions is intuitable by taking into consideration the human organism, which genetic heritage could quite entirely be transcribed into RNA, despite coding only in a minimal portion (about 3%) for proteins (Warner et al., 2018). A great majority of these transcripts therefore remain untranslated, originating non-coding genomic portions. RNA revolution has thus shed light on the regulatory activity of this widely different class of macromolecules that, along with some proteins, cooperate to control and finely orchestrate the genome expression (Connelly et al., 2016).

RNA polymer lengths range from small hairpins composed of a few tens of nucleobases to long non-coding RNAs sequences (lncRNAs) that can reach up to a few thousands nucleotides (Connelly et al., 2016). Differently from DNA, RNA usually exists as a single-stranded molecule that is not strictly limited by a Watson-Crick base pairing. In solution, ribonucleic acids explore a wide landscape of three-dimensional structures, characterizable by the presence of peculiar functional domains able to specifically recognize other nucleic acids, polypeptides, glyco-derivates, or cognates of small organic molecules (Draper, 1995; Cruz and Westhof, 2009; Salmon et al., 2014; Flynn et al., 2019).

From a topological point of view, the tertiary and quaternary structures that distinguish ribonucleic acids from their deoxyribonucleic counterpart make them more similar to proteins, a consideration that has paved the way for an attempt to pharmacologically modulate their biological functions through the discovery of small molecules interacting with RNA (SMIRNA) (Sucheck and Wong, 2000; Connelly et al., 2016). Interestingly, in recent research work, it has been estimated that pharmacologically modulating RNA would allow us to expanding—by more than an order of magnitude—the universe of targetable macromolecules, and this would thus considerably extend the portion of the druggable genome (Ecker and Griffey, 1999; Warner et al., 2018). Although RNA has been historically considered as an "undruggable" pharmaceutical target, the discovery that many drugs of undeniable therapeutic importance, especially antibiotics, act at this level has attracted the interest of the scientific community, resulting in greater effort being made toward the development of new tools for this purpose (Donlic and Hargrove, 2018; Disney, 2019). Furthermore, the orthogonality characterizing RNA homologous transcripts belonging to virus, prokaryote, and eukaryote genomes make RNA an interesting target for the purpose of achieving selectivity, especially in the field of anti-infectives compound development (Ecker and Griffey, 1999; Connelly et al., 2016). All these aspects, therefore, make the discovery of SMIRNAs extremely intriguing. A first pioneering approach to rationally design new RNA-targeting compounds, simply starting from the knowledge of the oligonucleotide sequence of pathological interest, was developed by the Disney research group and was successfully applied to a plethora of expanded repeating RNAs that are known to cause microsatellite disorders (Velagapudi et al., 2014; Disney et al., 2016). In addition, the quantitative structure-activity relationship (QSAR) model and chemical similarity search were initially exploited to *in-silico* identify or optimize new chemical probes targeting RNA (Disney et al., 2014). Since X-ray crystallography, NMR spectroscopy and, recently, Cryo-EM techniques have unveiled with an atomistic level of detail a multitude of three-dimensional RNA structures, the scientific community has begun to evaluate the applicability of structure-based drug design strategies (SBDD). These approaches, until now mainly validated on proteins targets, could enhance the rational design of SMIRNAs. Molecular docking represents one of the electives of *in silico* techniques, exploited both in the academic and industrial world, to accelerate the discovery and optimization of new drug candidates by evaluating the putative small molecules' binding mode and providing a way to perform a ranking of vast compound libraries. There are however many peculiarities of ribonucleic acids that affect both performance and accuracy of docking protocols, and this makes its application challenging. The polyanionic backbone of RNA determines a peculiar charge distribution on the polymer surface—quite different from the one characterizing proteins—to which the scoring functions were traditionally calibrated (Disney, 2019). Furthermore, docking protocols do not explicitly consider the role of solvent during the molecular recognition process, whereas structural data have highlighted how water molecules can stabilize RNA-ligand complexes, often mediating hydrogen bonds networks (Fulle and Gohlke, 2009). However, the aspect that mostly affects RNA-docking accuracy is the flexibility and the dynamic behavior characterizing ribonucleic acids, which are usually neglected by docking algorithms, thus limiting the discovery of compounds targeting a narrow region of the conformational space (Hermann, 2002; Fulle and Gohlke, 2009; Disney et al., 2014). An attempt to overcome these limitations was conducted by Stelzer et al., who performed a docking-based virtual screening on an RNA dynamic ensemble constructed by combining molecular dynamics simulations (MD) with NMR spectroscopy and reported the discovery of six molecules able to bind HIV-1 TAR with quite good affinity. MD simulations would represent a valuable computational tool with which to investigate different ligand–RNA recognition processes, fully considering both target flexibility and the solvent presence. Interestingly, molecular mechanics force fields (FF), such as AMBER or CHARMM, were revisited and refined during the last year to improve ribonucleotide simulation accuracy (Pérez et al., 2007; Denning et al., 2011). Nevertheless, the use of MD is mostly limited to the fluctuation exploration in the post-docking procedure since ligand–target associations are rare events that can be sampled only through long-timescale computationally expensive simulations. An implementation of classical MD, called supervised molecular dynamics (SuMD), was recently developed in our research group. SuMD is able to speed up the exploration of the ligand–receptor recognition pathways on a nanosecond timescale through the implementation of a tabu-like supervision algorithm (Sabbadin and Moro, 2014). The protocol was so far validated in different scenarios, including ion–protein, ligand–protein, and peptide–protein bound complexes, proving that it could reproduce the experimental determined final state with great geometric accuracy (Cuzzolin et al., 2016; Salmaso et al., 2017; Bissaro et al., 2019).

In this work, SuMD simulations were applied for the first time to investigate the recognition mechanism involving

ribonucleic acid macromolecules with the aim to extend the methodology applicability domain. This pilot study, which provided encouraging results, took into account a plethora of different ribonucleic complexes of pharmaceutical interest, the three-dimensional structures of which are known and available on the Protein Data Bank archive (Berman et al., 2000). SuMD methodology proved its ability in describing, with a reduced computational effort, the whole process of ligand–RNA recognition (from the unbound to the bound state), independently by the target topological complexity. As far as we know, this represents the first attempt to overcome methodological limitations within molecular docking when applied to ribonucleic acids, describing binding events through an all-atoms MD-based approach. This study confirms the possible use of SuMD as an innovative computational tool that can accelerate the discovery of new drug candidates and with peculiar attention to SMIRNAs.

MATERIALS AND METHODS

Software Overview

MOE suite (Molecular Operating Environment, version 2018.0101) was used to perform most of the general molecular modeling operations, such as RNA and ligand preparation. All these operations have been performed on an 8 CPU (Intel® Xeon® CPU E5-1620 3.50 GHz) Linux workstation. Molecular dynamics (MD) simulations were performed with an ACEMD engine (Harvey et al., 2009) on a GPU cluster composed of 18 NVIDIA drivers whose models go from GTX 980 to Titan V. For all the simulations, the ff14SB force field with χ modification tuned for RNA (χOL3) was adopted to describe ribonucleic acids, while a general Amber force field (GAFF) was adopted to parameterize small organic molecules (Wang et al., 2006; Sprenger et al., 2015; Tan et al., 2018).

Structures Preparation

The three-dimensional coordinates of each RNA–SMIRNA complex investigated were retrieved from the RCSB PDB database and prepared for SuMD simulations as herein described (Cuzzolin et al., 2016). For structures solved by NMR, which contain multiple conformations of the same complex, the one with the lowest potential energy (usually the first) was selected and then used. All complexes were then processed by means of an MOE protein structure preparation tool: missing atoms in nucleotide bases were built according to AMBER14 force field topology. Missing hydrogen atoms were added to X-Ray-derived complexes, and appropriate ionization states were assigned by Protonate-3D tool (Labute, 2009). Ligand coordinates (both small molecules and peptides) were moved at least 30 Å away from RNA binding cleft, a distance bigger than the electrostatic cut-off term used in the simulation (9 Å with Amber force field) to avoid premature interaction during the initial phases of the SuMD simulations.

Solvated System Setup and Equilibration

Each system investigated by means of SuMD contained an RNA target macromolecule, and the respective ligand, which was a SMIRNA or a peptide, moved far away from the binding site as previously described. The systems were explicitly solvated by a cubic water box with cell borders placed at least 15 Å away from any RNA/ligand atom, using TIP3P as a water model. To neutralize the total charge of each system, Na^+/Cl^- counterions were added to a final salt concentration of 0.154 M. The systems were energy minimized by 500 steps with the conjugate-gradient method, then 500,000 steps (1 ns) of NVT followed by 500,000 steps (1 ns) of NPT simulations were carried out, both using 2 fs as time step and applying harmonic positional constraints on RNA and ligand heavy atoms by a force constant of 1 kcal mol^{-1} $Å^{-2}$, gradually reducing with a scaling factor of 0.1. During this step, the temperature was maintained at 310 K by a Langevin thermostat with low dumping of 1 ps^{-1} and the pressure at 1 atm by a Berendsen barostat (Berendsen et al., 1984; Loncharich et al., 1992). The M-SHAKE algorithm was applied to constrain the bond lengths involving hydrogen atoms. The particle-mesh Ewald (PME) method was exploited to calculate electrostatic interactions with a cubic spline interpolation and 1 Å grid spacing, and a 9.0 Å cutoff was applied for Lennard–Jones interactions (Essmann et al., 1995).

Supervised Molecular Dynamics (SuMD) Simulations

Molecular dynamics simulations represent a well-validated computational tool that, through the numerical solution of the Newton equation of motion, makes it possible to describe the time-dependent evolution of a molecular system. Despite the impressive temporal resolution characterizing the technique, to capture pharmaceutically relevant events, such as the molecular recognition between a drug and its biological target, huge computational efforts are required. The SuMD protocol instead improves the efficiency with which a binding event is sampled, from a microsecond to a nanosecond timescale, by applying a tabu-like algorithm. In detail, short (600 ps long) unbiased MD trajectories are collected, and these monitor, during the entire simulation, the distance between the ligand center of mass with respect to the ribonucleic acid binding site; then, those distance points are fitted into a linear function. Only productive MD steps in which the computed slope is negative are maintained, thus indicating a ligand approach toward the RNA binding site. Otherwise, the simulation is restarted by randomly assigning the atomic velocities from the previous set of coordinates. The supervision algorithm controlled the sampling until the distance between the ligand and the ribonucleic binding site dropped below 5 Å, at which point it was disabled, and a short classical MD simulation was performed, allowing the system to relax. For each case study, up to a maximum of 10 SuMD binding simulations were collected, of which only the best was thoroughly analyzed and discussed in the manuscript. A detailed report on SuMD protocol performance can be found in the **Supplementary Material**. The three-dimensional RNA structures investigated in this study, along with the nucleotides selected for the computation of the respective binding cleft center of mass, are reported in **Figure 1**. In this implementation, the SuMD code is written in

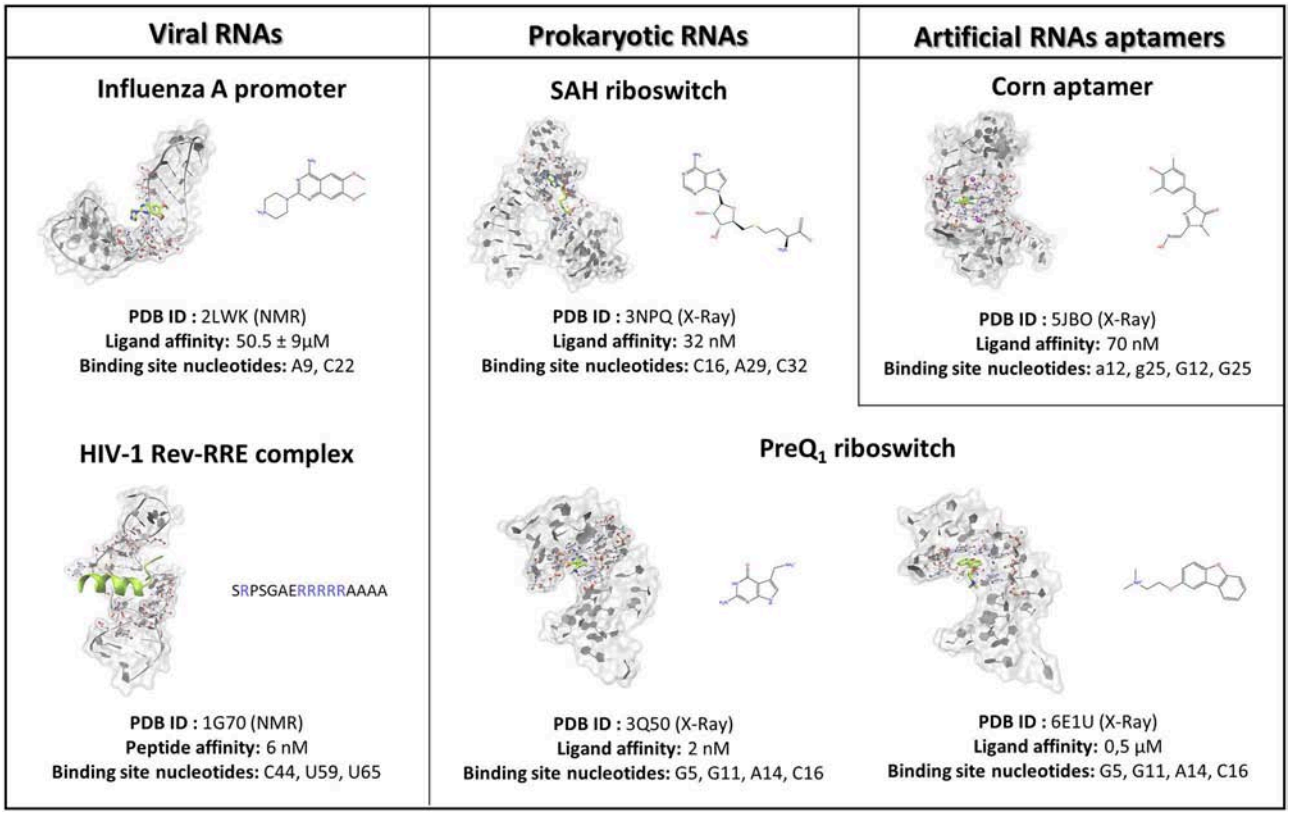


FIGURE 1 | The case studies selected for the SuMD methodological validation are herein summarized and subdivided into RNA of viral origin, prokaryotic origin, or artificially engineered aptamers. For each complex investigated, the three-dimensional structure is depicted, representing with a green color the reference ligand, together with the nucleobases selected to define the binding site position in the SuMD simulations. Finally, the chemical structures of each ligand are reported, along with the experimental datum of binding affinity. In the case of the peptide, the primary sequence is reported, highlighting the basic residues constituting the arginine reach motif (ARM) in a blue color.

Python programming languages and exploits the ProDy python package to perform the geometrical ligand–target supervision process (Bakan et al., 2011).

SuMD Trajectory Analysis

All the SuMD trajectories collected were analyzed by an in-house tool written in tcl and python languages, as described in the original publication (Salmaso et al., 2017). Briefly, the dimension of each trajectory was reduced, saving MD frames at a 20 ps interval; each trajectory was then superposed and aligned on the RNA phosphate atoms of the first frames and wrapped into an image of the system simulated under periodic boundary conditions. The geometric performance of SuMD methodology was evaluated, and it computed the ligand RMSD (Root mean square deviation) along with the entire simulation with respect to the experimental resolved three-dimensional complex. Furthermore, the RMSD of RNA structures were computed on the P atoms of the backbone and plotted over time, and these can be viewed in the **Supplementary Figures S1–S6A**. A ligand–RNA interaction energy estimation during the recognition process was calculated using an MMGBSA protocol, as implemented in AMBER 2014, and it plotted MMGBSA values over time (Miller et al., 2012). The MMGBSA values were also arranged according to the distances between ligand and ribonucleic target mass centers in the Interaction Energy Landscape plots (**Supplementary Figures S1–S6B**). Here, the distances between mass centers are reported on the x-axis, while the MMGBSA values are plotted on the y-axis, and these are rendered by a colorimetric scale going from blue to red for negative to positive energetic values. These graphs allow for the evaluation of the variation of the interaction energy profile at different ligand–RNA distances; this helps to individuate metastable binding states during the binding process. Furthermore, for each target investigated in this work, the nucleotides within a distance of 4 Å from the respective ligand atoms were dynamically selected to qualitatively and quantitatively evaluate the number of contacts during the entire binding process. The most contacted nucleotides were thus selected, to compute a per-nucleotide electrostatic and vdW interaction, and energy contribution, with the ribonucleic target. NAMD was used for post-processing computation of electrostatic interactions using an AMBER ff14SB force field. The cumulative electrostatic interactions were computed for the same target nucleotides by summing the energy values frame by frame along the trajectory, and the resulting graphs were reported to the lower-right of movies provided as **Supplementary Videos 1–6**. Representations

of the molecular structures were prepared with VMD software (Humphrey et al., 1996).

RESULTS AND DISCUSSION

To investigate the SuMD applicability domain and accuracy in the context of ribonucleic acid molecular recognition, a retrospective validation approach was selected, and it stressed the computational methodology ability in geometrically reproducing experimental binding modes of SMIRNA or small folded peptides. The three-dimensional structures of six ligand–RNA complexes solved both through X-Ray and NMR spectroscopy were retrieved from the RCSB PDB database and prepared for subsequent SuMD simulations moving ligands far away from the ribonucleic binding clefts, as accurately described in materials and methods section. The RNA structures, reported in **Figure 1**, were selected to span a vast plethora of pharmaceutically interesting ribonucleic targets, which vary between being of viral and bacterial origin, up to artificially engineered aptamers. Furthermore, the selected structures provide an overview of different peculiar three-dimensional RNA motifs, from a small stem-loop to a riboswitch characterized by a complex architecture. The results collected through SuMD simulations are then reported herein along with the geometric and interactives analysis performed. A summary of all the statistics regarding the simulation performances are reported in the **Supplementary Information**.

Targeting Viral RNAs (vRNAs)

The discovery and design of new antiviral compounds targeting viral proteins are complicated by the enormous variability affecting these macromolecules, an aspect representing the core of the drug resistance phenomenon. On the other hand, lncRNA regions belonging to viral genomes, being less affected by genetic mutations and having no counterpart in human organisms, are becoming attractive pharmaceutical targets. Aminoglycosides, antibacterial drugs known to inhibit protein synthesis acting at the level of the prokaryotic ribosome, have proven to be promiscuous molecules that are also able to bind lncRNA structural elements of viral genomes (Bernacchi et al., 2007). This experimental evidence has paved the way for the discovery of drug-like small molecules able to inhibit the replication for a plethora of pathological viral diseases, such as human immunodeficiency virus (HIV), hepatitis C virus (HCV), severe respiratory syndrome coronavirus (SARS CoV), and influenza A virus (Hermann, 2016).

Influenza a Virus Promoter

Influenza A represents a group of viruses differing from virulence and pathogenicity profiles that all belong to the Orthomyxoviridae family. The Influenza A genome comprises eight negative-sense single-stranded RNA segments (vRNA) encoding for 13 proteins (Coloma et al., 2009). The 5′-end and 3′-end terminal portions of each vRNA segment in the physiological condition fold together in a partial duplex, forming an arrangement called a promoter, which controls RNA-dependent RNA polymerase (RdRp) recognition and, thus, genome transcription and replication (Desselberger et al., 1980). Since the promoter sequences are highly conserved among Influenza A viruses and marginally affected by genetic variation that can enhance the onset of drug resistance, they represent a promising pharmaceutical target. The Varani research group, exploiting an NMR-based fragments screening approach, has identified 6,7-dimethoxy-2-(1-piperazinyl)-4-quinazolinamine (DPQ) as a promising scaffold for antiviral drug development as it is able to bind the Influenza A promoter region with a low micromolar affinity (K_d 50.5 $\pm$ 9 μM) and is also able to inhibit the virus replication in a comparable range of concentration (Lee et al., 2014). The SMIRNA binding mode was experimentally elucidated by means of NMR, as depicted in **Figure 1**, confirming DPQ recognition within the RNA major groove at the (A-A)-U internal loop level.

The SuMD algorithm was then applied to this first case study, in an attempt to investigate the entire DPQ binding mechanism, stressing at the same time the methodology accuracy in reproducing the experimental solved complex. A first interesting aspect is represented by the reduced time window of 30 ns required to sample a putative molecular recognition event between DPQ and its ribonucleic target (**Supplementary Video 1**). This result is quite impressive, especially if compared with classical MD simulations, which otherwise would require extensive computational efforts. At the end of the simulation, as depicted in the **Figure 2** graph, the SMIRNA has converged both from a geometrical and interactive point of view toward the NMR structure binding mode. The low $RMSD_{min}$ value of 2.6 Å, computed on DPQ heavy atoms, confirm, also in the case of nucleic acids, SuMD ability in predicting a reasonable binding hypothesis. This value must not be evaluated with excessive severity, having been calculated only with respect to one of the 16 conformations of the complex deposited on the PDB database. The solution NMR structure has indeed highlighted an important variability in the DPQ positioning within the RNA binding site, with an $RMSD_{max}$, computed on ligand-heavy atoms of 1.4 Å. Moreover, this approach makes it possible to peek at the entire molecular recognition process and to not focus merely on the final state. **Figure 2C** reports a time-dependent analysis performed on the nucleotides most frequently contacted during the simulation, reporting their cumulative contribution to binding, which is defined as the sum of each nucleotide electrostatic and van der Waals (vdW) interaction energy. It is encouraging to note how the nucleotides that computationally have shown a primary role in stabilizing the DPQ complex (A9–A11 and C21–G24) also correspond to those that have experimentally experienced the greatest chemical shift perturbations during NMR experiments. In addition, as reported in **Figures 2B,C** and on **Supplementary Figure S1**, SuMD simulation allows us to decipher the different role played by aforementioned nucleotides, some of them (A9–A11) participating only during the early phases of SMIRNA recognition (until 10 ns) and the other (C21–G24) stabilizing the complex within the ribonucleic cleft (after 10 ns). These results appear even more interesting if we consider the high flexibility characterizing the small RNA duplex. Despite the reduced time window explored by SuMD methodology, the

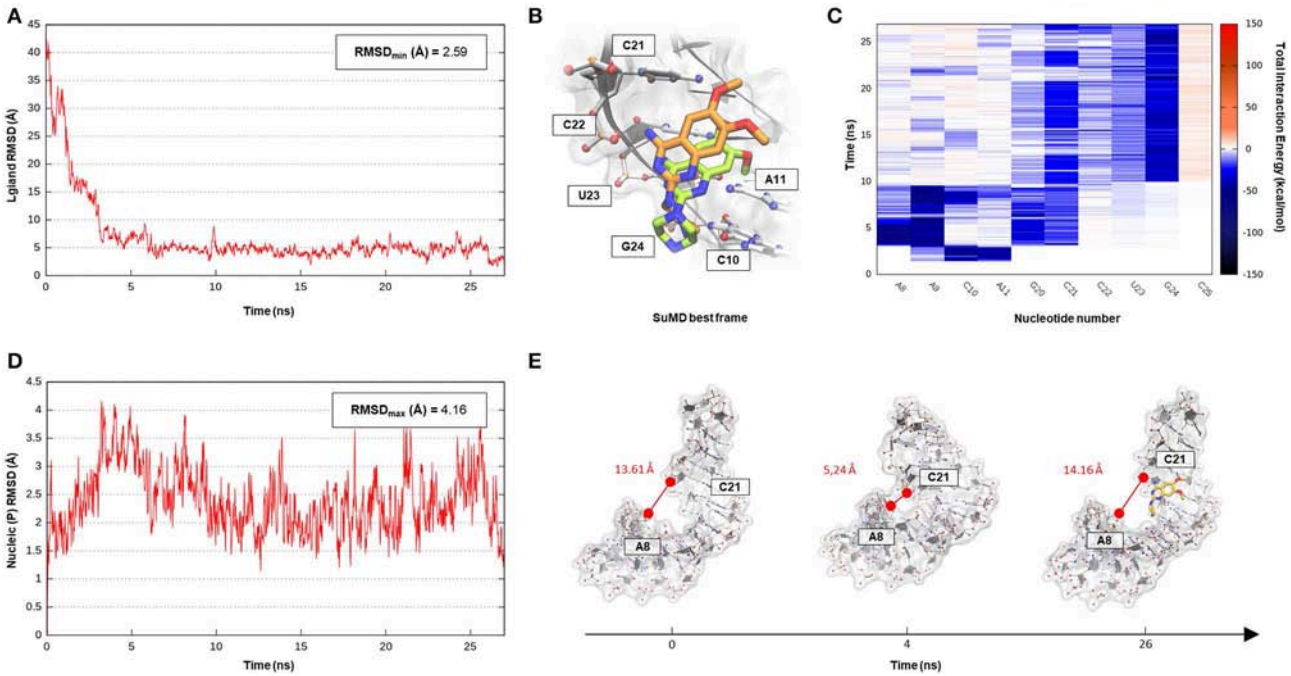


FIGURE 2 | This panel summarizes the recognition pathway of the DPQ fragment with the Influenza A promoter region. **(A)** RMSD of DPQ heavy atoms against the PDB reference. **(B)** Superimposition between the experimental NMR complex (PDB ID 2LWK, green-colored DPQ molecule) and the SuMD conformation with lowest RMSD along the trajectory (orange-colored molecule). The nucleotides surrounding the binding site are reported. **(C)** Dynamic total interaction energy (electrostatic + vdW) computed for most contacted RNA nucleobase. **(D)** RMSD of RNA phosphate atoms belonging to the backbone, computed against the PDB reference. **(E)** Flexibility characterizing the RNA structure during DPQ binding event, binding clef dimension was monitored as the distance dynamically occurring between two key nucleotides (A8 and C21).

structure has indeed shown a relevant $RMSD_{max}$ of 4.2 Å from the initial experimental coordinates (**Figure 2D**). In detail, after a few ns of simulation, the promoter duplex in the ligand-free form folds back on itself, and only DPQ binding allows the structure to return to the experimental linear conformation (**Figure 2E**). The same behavior was coherently captured also by NMR experiments, which previously highlighted how the RNA helical axis curvature changes upon ligand binding, enlarging the dimension of the binding cleft (Lee et al., 2014).

HIV-1 Rev-RRE Complex

The human immunodeficiency virus of type 1 (HIV-1) is a retrovirus belonging to the Lentivirus family, and it is responsible for acquired immunodeficiency syndrome (AIDS). RNA–protein interactions play a fundamental role in controlling the HIV replication cycle and, consequently, virulence profile (Battiste et al., 1996). HIV-1 Rev, in particular, is a small regulatory protein that drives the nuclear export of unspliced and partially spliced viral mRNAs transcripts. Rev protein mediated its function, recognizing a purine-rich bulge within stem-loop IIb of the Rev response element (RRE), a highly structured mRNA region within an *env* intron (DiMattia et al., 2010). The minimal binding domain in the Rev protein is constituted by a short α-helix folded peptide, which contains an arginine-rich binding motif (ARM), a domain known to be important also for tat-TAR (trans-acting region) interactions in HIV. Harada et al., exploiting an *in-vivo* strategy, have identified a class of specific RNA-binding peptides able to target HIV-1 Rev-RRE complex. Specifically, RSG-1.2, an α-helical peptide of 22 amino acids, was selected among a combinatorial library and subsequently engineered, providing a 7-fold increase in binding affinity and a 15-fold increase in selectivity toward the ribonucleic target, further resulting in an *in vivo* ability to completely disrupt the RNA–Rev protein interaction (Harada et al., 1996, 1997). The solution structure of an oligonucleotide portion derived from HIV-1 RRE-IIb stem domain in a complex with an RSG-1.2 peptide was solved through NMR, providing structural details about vRNA targeting by means of the small peptide (Gosser et al., 2001). We have therefore chosen this case study to validate SuMD performance in one of the most complex methodological scenarios, namely the molecular recognition between two highly flexible partners: a small α-helix folded peptide and a portion of ribonucleic acid. In addition, the predominant electrostatic component that both characterizes the RNA polyanionic backbone and the small polycationic peptide, which contain six Arg residues, makes the prediction of the binding mode even more complex. Despite the unfavorable premises, a few tens of ns proved to be sufficient for the SuMD protocol to sample a binding hypothesis for the RSG-1.2 peptide. During the simulation, as observable on **Supplementary Video 1**, the peptide was accommodated with the correct orientation within the HIV-1 RRE-IIb major groove reaching, as reported in **Figure 3A**, an $RMSD_{min}$ value

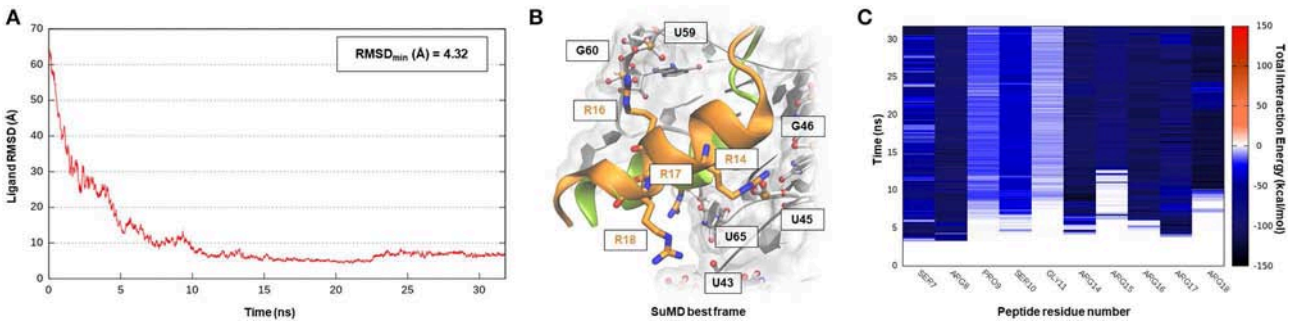


FIGURE 3 | This panel summarizes the recognition pathway of RSG-1.2 peptide with HIV-1 REE. **(A)** RMSD of RSG-1.2 Cα atoms against the PDB reference. **(B)** Superimposition between the experimental NMR complex (PDB ID 1G70, green-colored peptide) and the SuMD conformation with lowest RMSD along the trajectory (orange-colored peptide). The nucleotides surrounding the binding site, along with R residues belonging to ARM, are reported. **(C)** Dynamic total interaction energy (electrostatic + vdW) computed for most contacted RNA nucleobase.

of 4.3 Å, computed on Cα peptide atoms. Although the geometric accuracy is lower than the previous example, the SuMD simulation has allowed us to identify the main interactive hotspots stabilizing the complex. As hypothesized and confirmed by **Figure 3C**, the ARM motif plays a fundamental role in anchoring the RSG-1.2 peptide, with charged residue R 16, R17, and R18 mediating fork electrostatic interactions with the phosphate atoms of the ribonucleic backbone, in a coherent way with the experimentally solved structure. Furthermore, the analysis performed on the trajectory (**Supplementary Figure S2**) has highlighted the peculiar behavior of R14; its guanidinium side chain is deeply buried within the RNA groove, where, differently from the other charged residues, it stabilizes the peptide through a solvent-shielded hydrogen bond and vdW interactions, an aspect in great agreement with the experimental NMR data (Gosser et al., 2001).

Targeting Prokaryotic RNAs

In the last decades, the discovery that many aminoglycoside compounds clinically exploited to treat severe bacterial infections mediated their action by affecting the ribosome machinery confirmed the initial hypothesis of considering RNA, especially prokaryotic ones, as an appetible pharmaceutical target (Disney, 2019). However, the drugs that target ribosomes represent an exception, rather than a model: the abundance of ribosome macromolecules in the cytoplasmic compartment means, therefore, that even modest drug-binding affinity could result in acceptable therapeutic efficacy (Warner et al., 2018). Apart from ribosomes, a putative regulatory role of lncRNAs in bacterial systems has recently become increasingly clear. From a mechanistic point of view, it is possible to distinguish regulatory RNAs acting in *trans*, either by base-pairing with a complementary region in the target mRNA or by sequestration of an RNA-binding protein and regulatory sequences that, in contrast, are encoded as part of the mRNA for the gene they regulate, thus acting in *cis* (Sherwood and Henkin). Riboswitches, which are structured elements typically found in the 5′ untranslated regions (UTR) of mRNAs, represent an interesting example of the latter case (Tucker and Breaker, 2005). These RNA elements, through an aptameric portion, directly sense a physiological signal (ions, cofactors, or metabolites) and transmit the information to the gene expression machinery via a signal-dependent RNA conformational change (Sherwood and Henkin, 2016). The discovery that clinically approved antibacterial Roseflavin exerts part of its therapeutic action by binding the flavin mononucleotide (FMN) riboswitch, together with the increasing availability of structural data on riboswitches, has made these targets very interesting pharmaceutically (Pedrolli et al., 2012).

S-Adenosylhomocysteine Riboswitch

S-adenosyl-(L)-methionine (SAM) is a fundamental cofactor that serves as the primary methyl group donor in a large set of biochemical reactions. In bacteria, SAM homeostasis is so important to the point that at least six classes of RNA riboswitch regulatory elements have since now been characterized (Weinberg et al., 2010). Following SAM-mediated methylation, the by-product S-adenosyl-(L)-homocysteine (SAH) that is released, due to its high toxicity, must be readily degraded by SAH hydrolase (ahcY) enzymes. Recently, a new type of riboswitch was discovered, and it is able to sense and be responsible for the intracellular SAH concentration, upregulating the expression of ahcY enzymes in prokaryotes (Wang et al., 2008). The aptameric portion of the SAH riboswitch recognizes its cognate ligand with a quite high binding affinity of 32 nM and, surprisingly, also provides a discrete selectivity profile toward the original cofactor SAM (1,000-fold lower affinity), ensuring a fine regulation of the SAM/SAH metabolic cycle. The high-resolution crystal structure of the SAH riboswitch aptameric domain in complex with its cognate ligand was recently solved, elucidating the molecular basis for SAH substrate specificity (Edwards et al., 2010). This case study not only represents a pharmaceutical appealing prokaryotic RNA target but also provides the opportunity to stress the SuMD performance in a more complex binding site recognition, if compared to the simple duplex structures until now investigated. The SAH molecule indeed binds a small cleft located in the minor groove of the SAH riboswitch, which adopts an unusual "LL-type" pseudoknot conformation. Also, in this case, around 20 ns were sufficient for the SuMD

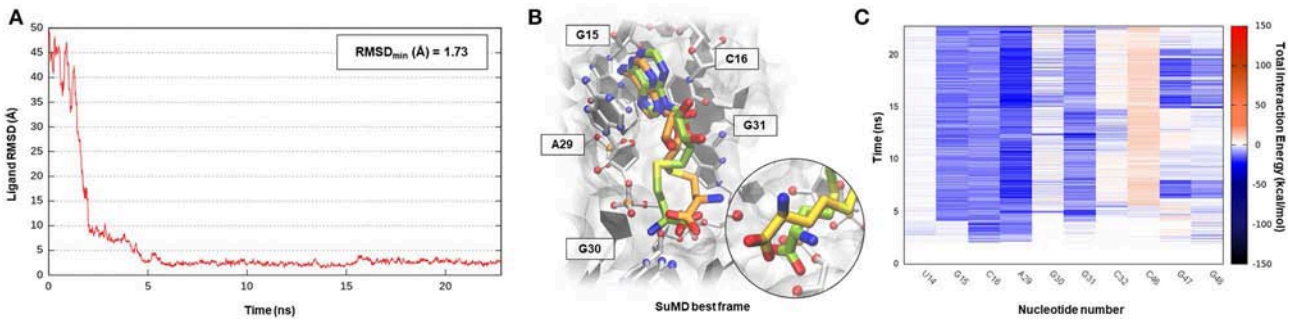


FIGURE 4 | This panel summarizes the recognition pathway of the SAH molecule with SAH riboswitch. **(A)** RMSD of SAH heavy atoms against the PDB reference. **(B)** Superimposition between the experimental X-Ray complex (PDB ID 3NPQ, green-colored SAH molecule) and the SuMD conformation with lowest RMSD along the trajectory (orange-colored molecule). The nucleotides surrounding the binding site are reported. Within the circular window, the SuMD conformation sampled for SAH tail is compared to a different crystallographic reference (PDB ID 3NPN). **(C)** Dynamic total interaction energy (electrostatic + vdW) computed for most contacted RNA nucleobase.

protocol to sample a putative molecular recognition trajectory (**Supplementary Video 3**). In detail, as reported in **Figure 4**, after only a few nanoseconds, SAH reached the riboswitch binding cleft reproducing the crystallographic complex with a notable geometric accuracy ($RMSD_{min}$ 1.7 Å). Then, the ligand conformation remained stable until the end of the simulation. From an interactive point of view, as reported in **Figure 4C** and also in **Supplementary Figure S3**, the SuMD trajectory analysis correctly highlighted the stabilizing role played by nucleotide C16 and A29, among which the adenine core of SAH is intercalated, providing the greatest vdW interactions. In contrast, the electrostatic contribution to binding analysis has revealed a divergent situation. Indeed, nucleobase G15, mediating a hydrogen bond network with an SAH adenine scaffold, is responsible for a great stabilizing contribution, whereas nucleotide C46 has shown during the entire simulation an unexpected repulsive contribution. The reason for this can be found in the conformation sampled by SuMD for the SAH homocysteine terminal tail. As depicted by **Figure 4B**, the carboxylic moiety of the ligand spatially approaches the C46 pyrimidine carbonyl, whereas in the crystallographic structure (green representation), through a simple bond rotation, the interaction is instead mediated by the vicinal amino group. Curiously, the same research group also deposited on the PDB database a worst resolution structure of the complex under investigation (PDB ID 3NPN), reporting the same apparently energetic unfavored SAH conformation described by the SuMD protocol (**Figure 4B**, circular window), thus validating the goodness of the sampling and the flexibility characterizing the ligand tail.

Pre-queuosine$_1$ Riboswitch

Pre-queosine$_1$ ($PreQ_1$), or 7-aminomethyl-7-deazaguanine, is a metabolic intermediate in the synthetic pathway that, starting from guanosine-5′-triphosphate (GTP) nucleotide, originates the hypermodified guanine derivate queuosine (Q). Q has been detected both in eubacteria and eukaryotic organisms where it occupies the anticodon wobble position of tRNAs specific for the amino acid asparagine, aspartate, histidine, and tyrosine (Roth et al., 2007). Q modification has been related to an improvement in translation fidelity as well as bacterial pathogenicity. Interestingly, only prokaryotes can synthesize Q via a multistep reaction, whereas eukaryotes are obliged to assimilate the nucleoside through the diet (Eichhorn et al., 2014). In bacteria like *Bacillus subtilis (Bs)* or *Thermoanaerobacter tengcongenesis (Tt)*, the expression of genes responsible for Q biosynthesis is negatively modulated by the intermediate $PreQ_1$ intracellular concentration. $PreQ_1$, binding to a small aptameric RNA motif composed of 34 nucleotides determines the folding of the $PreQ_1$ riboswitch in an "H-type" pseudoknot structure in which more than half of the nucleobases engage in triplet or quartet interactions (Rieder et al., 2010; Jenkins et al., 2011). The three-dimensional structure of the class I $PreQ_1$ riboswitch in complex with its cognate ligand was solved by X-ray crystallography (PDB ID 3Q50), and this allowed us to speculate about the quite impressive binding affinity characterizing this endogenous precursor ($K_d = 2$ nM) (Edwards et al., 2010). Even in this case, <40 ns of SuMD simulation proved to be sufficient in describing a binding event between the metabolic intermediate $PreQ_1$ and its related riboswitch (**Supplementary Video 4**). As observable in **Figure 5A**, $PreQ_1$ recognition mainly articulates in three well-distinguishable phases. In the beginning, the ligand approaches the riboswitch binding site vestibule where it negotiates for about 15 ns the accommodation in the deep cleft before converging, with great geometric accuracy ($RMSD_{min}$ 1.3 Å), toward the solved crystallographic conformation. This behavior has also been captured by the interaction energy graph (**Supplementary Figure S4B**), highlighting the presence of two major sites visited during the recognition trajectory, i.e., the canonical binding cleft and the aforementioned external vestibular region, located about 10 Å apart. It is interesting to note the comparable interaction energy characterizing these two distal sites, which are distinguishable for their different degrees of solvent exposition. In addition, the dynamic interaction fingerprint reported in **Figure 5C**, elucidates the role played by the binding site nucleotides during recognition in a coherent way with respect to the results reported on the original publication.

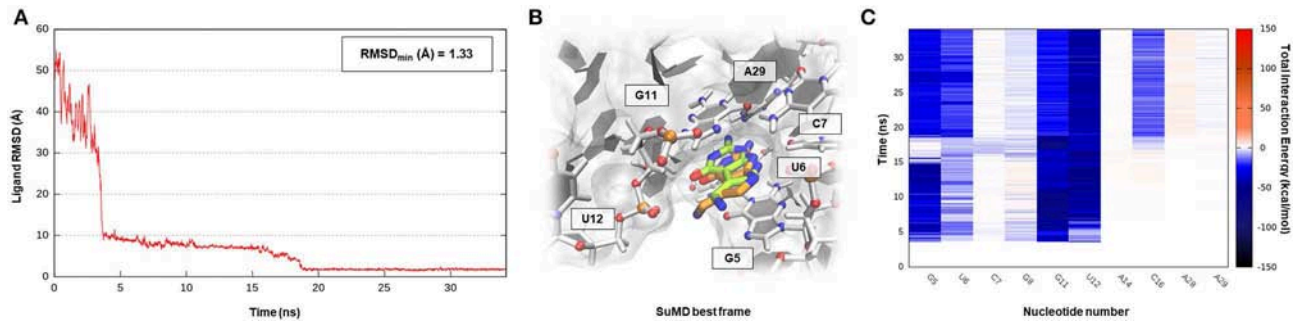


FIGURE 5 | This panel summarizes the recognition pathway of the $PreQ_1$ molecule with $PreQ_1$-1 riboswitch. **(A)** RMSD of $PreQ_1$ heavy atoms against the PDB reference. **(B)** Superimposition between the experimental X-Ray complex (PDB ID 3Q50, green-colored $PreQ_1$ molecule) and the SuMD conformation with lowest RMSD along the trajectory (orange-colored molecule). The nucleotides surrounding the binding site are reported. **(C)** Dynamic total interaction energy (electrostatic + vdW) computed for most contacted RNA nucleobase.

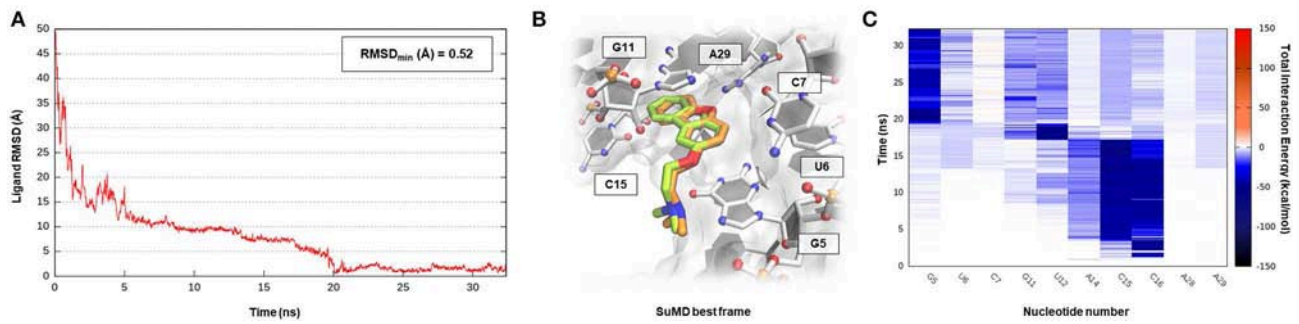


FIGURE 6 | This panel summarizes the recognition pathway of HMJ molecule with $PreQ_1$-1 riboswitch. **(A)** RMSD of HMJ heavy atoms against the PDB reference. **(B)** Superimposition between the experimental X-Ray complex (PDB ID 6E1U, green-colored HMJ molecule) and the SuMD conformation with lowest RMSD along the trajectory (orange-colored molecule). The nucleotides surrounding the binding site are reported. **(C)** Dynamic total interaction energy (electrostatic + vdW) computed for most contacted RNA nucleobase.

All the cases considered so far have confirmed the ability of SuMD to predict reasonable binding hypotheses for different ligands when exploiting as starting point the experimental structures of the ribonucleic targets in which each of these ligands were originally co-crystallized. From a pharmaceutical and applicative perspective, however, it is often required to rationalize the binding mode of compounds that are in most of the cases different from the ones now co-crystallized. It has thus become crucial to understand how the choice of the initial RNA target conformation could affect SuMD performance. The studies performed by the Schneekloth Jr. group in the attempt to experimentally asses the druggability profile of $PreQ_1$-I riboswitch through synthetic organic molecules have then given us an opportunity to further explore this question. In a recent scientific work, it the discovery of HMJ was indeed reported; this is a dibenzofuran derivative that, despite the not obvious chemical similarity with $PreQ_1$, exhibits a sub-micromolar affinity to the RNA target (K_d = 0.5 μM) and the ability to induce premature transcriptional termination (Connelly et al., 2019). The three-dimensional structure determination of the complex was, however, quite difficult and was achieved only by designing a hybrid riboswitch aptamer sequence in which the nucleobase A14, as well as the two vicinal ones, were removed (PDB ID 6E1U). Since this structure lacked a key binding site nucleotides, it represent a non-optimal starting point for a computational study; we therefore decided to investigate the HMJ binding mechanism, exploiting the high-quality riboswitch structure originally solved in the presence of $PreQ_1$ and then comparing the accuracy of the prediction with the experimental solved data. Encouragingly, even for such a system, the SuMD protocol has succeeded in sampling, in about 30 ns, an extremely accurate binding hypothesis for HMJ, whose $RMSD_{min}$ was computed with respect to reference structure (PDB ID 3Q50) and has reached the impressive value of 0.5 Å (**Figure 6A**, **Supplementary Video 5**). From the analysis of the trajectory, it was furthermore possible to confirm how the benzofuran ligand competes with $PreQ_1$ for the riboswitch binding site. As depicted by **Figure 6C**, and as is coherent with experimental evidence, HMJ makes a strong stabilizing interaction with the nucleobases G5, G11, and C16, which define the "floor" and the "ceiling" of the binding cleft where the aromatic core stacks, and nucleobase U6, C15, and A29, which shape instead the binding cavity borders. Moreover, the Interaction Energy Landscape (**Supplementary Figure S5B**) highlights a binding profile similar to the one previously described for the cognate ligand $PreQ_1$, confirming the vestibular region's role in recruiting the riboswitch binding partners.

Targeting Artificial RNA Aptamers Containing G-Quadruplex Motifs

The discovery, made in 1994, that the green fluorescent protein (GFP) from the jellyfish *Aequorea victoria* could be used as a marker for protein localization and expression has revolutionized molecular biology to the point that, in 2008, the discovery earned a Nobel prize (Swaminathan, 2009). However, since a minimal portion of the human genome is translated into proteins while most of it is transcribed into RNA, being able to investigate the dynamic and spatial properties of the human transcriptome has become essential. As there are no known naturally fluorescent RNAs, a series of *in vitro* engineered ribonucleic tags able to fold into peculiar three-dimensional structures were selected (Trachman and Ferré-D'Amaré, 2019). These RNAs, through an aptameric domain, can bind fluorophore molecules, increasing their spectroscopic signal and hence allowing for the dynamic monitoring of nucleic acid expression and localization in cells. Most of the fluorophore RNA binding sites, despite the different overall architecture, have evolutionarily converged on G-quadruplex motifs, supporting their important role in enhancing the fluorescence phenomenon, in a similar way to how the β-barrel domains characterize GFPs (Warner et al., 2014).

Corn Aptamer

Corn is one recently developed RNA aptamer engineered *in vitro* to bind 3,5-difluoro-4-hydroxybenzylidene imidazolinone-2-oxime (DFHO), a fluorophore analogous of red fluorescent protein (RFP) (Warner et al., 2017). Corn-DFHO differs from other similar RNA tags for its limited light-induced cytotoxicity, its minimal background fluorescence, and its increased photostability, thus representing a valuable imaging tool. Corn aptamer is characterized by an atypical three-dimensional structure elucidated by X-ray crystallography and biophysical experiments. How it is observable in **Figure 1** that two RNA segments join together in a quasi-symmetric homodimer structure (1:2 chromophore:RNA stoichiometry) at the interfaces where a single DFHO molecule is tightly bound (K_d = 70 nM), stacking between two G-quadruplex planes stabilized by the presence of K^+ ions (Warner et al., 2017). Despite the lack of therapeutic application for this aptamer, which is instead more suitable for molecular biology studies, the investigation of such a complex binding site recognition can be considered as a proof of concept to validate G-quadruplex motif targeting through an SuMD approach. Nucleotide quartet structures, which presence have been extensively characterized in the telomeric terminal portion of eukaryotes chromosomes and within gene promoter regions, are indeed acquiring increasing attention, as they could represent promising pharmaceutical targets (Balasubramanian and Neidle, 2009). As shown in **Supplementary Video 6**, SuMD methodology has produced a putative binding trajectory for DFHO in <30 ns, converging with an impressive geometrical accuracy toward the experimental solved complex ($RMSD_{min}$ 0.34 Å) (**Figure 7A**). Moreover, the Dynamic Total Interaction Energy plot reported on **Figure 7C**, strongly retraces the interactive pattern already described on the original scientific work, highlighting the role played by nucleotide G12, G25 (first protomer), and g25 (second protomer) in circumscribing the sandwich cavity within which the aromatic chromophore stacks. Nucleobase A14 (first protomer) and a11 (second protomer) instead mediated a hydrogen bond network with oxime and imine moieties of the DFHO ligand, respectively. SuMD simulation has also illuminated how the entire binding process is not driven by the electrostatic contribution, as often it happens for SMIRNA, but is instead controlled by the vdW interactions (**Supplementary Figure S6**). From this perspective, Corn aptamer represents an unusual, but potentially revolutionary case study, as it distorts an old paradigm that has now since affected the identification of putative RNA binders. DFHO has indeed demonstrated how even apolar or anionic molecules can target ribonucleic acids reaching a nanomolar binding affinity. This provides the opportunity to expand the chemical space explorable by SMIRNA beside that of the well-known, but often problematic, polycationic compounds.

CONCLUSION

Over the last decades, among all the biological macromolecules, proteins have represented the target of choice for the

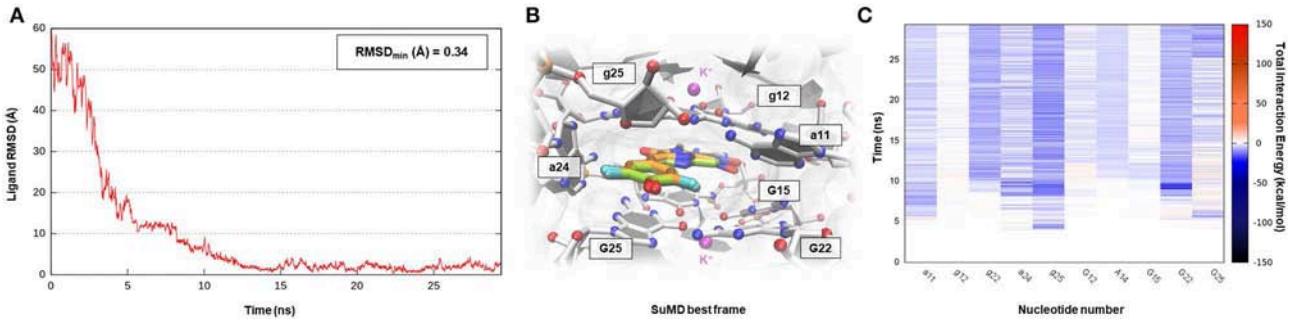


FIGURE 7 | This panel summarizes the recognition pathway of DFHO molecules with the Corn aptamer. **(A)** RMSD of DFHO heavy atoms against the PDB reference. **(B)** Superimposition between the experimental X-Ray complex (PDB ID 5BJO, green-colored DFHO molecule) and the SuMD conformation with lowest RMSD along the trajectory (orange-colored molecule). The nucleotides surrounding the binding site are reported. **(C)** Dynamic total interaction energy (electrostatic + vdW) computed for most contacted RNA nucleobase.

development of new drug candidates. Nucleic acids, on the other hand, have so far represented a less attractive target due to the difficulty in guaranteeing a selective recognition mechanism. The recent discovery of peculiar and physiologically stable three-dimensional conformation characterizing RNAs oligomers has, however, paved the way for the investigation of SMIRNA. The increasing availability of structural data for a wide range of relevant therapeutic ribonucleic targets has promoted the application of well-validated SBDD computational approaches, such as molecular docking, also in this field. However, the remarkable flexibility and the peculiar electrostatic potential, which distinguish nucleic acids from proteins, have readily highlighted the limitation of many of these methodologies. MD simulations would allow us to overcome some of the aforementioned problems; however, the computational cost required to capture rare events such as ligand binding has so far limited their routine utilization.

In this work, we have investigated the applicability domain of SuMD in the field of pharmaceutically relevant RNA polymers. The performances of the protocol were measured as the geometrical accuracy, expressed in terms of RMSD, with which an experimentally solved complex is predicted by the SuMD simulation. Case studies in this research were chosen in such a way as to span very different ribonucleic secondary, tertiary, and even quaternary structures, starting from small duplex stem-loops up to pseudoknot or aptameric homodimers, which contain G-quadruplex motifs. Furthermore, the recognition of different ligands was investigated, both small organic molecules and folded α-helical peptides.

Although this work must be considered as a preliminary investigation and the number of examples taken into consideration cannot guarantee statistical robustness, it is encouraging to note how, in all the six ribonucleic complexes simulated, SuMD correctly reproduced the experimentally solved final state starting from the unbound state in few hours of simulation. The accuracy of the protocol varies significantly in a system-dependent manner, but, in all the cases, it was possible to collect valuable interactive and energetic information about the nucleotides dynamically involved in the recognition process. Curiously, the RNA target in which the architecture of the binding site is not very complex, such as the stem-loop domain of Influenza A promoter and HIV-1 RRE, are those in which the computational protocol experienced the poorest geometric accuracy in reproducing the ligand-binding mode. A separate consideration must be made for the latter complex (PDB ID 1G70) since the recognition between two extremely flexible entities, i.e., the small peptide and the RNA duplex, represents a very challenging case. However, the results obtained, with an $RMSD_{min}$ lower than 5 Å, are in line with those previously described when applying SuMD methodology to peptide–protein recognition. Moving toward more complex binding sites, such as the one that characterizes pseudoknot riboswitch structures or G-quadruple-shaped clefts, the geometric accuracy of the method progressively improves, with the best results obtained in the artificial aptameric structure ($RMSD_{min}$ 0.34 Å). These findings are in agreement with a recent perspective work that assessed how the complexity of an RNA binding site, measured in terms of information content, could represent a valuable discriminant to individuate druggable oligonucleotides (Warner et al., 2018). Indeed, the three-dimensional complexity of a binding site makes ribonucleic pocket more similar to a protein-like environment rather than an ordered and repetitive structure like that characterizing DNA.

Furthermore, the high conformational flexibility that has characterized all the investigated ribonucleic structures (RMSD computed on RNA backbone are reported on **Supplementary Material**) during SuMD simulations has evidenced the importance of adopting techniques able to consider the flexibility of both macromolecules and ligands to better describe such complex molecular recognition. In conclusion, we have shown how SuMD can be a valid computational method to generate binding hypothesis for ribonucleic targets in a nanosecond timescale, explicitly considering both the role of the solvent and the flexibility of the macromolecule. SuMD simulation results could not only help with the interpretation and investigation of the complex mechanism of recognition characterizing SMIRNA, especially when structural information is not available, but they could also guide the rational discovery and optimization of these compounds.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

MB carried out the experiment. MB wrote the manuscript with support from MS. MS and SM supervised the project. MB and SM conceived the original idea.

ACKNOWLEDGMENTS

We are very grateful to Chemical Computing Group, OpenEye, and Acellera for the scientific and technical partnership, and we thank NVIDIA for the donation of the Titan Xp and Titan V GPUs used to perform the computational studies.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2020.00107/full#supplementary-material

REFERENCES

Bakan, A., Meireles, L. M., and Bahar, I. (2011). ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics* 27, 1575–1577. doi: 10.1093/bioinformatics/btr168

Balasubramanian, S., and Neidle, S. (2009). G-quadruplex nucleic acids as therapeutic targets. *Curr. Opin. Chem. Biol.* 13, 345–353. doi: 10.1016/j.cbpa.2009.04.637

Battiste, J. L., Mao, H., Rao, N. S., Tan, R., Muhandiram, D. R., Kay, L. E., et al. (1996). α helix-RNA major groove recognition in an HIV-1 rev peptide-RRE RNA Complex. *Science* 273, 1547–1551. doi: 10.1126/science.273.5281.1547

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81, 3684–3690. doi: 10.1063/1.448118

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Bernacchi, S., Freisz, S., Maechling, C., Spiess, B., Marquet, R., Dumas, P., et al. (2007). Aminoglycoside binding to the HIV-1 RNA dimerization initiation site: thermodynamics and effect on the kissing-loop to duplex conversion. *Nucleic Acids Res.* 35, 7128–7139. doi: 10.1093/nar/gkm856

Bissaro, M., Bolcato, G., Deganutti, G., Sturlese, M., and Moro, S. (2019). Revisiting the allosteric regulation of sodium cation on the binding of adenosine at the human A2A adenosine receptor: insights from supervised molecular dynamics (SuMD) simulations. *Molecules* 24:2752. doi: 10.3390/molecules24152752

Coloma, R., Valpuesta, J. M., Arranz, R., Carrascosa, J. L., Ortín, J., and Martín-Benito, J. (2009). The structure of a biologically active influenza virus ribonucleoprotein complex. *PLoS Pathog.* 5:e1000491. doi: 10.1371/journal.ppat.1000491

Connelly, C. M., Moon, M. H., and Schneekloth, J. S. (2016). The emerging role of RNA as a therapeutic target for small molecules. *Cell Chem. Biol.* 23, 1077–1090. doi: 10.1016/j.chembiol.2016.05.021

Connelly, C. M., Numata, T., Boer, R. E., Moon, M. H., Sinniah, R. S., Barchi, J. J., et al. (2019). Synthetic ligands for $PreQ_1$ riboswitches provide structural and mechanistic insights into targeting RNA tertiary structure. *Nat. Commun.* 10:1501. doi: 10.1038/s41467-019-09493-3

Cruz, J. A., and Westhof, E. (2009). The dynamic landscapes of RNA architecture. *Cell* 136, 604–609. doi: 10.1016/j.cell.2009.02.003

Cuzzolin, A., Sturlese, M., Deganutti, G., Salmaso, V., Sabbadin, D., Ciancetta, A., et al. (2016). Deciphering the complexity of ligand–protein recognition pathways using supervised molecular dynamics (SuMD) simulations. *J. Chem. Inf. Model.* 56, 687–705. doi: 10.1021/acs.jcim.5b00702

Denning, E. J., Priyakumar, U. D., Nilsson, L., and Mackerell, A. D. (2011). Impact of 2′-hydroxyl sampling on the conformational properties of RNA: update of the CHARMM all-atom additive force field for RNA. *J. Comput. Chem.* 32, 1929–1943. doi: 10.1002/jcc.21777

Desselberger, U., Racaniello, V. R., Zazra, J. J., and Palese, P. (1980). The 3′ and 5′-terminal sequences of influenza A, B and C virus RNA segments are highly conserved and show partial inverted complementarity. *Gene* 8, 315–328. doi: 10.1016/0378-1119(80)90007-4

DiMattia, M. A., Watts, N. R., Stahl, S. J., Rader, C., Wingfield, P. T., Stuart, D. I., et al. (2010). Implications of the HIV-1 Rev dimer structure at 3.2 Å resolution for multimeric binding to the Rev response element. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5810–5814. doi: 10.1073/pnas.0914946107

Disney, M. D. (2019). Targeting RNA with small molecules to capture opportunities at the intersection of chemistry, biology, and medicine. *J. Am. Chem. Soc.* 141, 6776–6790. doi: 10.1021/jacs.8b13419

Disney, M. D., Winkelsas, A. M., Velagapudi, S. P., Southern, M., Fallahi, M., and Childs-Disney, J. L. (2016). Inforna 2.0: a platform for the sequence-based design of small molecules targeting structured RNAs. *ACS Chem. Biol.* 11, 1720–1728. doi: 10.1021/acschembio.6b00001

Disney, M. D., Yildirim, I., and Childs-Disney, J. L. (2014). Methods to enable the design of bioactive small molecules targeting RNA. *Org. Biomol. Chem.* 12, 1029–1039. doi: 10.1039/C3OB42023J

Donlic, A., and Hargrove, A. E. (2018). Targeting RNA in mammalian systems with small molecules. *Wiley Interdiscip. Rev. RNA* 9:e1477. doi: 10.1002/wrna.1477

Draper, D. E. (1995). Protein-RNA recognition. *Annu. Rev. Biochem.* 64, 593–620. doi: 10.1146/annurev.bi.64.070195.003113

Ecker, D. J., and Griffey, R. H. (1999). RNA as a small-molecule drug target: doubling the value of genomics. *Drug Discov. Today* 4, 420–429. doi: 10.1016/S1359-6446(99)01389-6

Edwards, A. L., Reyes, F. E., Héroux, A., and Batey, R. T. (2010). Structural basis for recognition of S-adenosylhomocysteine by riboswitches. *RNA* 16, 2144–2155. doi: 10.1261/rna.2341610

Eichhorn, C. D., Kang, M., and Feigon, J. (2014). Structure and function of $preQ_1$ riboswitches. *Biochim. Biophys. Acta - Gene Regul. Mech.* 1839, 939–950. doi: 10.1016/j.bbagrm.2014.04.019

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *J. Chem. Phys.* 103, 8577–8593. doi: 10.1063/1.470117

Flynn, R. A., Smith, B. A. H., Johnson, A. G., Pedram, K., George, B. M., Malaker, S. A., et al. (2019). Mammalian Y RNAs are modified at discrete guanosine residues with N-glycans. *bioRxiv[Preprint]* 787614. doi: 10.1101/787614

Fulle, S., and Gohlke, H. (2009). Molecular recognition of RNA: challenges for modelling interactions and plasticity. *J. Mol. Recognit.* 23, 220–231. doi: 10.1002/jmr.1000

Gosser, Y., Hermann, T., Majumdar, A., Hu, W., Frederick, R., Jiang, F., et al. (2001). Peptide-triggered conformational switch in HIV-1 RRE RNA complexes. *Nat. Struct. Biol.* 8, 146–150. doi: 10.1038/84138

Harada, K., Martin, S. S., and Frankel, A. D. (1996). Selection of RNA-binding peptides *in vivo*. *Nature* 380, 175–179. doi: 10.1038/380175a0

Harada, K., Martin, S. S., Tan, R., and Frankel, A. D. (1997). Molding a peptide into an RNA site by *in vivo* peptide evolution. *Proc. Natl. Acad. Sci. U.S.A.* 94, 11887–11892. doi: 10.1073/pnas.94.22.11887

Harvey, M. J., Giupponi, G., and Fabritiis, G., De (2009). ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.* 5, 1632–1639. doi: 10.1021/ct9000685

Hermann, T. (2002). Rational ligand design for RNA: the role of static structure and conformational flexibility in target recognition. *Biochimie* 84, 869–875. doi: 10.1016/S0300-9084(02)01460-8

Hermann, T. (2016). Small molecules targeting viral RNA. *Wiley Interdiscip. Rev. RNA* 7, 726–743. doi: 10.1002/wrna.1373

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual molecular dynamics. *J. Mol. Graph.* 14, 33–38. doi: 10.1016/0263-7855(96)00018-5

Jenkins, J. L., Krucinska, J., McCarty, R. M., Bandarian, V., and Wedekind, J. E. (2011). Comparison of a $PreQ_1$ riboswitch aptamer in metabolite-bound and free states with implications for gene regulation. *J. Biol. Chem.* 286, 24626–24637. doi: 10.1074/jbc.M111.230375

Labute, P. (2009). Protonate3D: assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins* 75, 187–205. doi: 10.1002/prot.22234

Lee, M.-K., Bottini, A., Kim, M., Bardaro, M. F., Zhang, Z., Pellecchia, M., et al. (2014). A novel small-molecule binds to the influenza A virus RNA promoter and inhibits viral replication. *Chem. Commun.* 50, 368–370. doi: 10.1039/C3CC46973E

Loncharich, R. J., Brooks, B. R., and Pastor, R. W. (1992). Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N′-methylamide. *Biopolymers* 32, 523–535. doi: 10.1002/bip.360320508

Miller, B. R., McGee, T. D., Swails, J. M., Homeyer, N., Gohlke, H., and Roitberg, A. E. (2012). MMPBSA.py: an efficient program for end-state free energy calculations. *J. Chem. Theory Comput.* 8, 3314–3321. doi: 10.1021/ct300418h

Morris, K. V., and Mattick, J. S. (2014). The rise of regulatory RNA. *Nat. Rev. Genet.* 15, 423–437. doi: 10.1038/nrg3722

Pedrolli, D. B., Matern, A., Wang, J., Ester, M., Siedler, K., Breaker, R., et al. (2012). A highly specialized flavin mononucleotide riboswitch responds differently to similar ligands and confers roseoflavin resistance to *Streptomyces davawensis*. *Nucleic Acids Res.* 40, 8662–8673. doi: 10.1093/nar/gks616

Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, T. E., Laughton, C. A., et al. (2007). Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* 92, 3817–3829. doi: 10.1529/biophysj.106.097782

Rieder, U., Kreutz, C., and Micura, R. (2010). Folding of a transcriptionally acting $PreQ_1$ riboswitch. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10804–10809. doi: 10.1073/pnas.0914925107

Roth, A., Winkler, W. C., Regulski, E. E., Lee, B. W. K., Lim, J., Jona, I., et al. (2007). A riboswitch selective for the queuosine precursor $preQ_1$ contains

an unusually small aptamer domain. *Nat. Struct. Mol. Biol.* 14, 308–317. doi: 10.1038/nsmb1224

Sabbadin, D., and Moro, S. (2014). Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR–ligand recognition pathway in a nanosecond time scale. *J. Chem. Inf. Model.* 54, 372–376. doi: 10.1021/ci400766b

Salmaso, V., Sturlese, M., Cuzzolin, A., and Moro, S. (2017). Exploring protein-peptide recognition pathways using a supervised molecular dynamics approach. *Structure 25,* 655–662.e2. doi: 10.1016/j.str.2017.02.009

Salmon, L., Yang, S., and Al-Hashimi, H. M. (2014). Advances in the determination of nucleic acid conformational ensembles. *Annu. Rev. Phys. Chem.* 65, 293–316. doi: 10.1146/annurev-physchem-040412-110059

Sherwood, A. V., and Henkin, T. M. (2016). Riboswitch mediated gene regulation: novel RNA architectures dictate gene expression responses. *Annu. Rev. Microbiol.* 70, 361–374. doi: 10.1146/annurev-micro-091014-104306

Sprenger, K. G., Jaeger, V. W., and Pfaendtner, J. (2015). The general AMBER force field (GAFF) can accurately predict thermodynamic and transport properties of many ionic liquids. *J. Phys. Chem. B* 119, 5882–5895. doi: 10.1021/acs.jpcb.5b00689

Sucheck, S. J., and Wong, C.-H. (2000). RNA as a target for small molecules. *Curr. Opin. Chem. Biol.* 4, 678–686. doi: 10.1016/S1367-5931(00)00142-3

Swaminathan, S. (2009). GFP: the green revolution. *Nat. Cell Biol.* 11:S20. doi: 10.1038/ncb1953

Tan, D., Piana, S., Dirks, R. M., and Shaw, D. E. (2018). RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc. Natl. Acad. Sci. U.S.A.* 115, E1346–E1355. doi: 10.1073/pnas.1713027115

Trachman, R. J., and Ferré-D'Amaré, A. R. (2019). Tracking RNA with light: selection, structure, and design of fluorescence turn-on RNA aptamers. *Q. Rev. Biophys.* 52:e8. doi: 10.1017/S0033583519000064

Tucker, B. J., and Breaker, R. R. (2005). Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.* 15, 342–348. doi: 10.1016/j.sbi.2005.05.003

Velagapudi, S. P., Gallo, S. M., and Disney, M. D. (2014). Sequence-based design of bioactive small molecules that target precursor microRNAs. *Nat. Chem. Biol.* 10, 291–297. doi: 10.1038/nchembio.1452

Wang, J., Wang, W., Kollman, P. A., and Case, D. A. (2006). Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* 25, 247–260. doi: 10.1016/j.jmgm.2005.12.005

Wang, J. X., Lee, E. R., Morales, D. R., Lim, J., and Breaker, R. R. (2008). Riboswitches that sense S-adenosylhomocysteine and activate genes involved in coenzyme recycling. *Mol. Cell* 29, 691–702. doi: 10.1016/j.molcel.2008.01.012

Warner, K. D., Chen, M. C., Song, W., Strack, R. L., Thorn, A., Jaffrey, S. R., et al. (2014). Structural basis for activity of highly efficient RNA mimics of green fluorescent protein. *Nat. Struct. Mol. Biol.* 21, 658–663. doi: 10.1038/nsmb.2865

Warner, K. D., Hajdin, C. E., and Weeks, K. M. (2018). Principles for targeting RNA with drug-like small molecules. *Nat. Rev. Drug Discov.* 17, 547–558. doi: 10.1038/nrd.2018.93

Warner, K. D., Sjekloa, L., Song, W., Filonov, G. S., Jaffrey, S. R., and Ferré-D'Amaré, A. R. (2017). A homodimer interface without base pairs in an RNA mimic of red fluorescent protein. *Nat. Chem. Biol.* 13, 1195–1201. doi: 10.1038/nchembio.2475

Weinberg, Z., Wang, J. X., Bogue, J., Yang, J., Corbino, K., Moy, R. H., et al. (2010). Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol.* 11:R31. doi: 10.1186/gb-2010-11-3-r31

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Bissaro, Sturlese and Moro. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 28 February 2020
doi: 10.3389/fchem.2020.00142

LigBuilder V3: A Multi-Target *de novo* Drug Design Approach

Yaxia Yuan[1], *Jianfeng Pei*[2]* *and Luhua Lai*[1,2,3]*

[1] *Beijing National Laboratory for Molecular Sciences, State Key Laboratory for Structural Chemistry of Unstable and Stable Species, College of Chemistry and Molecular Engineering, Peking University, Beijing, China,* [2] *Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China,* [3] *Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China*

With the rapid development of systems-based pharmacology and poly-pharmacology, method development for rational design of multi-target drugs has becoming urgent. In this paper, we present the first *de novo* multi-target drug design program LigBuilder V3, which can be used to design ligands to target multiple receptors, multiple binding sites of one receptor, or various conformations of one receptor. LigBuilder V3 is generally applicable in *de novo* multi-target drug design and optimization, especially for the design of concise ligands for protein targets with large difference in binding sites. To demonstrate the utility of LigBuilder V3, we have used it to design dual-functional inhibitors targeting HIV protease and HIV reverse transcriptase with three different strategy, including multi-target *de novo* design, multi-target growing, and multi-target linking. The designed compounds were computational validated by MM/GBSA binding free energy estimation as highly potential multi-target inhibitors for both HIV protease and HIV reverse transcriptase. The LigBuilder V3 program can be downloaded at "http://www.pkumdl.cn/ligbuilder3/".

Keywords: *De novo* design, Multi-target drug design (MTDD), multi-target drug optimization, Dual-functional inhibitors, LigBuilder

OPEN ACCESS

Edited by:
Jose L. Medina-Franco, National Autonomous University of Mexico, Mexico

Reviewed by:
Francesca Grisoni, ETH Zürich, Switzerland
Tingjun Hou, Zhejiang University, China

***Correspondence:**
Jianfeng Pei
jfpei@pku.edu.cn
Luhua Lai
lhlai@pku.edu.cn

Specialty section:
This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

Received: *29 October 2019*
Accepted: *14 February 2020*
Published: *28 February 2020*

Citation:
Yuan Y, Pei J and Lai L (2020) LigBuilder V3: A Multi-Target de novo Drug Design Approach. Front. Chem. 8:142. doi: 10.3389/fchem.2020.00142

INTRODUCTION

For most of the twentieth century, drug discovery process was dominated by a reductionist "one disease, one target, one molecule" philosophy (Alcaro et al., 2019). Researchers and pharmaceutical industries around the world have been struggling to develop highly specific regulators against particular targets, which are generally expected to achieve higher potencies while reducing the risk of off-target related side effects (Eaton et al., 1995; Morphy and Rankovic, 2009; Hughes et al., 2011). Although successful drugs have been brought to market with this approach, new drug R&D aiming novel targets was noticeable slowdown and fewer drugs were approved over the last decades (Scannell et al., 2012; Ramsay et al., 2018), which implies the limitation and deficiency of previous single-target drug discovery strategy. Due to the complexity of biological network (Gerstein et al., 2012), disease usually involves multiple factors and biological pathways, so agents that directly interfere individual molecular targets often lack effectiveness at treating complex diseases (Brown and Superti-Furga, 2003; Kamb et al., 2007; Cavalli et al., 2008; He et al., 2016). Moreover, the upstream components of pathways have to be regulated if only one target is aimed at in a multiple pathology related disease, which is more likely to cause unexpected side effects. Consequently, researchers and pharmaceutical industries have been turning their attention to develop therapies that modulate multiple targets simultaneously (Reddy and Zhang, 2013; Zhang et al., 2017; Kumar and Sharma, 2018). Combination therapy and multi-target therapy were proposed to address this problem.

Combination drugs, which is defined as a concerted pharmacological intervention of multiple targets with several compounds, have been used increasingly to treat many types of diseases, such as viral and bacterial infection, cancer, hypertension, and atherosclerosis (Giles et al., 2014; Von Hoff et al., 2014; Blonde et al., 2015; Lu et al., 2018). Although the combination therapy is proposed to set up a new direction for drug discovery, it is not a new concept. In fact, using multi-component mixture extracted from natural products is a historical therapy in traditional medical treatments. Besides, the highly active antiretroviral therapy (HAART) (Lu et al., 2018), which is also known as the "AIDS cocktail," has been the first-line anti-AIDS treatment since the end of last century (Bhatti et al., 2016). Many combination drugs have been launched to market and proved to be effective therapies for complex diseases, however, poor patient compliance has been raised especially in treatment of asymptomatic diseases such as hypertension (Eisen et al., 1990). An alternative way to simplify drug dosing is to mix multiple drug components into single co-formulated tablet, but different PK/PD property of each component may complicate the formulation and raise the risk of drug-drug interaction, and increase the risk and cost of such fix dose combinations strategy (Morphy and Rankovic, 2009).

Multi-target drug, which is defined as single compound that interacts with multiple targets simultaneously, has been paid much attention recently. Multi-target therapy is expected to be new and more effective medications for a variety of complex diseases even with relatively weak activity (Korcsmaros et al., 2007; Zimmermann et al., 2007). The uniform chemical component of multi-target drug will introduce lower risk of drug-drug interaction comparing with multi-components strategy. Moreover, although the discovery process of multi-target drug will be more complicated in the design and optimization stage due to the increased constraints from multiple targets, the risk and costs for the most expensive clinic trial stage are in principle similar with traditional single-target drug development. Consequently, many methods for multi-target ligand discovery were developed (Morphy et al., 2004; Zhan and Liu, 2009; Abdolmaleki et al., 2017; Zhang et al., 2017), such as multi-target QSAR (González-Díaz et al., 2006), fragment linker strategy (Morphy and Rankovic, 2006), framework combination (Morphy and Rankovic, 2006; Chen et al., 2011), and common pharmacophore based virtual screening and cross screening (Wei et al., 2008). Among them, framework combination and cross screening are both widely used approaches for discovering of multi-target lead (Morphy and Rankovic, 2005, 2009, 2010; Wu et al., 2012; Lepailleur et al., 2014; Bottegoni et al., 2016). Framework combination approach is based on the integration of multiple compounds via the fusion of common or similar sub-structure. Although the combined molecule from this approach is usually much smaller than directly linking two distinct structures with flexible chain, the median ligand efficiency is typically lower than general preclinical compounds which may lead to possible poor oral pharmacokinetics (Morphy and Rankovic, 2007). An alternative way is to screen multiple targets with the same compound library and select the consensus hints, namely, cross screening (Geppert et al., 2010). Although reported compounds derived by cross screening are better in ligand efficiency than that of framework combination approach, they are still statistically less efficient than general preclinical compounds. Considering the requirement of interacting with distinct binding sites, we are not surprising in the relative low ligand efficiency of multi-target compounds designed by the above methods (Morphy and Harris, 2012). Therefore, it is critical for multi-target compounds to be "highly integrated" that could make the most of each component group in multiple interactions. Moreover, the optimization of multi-target lead is far more complicated than that of single-target lead, because the "optimization landscape" of multi-target lead is no longer a simple stepwise "group-activity" profile in single-target lead optimization. The requirement of binding affinity balance for multiple binding will significantly reduce the available chemical space of the lead structure, as a result, stepwise optimization in multi-target optimization easily leads to "the blind alley," namely, local minima. The increased dimensions in "optimization landscape" of multi-target lead optimization make the stepwise strategy less efficient, and implies that a more global and extensive structure sampling is necessary in optimization, which may be difficult to be achieved by manual work. It also suggests that a "one-step" design rather than routine "optimizing-bioassay" cycle is more suitable for multi-target drug discovery process. Therefore, the efficient discovery strategy of "highly integrated" ligand for unrelated targets remains challenging and a general strategy of multi-target rational drug design for dissimilar targets needs to be developed.

We developed an innovative multi-target design method, called LigBuilder V3, which enables the *de novo* design and molecular optimization algorithm to handle multiple targets. The chemical space exploration algorithm inherited from LigBuilder V2 (Yuan et al., 2011) has been upgraded to explore more sophisticated structure space of multi-target ligands. As we design the multi-target ligands from scratch with the consideration of multiple interactions of each component group, high ligand efficiency is expected to be achieved with this *de novo* design approach, which is very important for multi-target drugs. Multi-target lead optimization is also implemented in LigBuilder V3, which can help researchers to find possible multi-target optimization solutions. Furthermore, we apply an "ensemble linking" strategy to promote the efficiency of "fragment linking" algorithm and make it available in linking fragments for multi-target design, which is helpful in highly efficient recombination of known ligands and framework combination.

METHOD AND ALGORITHM

Data Structure and Definition

LigBuilder V3 implements the same genetic algorithm (GA) (Fraser, 1957; Bremermann, 1958; Holland, 1975; Whitley, 1994) used in LigBuilder V2. GA is an optimization algorithm inspired by the process of natural selection, and it mimics the evolution of a population under selection pressure. LigBuilder V3 uses the overlapping generation model of GA, that is, new generation of individuals are evolved from previous population and then replace their parents with GA iteration. For a typical overlapping generation model of GA, roulette wheel selection approach

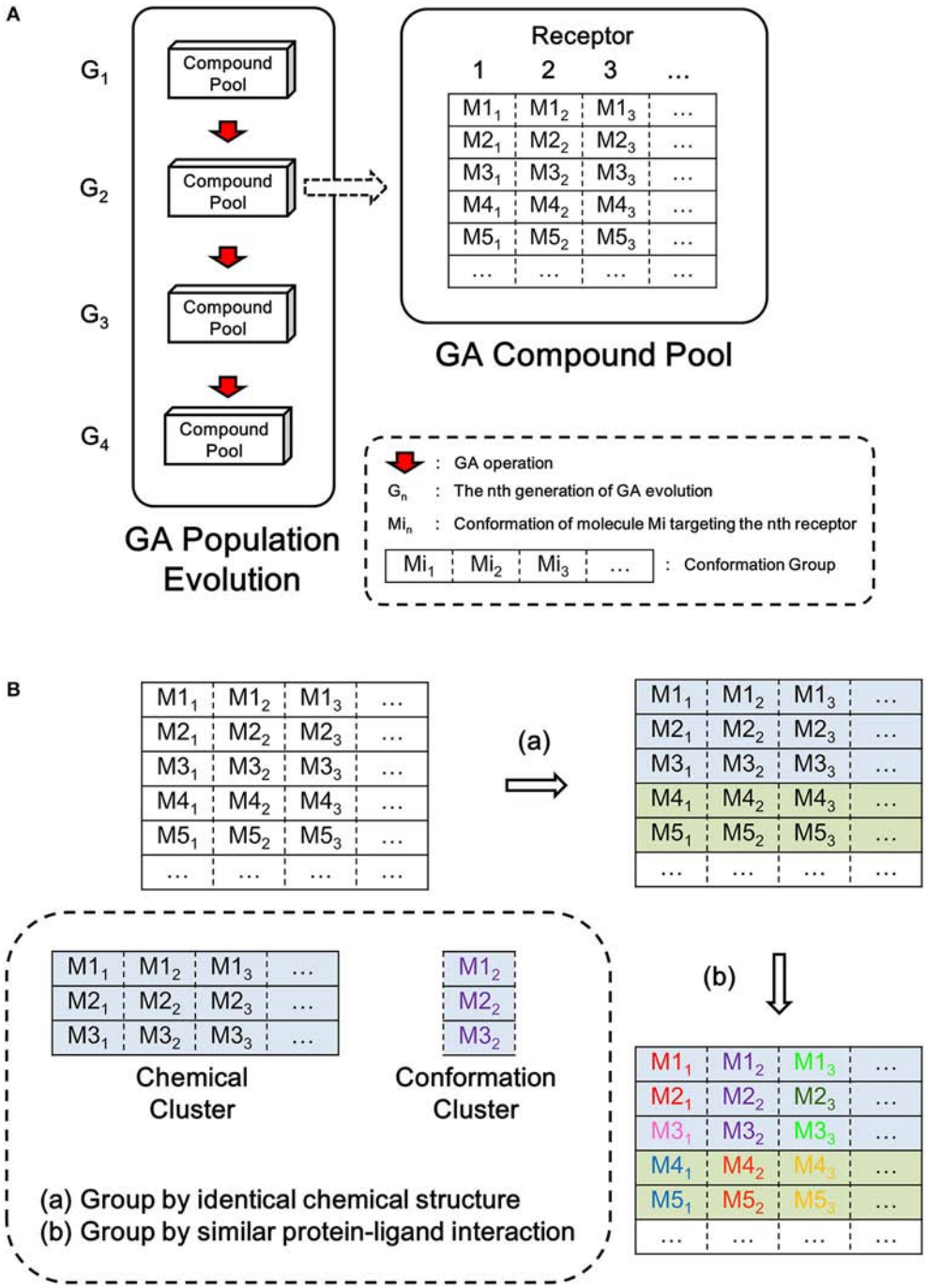


FIGURE 1 | Sketch map of data structure in LigBuilder V3. **(A)** The overview structure of GA evolution and GA compound pool. The GA compound pool includes the ensemble of molecules evolved in GA population. For multi-target design, conformations for each receptor are listed in corresponding column, as a result, each row could represent a solution of multi-target inhibition, and it could be viewed as the basic unit in multi-target GA population, which has similar status as "molecule" in single-target GA population. To avoid confusion, we define each row as a "conformation group" instead of "molecule" or "conformation." **(B)** The definition of chemical cluster and conformation cluster. The compound pool is clustered on two levels: **(a)** all molecules sharing identical chemical structures will be clustered as chemical cluster, which could be synthesized by same route; **(b)** conformations with similar protein-ligand interaction will be clustered as conformation cluster, which represent same interaction mode.

is used to select 10% members from current population as parent for evolving next generation, and all members in current population will be discarded. To balance quality and diversity of population, LigBuilder exempt the top 10% members in current population from elimination, that is, these top members will be directly transferred to next generation. So the quality of member

in offspring generation will be better, at least equal to parent generation. We define the GA compound pool as the ensemble of molecules in the newest generation of GA population evolution. The overview of the data structure used in GA evolution is described in **Figure 1A**.

As each molecule produced by multi-target design method involves multiple proteins, we should consider the multiple conformations of the molecule that bind to its corresponding targets. This is different to single target drug design method. To avoid confusion, we use "Conformation Group" instead of "Molecule" and "Conformation" to indicate the multiple binding conformations of multi-target molecule in this manuscript (**Figure 1A**).

We also define the "chemical cluster" and "conformation cluster" to depict the relationship in chemical structure and binding conformation among multi-target molecules (**Figure 1B**). From the perspective of synthesis, molecules with the same chemical structure could be considered as identical. But from the perspective of protein-ligand interaction, the conformations of ligand must be taken into account because the binding of ligand is based on spatial interaction between atoms from ligand and protein. Therefore, we cluster all conformations at two levels: (1) chemical clusters: each conformation of a chemical cluster shares the same two-dimensional (2D) structure, and they could be synthesized via the same reactions estimated by the synthesis-accessibility analysis module inherited from LigBuilder V2; (2) conformation clusters: all the conformations in a conformation cluster also share the same 2D structure, thus the conformation cluster is a subset of chemical cluster. All conformations in a conformation cluster are similar with each other, so they could be consider as sharing same interaction mode. Although all conformations of a conformation cluster are interchangeable from the perspective of interaction mode, we have to keep these "duplicates," because they may provide necessary local perturbation, for example, the members in a conformation cluster may have various orientations of hydrogen atoms. The orientations of hydrogen atoms usually have little effect on protein-ligand binding except being involved in hydrogen bond forming, but it is much sensitive in further evolution of molecules because the hydrogen atom is responsible for growing site for connecting newly added fragments.

Multi-Target Seed Structure Mapping

Seed structure is the starting point structure for lead optimization. The preparation of seed structures for single target lead optimization is straightforward, however, additional steps are needed for preparation of seed structures for multi-target design. As each "multi-target seed structure" indicates a conformation group which is composed by the different binding conformation of the ligand to each target, therefore it is necessary to make one-to-one correspondence between atoms of each member in the conformation group. Because only hydrogen atoms are possible connection site in the whole design process, the seed structure mapping is based on the mapping of hydrogen atoms, namely, hydrogen mapping. Due to the symmetry of molecule, there may be more than one possible solution of hydrogen mapping between two structures.

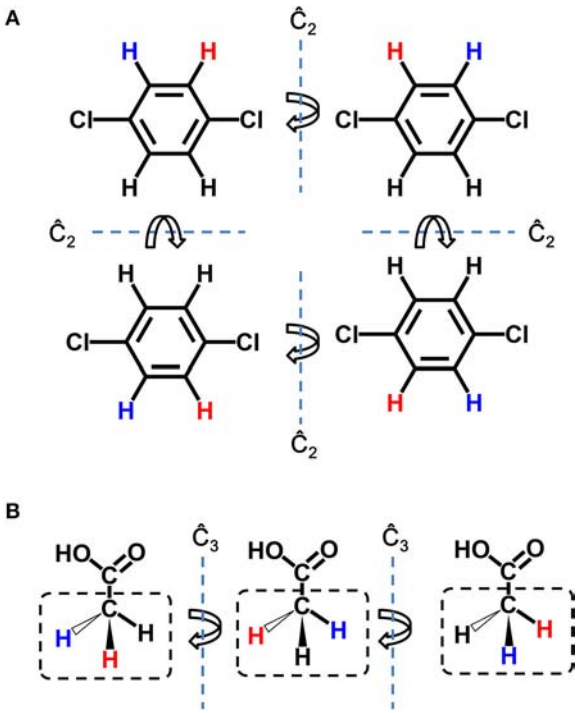


FIGURE 2 | Symmetry in hydrogen mapping. **(A)** Example for the symmetry of structure. 2 C_2 symmetry axis of 1,4-dichlorobenzene conduce to 4 possible hydrogen mappings. **(B)** Example for the symmetry of group. A C_3 symmetry axis of methyl group of acetic acid conduces to three possible hydrogen mappings. The hydrogen atom colored in red and blue indicates the first hydrogen and last hydrogen in hydrogen mapping, respectively.

As depicted in **Figure 2**, two types of symmetry should be taken into account, i.e., the hydrogen symmetry of molecule and the hydrogen symmetry of group. The molecular hydrogen symmetry refers to the rotation symmetry of all hydrogen in the molecule, and the hydrogen symmetry of group refers to the rotation symmetry of multiple hydrogen atoms that connected to one heavy atom. **Figure 2A** shows two C_2 symmetry axises of 1,4-dichlorobenzene, which conduce to 4 possible hydrogen mappings. **Figure 2B** shows a C_3 symmetry axis of the methyl group of acetic acid, which conduces to 3 possible hydrogen mappings. We should note that although some molecules such as the acetic acid are not chiral, the potential chirality is taken into account for the hydrogen mapping in LigBuilder V3, because the further growing operation may bring in chirality to the carbon atom. In other words, both 2D topological and three dimensional (3D) structural information are considered in hydrogen mapping.

Although all the hydrogen atoms in ligand are possible fragment growing site, not every hydrogen atom could serve as growing site because of steric hindrance or user's preference. Thus, LigBuilder V3 only reserves the possible hydrogen mappings with maximal growing sites mapped, named growing site mapping, which increase the possibility of further growing operation to the greatest extent. If there is no possible

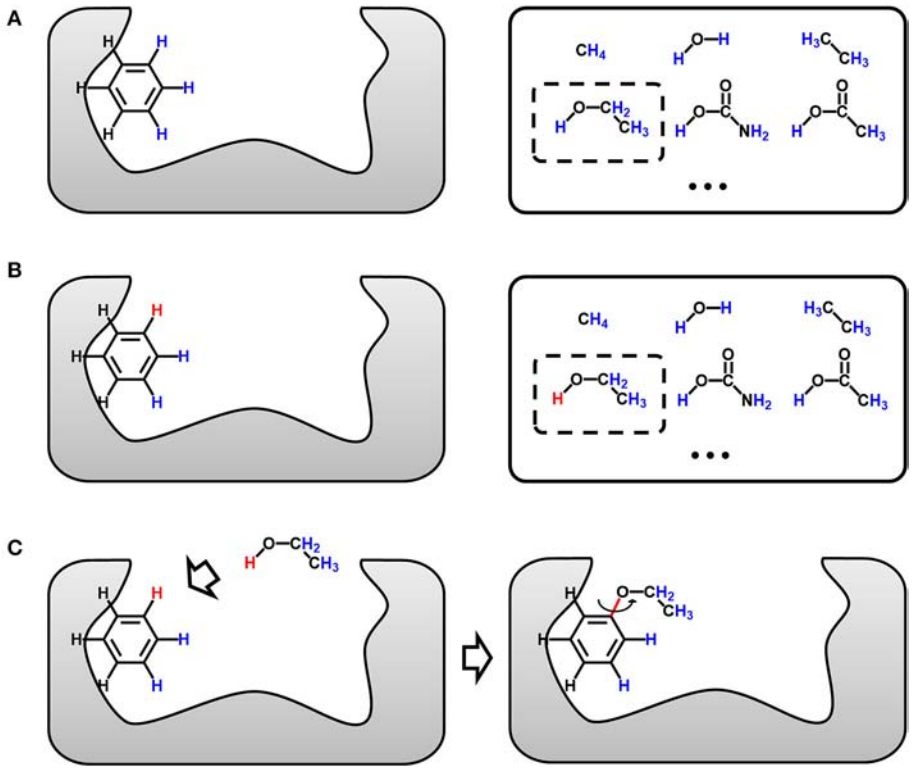


FIGURE 3 | Sketch map of the growing operation. **(A)** The seed structure in the binding site and chosen fragment (dashed box) in the building block library (solid box). All potential growing sites are colored in blue. **(B)** The randomly chosen growing site of seed structure and chosen fragment are colored in red. **(C)** The chosen fragment is moved and then attached to the seed structure. The favorable conformations are determined by uniformly 3 degree-step sampling of the torsion angle along the newly formed bond, which is colored in red.

hydrogen mapping or growing sites mapping, the corresponding conformation group will be ignored. However, if there are more than one solutions of rational hydrogen mappings, LigBuilder V3 will regard them as different seed structures and use them independently in subsequentially design process.

For the case that involves more than two targets, LigBuilder V3 makes hydrogen mapping between the conformation for the first target and each of the remaining targets one by one. As a result, all the rest conformations of the conformation group are mapped to the first conformation, so it is feasible to the find the common growing site mappings of the whole conformation group.

Multi-Target Growing

Lead optimization is the fundamental function of LigBuilder series. Both LigBuilder V1 and V2 provide the "Growing" strategy, which generates derivatives based on the lead structure (i.e., "seed" structure) that has been pre-placed into the binding pocket. In the present study, we extend the "Growing" strategy to multi-target growing (multi-target lead optimization).

Figure 3 is the sketch map of single-target growing operation, which is the basis of multi-target growing operation. The gray area on the left in **Figure 3A** represents the binding site of the target, and the benzene is a representative seed structure. Molecules in the solid box on the right are privileged fragments, which could serve as the building blocks for assembling new structure. Although all hydrogen atoms are feasible for attaching fragments, only a few of them are potential connection site without steric hindrance. Taking **Figure 3A** for example, hydrogen atoms of benzene face to the vacant region of binding site are colored in blue, which indicate the potential growing sites, and the others near to the receptor atoms will be ignored. Meanwhile, all the hydrogen atoms of building blocks will be considered as potential connection site by default. Users can also assign or block certain "growing sites" on seed structures and building blocks to customize the style of molecule. As the seed structure and building block library has been prepared, LigBuilder will randomly choose a fragment from the building block library (the dashed box in **Figure 3A**), and then randomly choose a potential growing site on the seed structure and the chosen building block, respectively (red hydrogen atoms in **Figure 3B**). The building block will be attached to the seed structure along the direction of selected hydrogen atoms (red hydrogen in **Figure 3B**). With uniformly 3 degree-step sampling of the torsion angle along the newly formed bond (red bond in **Figure 3C**), several favorable conformations with local minimal energies will be reserved as candidates in consideration of the flexibility of molecule. GA is applied

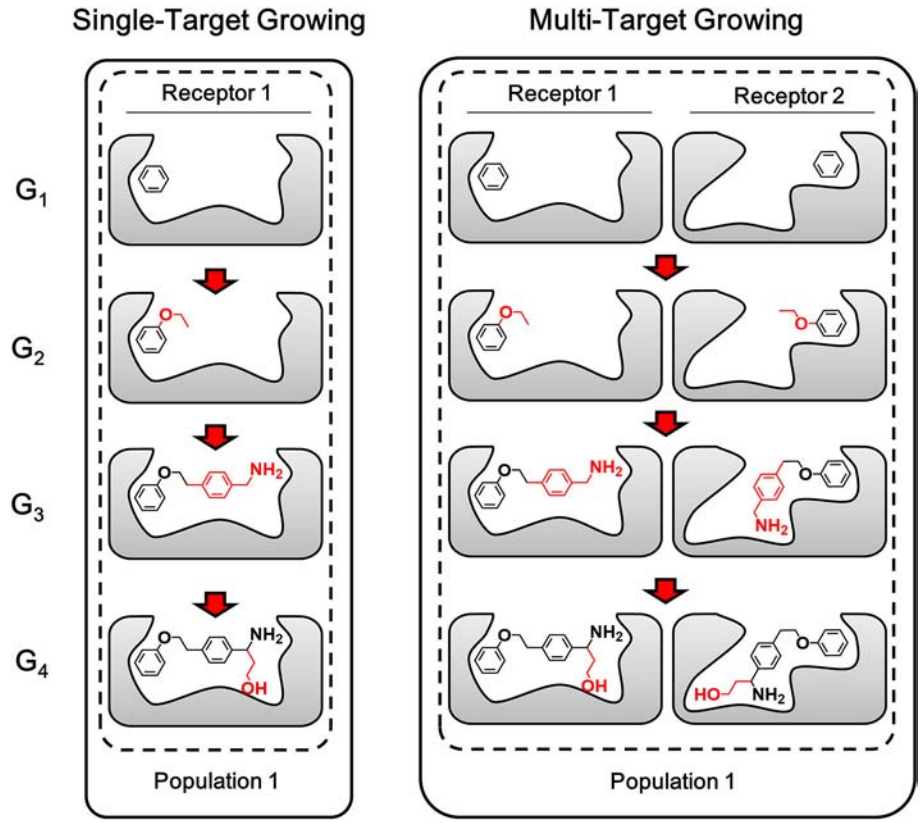


FIGURE 4 | The sketch map of growing process. The multi-target growing operation could be considered as multiple synchronous single-target growing operation. The fragments grown in each step are colored in red (Only one conformation group of each generation are showed here as representatives of the compound pool).

to select elites from these candidates, and these elites will serve as the seed structures for the next growing cycle. This repeated process for each ligand continues until: (1) the ligand is fully designed and there is no available space for adding any new chemical group; (2) the ligand reaches the limitation of molecular weight, which is 480 Da by default; (3) the GA generation number reaches a maximal number, which is 15 by default.

Different from single-target design, the lead structure for multi-target design should be prepared as "seed" conformation group, which is composed by the binding conformation of the lead structure to each target. With a simultaneous operation of growing chemically identical building block on the same site of each member in the conformation group, compounds generated by LigBuilder V3 are expected to be capable of binding to multiple targets. As depicted in **Figure 4**, multi-target growing could be considered as multiple synchronous single-target growing operation. The identical building block and the same growing site in the growing operation will maintain the consistency of 2D structures of the conformation group. Meanwhile, the 3D conformation of ligand is only restrained by its corresponding targets, that is, the conformation in each conformation group is optimized and evaluated independently. Therefore, this strategy could utilize the flexibility of ligand to improve the capability of multi-target binding. Genetic Algorithm (GA) is also applied to manipulate the growing cycle in the same manner as single target growing.

Ensemble Linking

Although assembling several bioactive fragments to generate potent ligand is very promising, the computational method of linking proximal fragments covalently is fraught with challenges. To avoid affecting respective bioactivity, the orientation and position of fragments should be changeless. Therefore, the feasibility of linking is severely limited by the rigid restriction of bond length and bond angle in molecule. Besides, the unfavorable energy of torsion may further reduce the feasibility. Although there may be some solutions existed in huge chemical space, the low efficiency in finding these solutions narrows the application of fragments linking. As a result, there are few successful cases of fragments linking, except using flexible chain as linker. However, although flexible chain could be used to relax the rigid restriction of linking, it may increase the amount of accessible conformation of the structure which brings in unfavorable entropy change during binding process, thus the linked fragments usually do not bind as the same degree as the sum of the individual fragments.

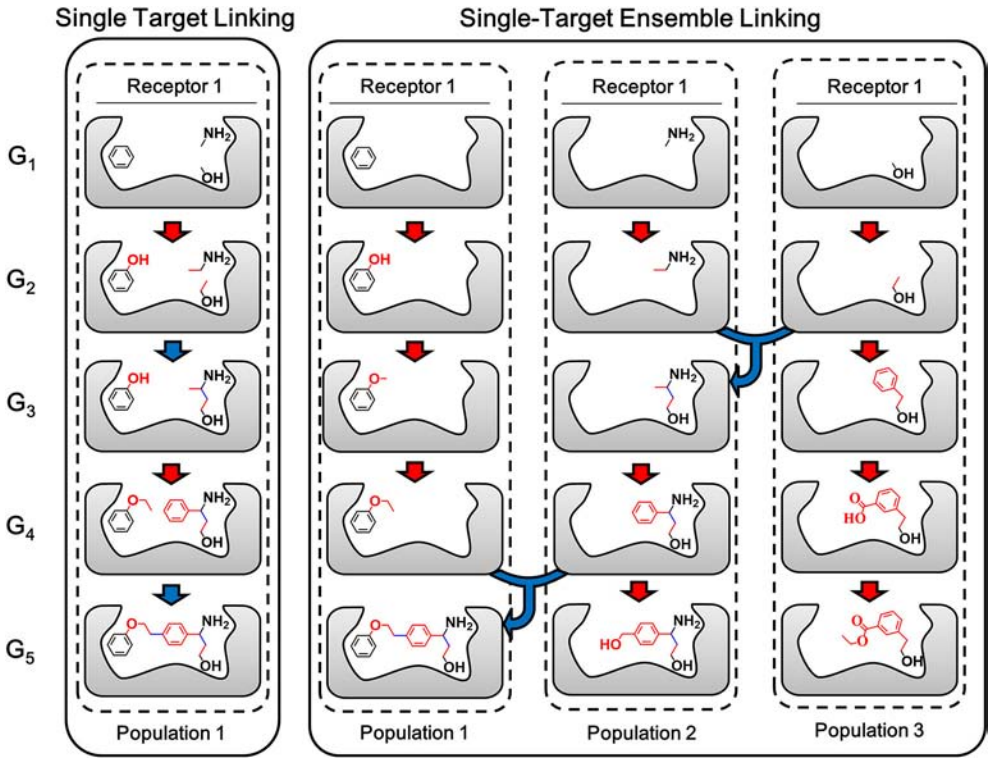


FIGURE 5 | The sketch map of linking and ensemble linking process. The red arrow indicates the growing operation while blue arrow indicates the linking operation. In accordance with the color of arrows, atoms, and bonds formed in growing and linking operation are colored in red and blue, respectively. For previous linking algorithm, the aggregation of certain fragments is regarded as a whole "seed," so the linking process will be constrained by every fragment in the aggregation. Instead, ensemble linking algorithm applies a more flexible linking strategy, which attempts to derive new structures from each of seed fragments independently and then find possible way of linking among these structures. The different populations indicated different independent GA threads.

Moreover, the excessive flexibility of structure may reduce the specificity of ligand.

The ensemble linking algorithm is developed in LigBuilder V3 to improve the efficiency of fragments linking process, which is expected to make this method more practical. To demonstrate the details of the new algorithm, the sketch map of the linking algorithm used in previous versions of LigBuilder series and ensemble linking algorithm used in LigBuilder V3 are compared in **Figure 5**. Previous linking algorithm applies a direct strategy of linking, which aims to linking certain fragments with many building blocks. Instead, ensemble linking algorithm applies a more flexible linking strategy, which attempts to derive new structures from each of seed fragments independently and then find possible way of linking among these structures. Although both algorithms are capable of generating the same final structure in **Figure 5**, ensemble linking strategy is expected to be more efficient. For the general linking algorithm, it is straightforwardly requiring that all the given fragments should be linked, which is usually hard especially for multi-target linking. To overcome this problem, ensemble linking algorithm is based on extra linking fragments, and automatically find the apportioned combination of fragments, which would improve the possibility of finding solution for linking fragments. To be specific, for general linking algorithm, the number of fragments used for linking is limited, for example, user poses 3 fragments into the ligand binding site, and the linking algorithm attempts to find suitable linkers to connect all these 3 specific fragments. For ensemble linking, user could pose several thousands of fragments into the ligand binding site, and the ensemble linking algorithm attempts to find suitable linkers to connect any 3 fragments among all available fragments. Obviously, the ensemble linking algorithm will significantly increase the possibility of finding suitable solution for linking 3 fragments comparing with general linking algorithm. Besides, with dissociation of combined seed fragments in linking algorithm, the whole linking process would be more robust, it would not be dragged by improper derivation or conformation of individual fragment. Moreover, it raises the possibility of comprehensive utilization of more bioactive fragments without exhaustive combination. As a result, LigBuilder V3 could be applied to find possible solutions of linking among hundreds of fragments, which further improves the success rate of linking.

The ensemble linking algorithm will generate many derivative candidates based on the linking fragments to enhance the possibility for finding solution, obviously, although the possibility of linking increases with the number of candidates, the computation cost also will increase by the same rate. So it is important to generate candidates more effective rather than

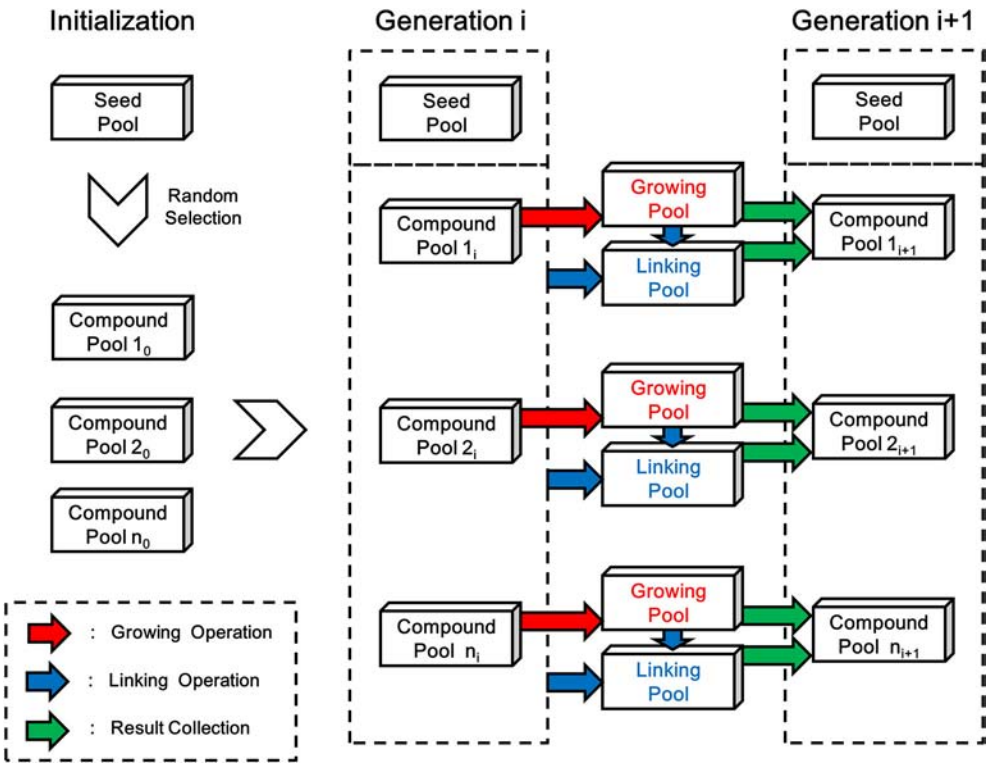


FIGURE 6 | The flow chart of linking process in LigBuilder V3. The structures of initial compound pool are randomly selected from the seed pool in the initialization stage. Then LigBuilder V3 perform the growing and linking operation on each compound pool for producing new generation of compound pools. The growing operation brings in derivation on molecules in previous compound pool and result in a "Growing pool," which is colored in red. Then the linking operation will be performed for finding possible way of linking between these newly formed structures and previous existed structures. The "previous existed structures" in each generation is denoted with the dashed box. As similar with growing operation, the linking results will be collected into a "Linking pool" which is colored in blue. At last, the "Growing pool" and "Linking pool" will be merged into a new generation of compound pool.

increase the number of candidates to improve the efficiency. As all candidates in the GA population have similar molecular weight because they are generated by GA evolution with same number of generations, the linking possibility will reach the peak when candidates in the population occupy about half of the binding site. But it rapidly falls when candidates in the population are too large to be integrated in limited space of binding site. In addition, the linking possibility is also low when the candidates in the population are too small which may make them far away from each other for linking. So we applied a stagger strategy that operating several independent GA threads simultaneously, meanwhile, the starting of each GA threads are staggered so as to make them be in various generation of GA process. That is, ensemble linking algorithm will not only perform "intra-linking" among candidates in a GA process, but also perform "inter-linking" among candidates in different GA threads. With this strategy, the high diversity of molecular weight distribution among all candidates could bring in higher linking possibility and efficiency.

As depicted in **Figure 6**, each generation of ensemble linking can be decomposed into two steps, i.e., the growing step and the linking step. LigBuilder V3 performs the growing operation on all compounds from each compound pool in the growing step, and then finds possible way of linking between the newly formed compounds and all previous existed compounds including seed pool (dashed box in **Figure 6**) in the linking step. Although compounds generated in both steps will be collected together into new generation of compound pools, the compounds generated in the linking step (linking pools in **Figure 6**) will have a certain level of priority in GA process, which make the ensemble linking algorithm trends to link fragments rather than grow for derivation. To be specific, the compounds generated in the linking step indicates a "linking" operation is occurred, on the contrary, compounds generated in the growing step do not link with other fragments in this step. So LigBuilder will elevate the fitness score of compounds from linking step, which encourage the linking behavior. The structures of initial compound pool are randomly selected from the seed pool. After the initialization, LigBuilder V3 will repeat the ensemble linking process for each ligand until: 1) the ligand is fully designed and there is no available space for any new chemical group; 2) the ligand reaches the limitation of molecular weight, which is 480 Da by default; 3) the GA generation number reaches a maximal number, which is 15 by default.

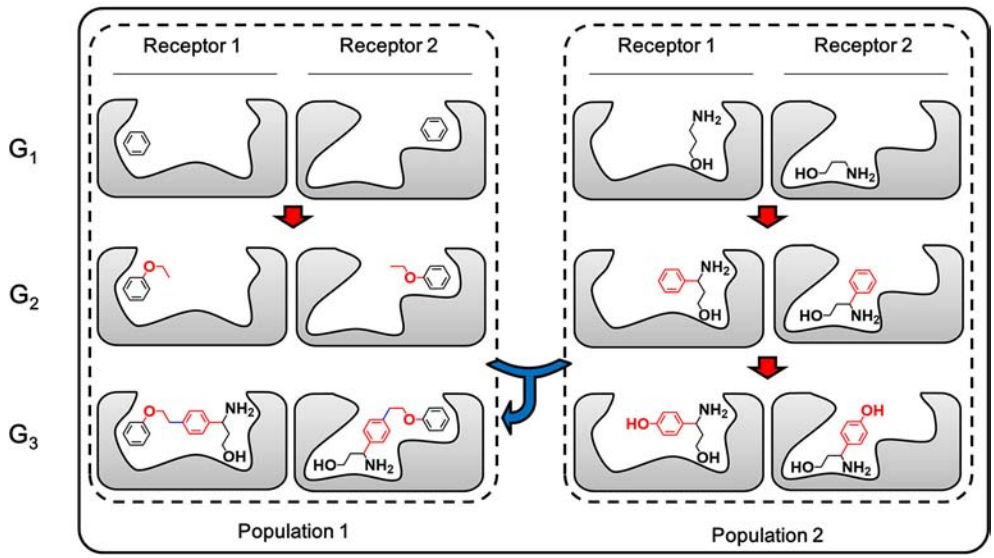


FIGURE 7 | The multi-target ensemble linking operation could be considered as multiple synchronous single-target ensemble linking operation. The red arrow indicates the growing operation while blue arrow indicates the linking operation. In accordance with the color of arrows, atoms, and bonds formed in growing and linking operation are colored in red and blue, respectively.

Multi-Target Linking

A further challenge lies in designing multi-target ligand is linking fragments that interacting with multiple targets. Although many successes occurred in designing single target ligand by fragments linking strategy, few research focus on multi-target linking method. Comparing with lead compound, potential active fragments are much easier to pick by fragment-based approach, such as NMR, DSF, X-ray crystallography, surface plasmon resonance and mass spectrometry (Mashalidis et al., 2013). In addition, computational methods such as fragment docking (Wang et al., 2015) or CrystalDock (Durrant et al., 2011) are also effective ways to identify lead-fragments. Moreover, small fragments are much more likely to interact with multiple targets due to its lower specificity. Therefore, it is feasible and promising to design multi-target ligand by integrating several fragments. So we try to improve our ensemble linking algorithm to handle multi-target fragments linking in LigBuilder V3.

As with the multi-target growing algorithm, the fragments for multi-target linking should also be prepared as "seed" conformation group. Because the ensemble linking algorithm in LigBuilder V3 handle the fragments independently, users no longer have to predetermine which fragments will be linked together in the stage of seed preparation. That is, the procedure of conformation group preparation for linking is same as that for growing. The only difference is that at least 2 conformation groups should be prepared for linking and at least 1 for growing.

Multi-target linking algorithm is based-on the ensemble linking algorithm described in above section, therefore both the growing step and linking step which make up the ensemble linking algorithm will be extended to multiple targets. The growing step of multi-target ensemble linking is exactly the same with multi-target growing, while the linking step applies the same strategy of "multi-target operation" used in multi-target growing. As described above, multi-target growing operation is a simultaneously operation of growing chemical identical building block on the same site of each member in the conformation group. In a similar way, multi-target linking operation in linking step is a simultaneously pairwise operation of linking corresponding members from two conformation groups on the same linking sites (**Figure 7**). That is, the first member of conformation group A will be linked with the first member of conformation group B. Then the second member of both conformation group will be linked together on the same linking sites of the first member. This pairwise process will be repeated until all members have been linked or any failure occurs due to steric hindrance or molecular tension. As a result, these linked structures are expected to be capable of binding to multiple receptor, while they share identical chemical structures.

Multi-Target *de novo* Design

LigBuilder V3 inherits the "Chemical Space Exploring Algorithm" (CSEA) from LigBuilder V2 to create novel scaffolds and structures. In LigBuilder V2, CSEA works in the following way: (1) an sp^3 carbon with 4 hydrogen atoms will be randomly posed in the binding site and serve as the starting point of constructing new molecules with growing operation; (2) newly designed molecules will be split into fragments; (3) fragments with high predicted binding affinity, that is, high contribution fragments will be selected for updating the "seed structure pool," which is used to supply seeds for subsequent design cycles;

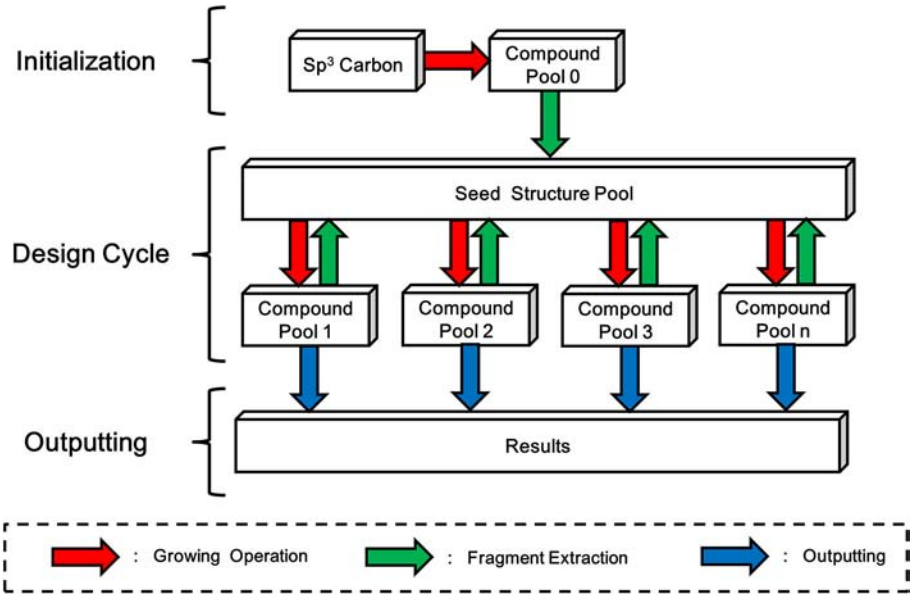


FIGURE 8 | Schema of Chemical Space Exploring Algorithm (CSEA). A randomly positioned sp^3 carbon will be taken as starting point of growing operation for generating the initial compound pool. Then fragments extracted from compound pool will be used to construct a seed structure pool, which contains seed structures that more favorable than the initial sp^3 carbon. As new compound pools could be generated based-on structures in seed structure pool, while fragments extracted from these newly formed compounds will be used to update the seed structure pool, LigBuilder V3 avoids the dependency of initial structure and be promising in exploring larger chemical space.

(4) a structure from "seed structure pool" is randomly select as the starting point of constructing new molecules with growing operation, then the 2–4 steps will be repeated. For LigBuilder V3, CSEA is extended to multi-target design purpose by applying multi-target growing operation instead of single-target growing operation. Meanwhile, the fragment extraction process of single conformation will also be replaced by fragment extraction of "conformation group" (**Figure 8**). With the self-circulation seed generating feature, CSEA can help to avoid the limitations associated with pre-assigned seed structures and explore a broader chemical space, thus greatly improving the novelty and efficiency of design.

As the seed structure pool is used to collect and provide initial fragments of design process, the quality of seed structure pool may significantly affect the design results. Since CSEA could provide a mass of potential seed fragments during the design process, the seed structure pool will be updated to achieve higher binding affinity while maintain diversity of seeds. Then the CSEA will have a higher starting point for generating potent structures, which in turn produce better seed fragments for updating the seed structure pool. Therefore, the seed structure pool will keep evolving during the whole design process, which iteratively optimize the performance of design.

Seed Structure Extraction

The most direct way of extracting component fragments from chemical structure is splitting the molecule by iterating over all single bonds. However, the traversal extraction method would take a lot of computing time when handle millions of compounds, which is a common order of magnitude in CESA process of LigBuilder V3. Therefore, we develop a simplified extraction algorithm for acceleration. To balance the representativeness of fragments and extraction speed, we only focus on the molecular scaffold and key interaction group, which are major determinants of molecular conformation and protein-ligand interaction. As a result, only five categories of fragments are considered in CESA: (1) Single atoms; (2) rigid scaffold (rigid chain and rigid ring system); (3) flexible scaffold (flexible ring system); (4) interaction group (e.g., carboxyl group); (5) scaffold with connected interaction group. If a fragment could be classified into more than one category, it will be put into the category with smallest category number. With this algorithm, LigBuilder V3 is also capable of extracting fragments from known ligands as seed structures, which is convenient for fragments linking or lead optimization.

For multi-target design, as the seed structure is composed of multiple conformations (conformation group) instead of single conformation, seed structures extracted from known ligands of each targets should be paired to construct conformation groups first. LigBuilder V3 extracts fragments from all known ligands in the same way of extracting fragments in single-target design, and then hydrogen mapping algorithm will be implemented among these fragments to find all possible combinations. That is, if a fragment is present in ligand of every targets, it is a common fragment and LigBuilder will mapping this fragment to construct multi-target fragments group as seed structures. However, if this fragment is absent in ligands for any one of the targets, it is not common fragment and the fragment will be discarded.

Multi-Target Ligand Efficiency

Ligand efficiency (LE) is frequently used to prioritize hits from HTS, and it can be regarded as a guide for selecting efficient fragments for further optimization. It is much more important for multi-target design because the ligand efficiency is also the index of "integration degree," which is a more critical index for multi-target drugs. There are various definitions of LE but most widely used approximation is described as the average free energy of binding per heavy atom or average pIC_{50} per heavy atom, which are demonstrated as follows (Hopkins et al., 2004):

$$LE = \frac{-\Delta G}{HAC} \quad or$$
$$LE = \frac{-\log(IC_{50})}{HAC} \qquad (HAC\ is\ heavy\ atom\ count)$$

For multi-target ligand, multi-target ligand efficiency (MLE) could be derived with a similar form of LE. As the multi-target ligand causes multiple binding free energy, MLE could be described as the summation of average free energy of binding per heavy atom or summation of average pIC_{50} per heavy atom, which are demonstrated as follows:

$$MLE_N = \frac{\sum -\Delta G_n}{HAC} \quad or$$
$$MLE_N = \frac{\sum -\log\left(IC_{50_n}\right)}{HAC} \quad (N\ is\ the\ number\ of\ target)$$

The ligand efficiency is not comparable if the target number is different, so we use the subscript for MLE to indicate the condition of rational comparison and make it distinct from LE of single-target ligands. We should note that the MLE is insufficient for the performance evaluation of multi-target ligand, because the uneven activity of targeting individual binding site may be obscured by summation. However, it is much complicated to evaluate the efficiency of multi-target ligand, because it depends on the specific biological network that it is involved in. So the MLE would be only considered as an index of average efficiency for selecting potential multi-target lead structures.

Other Functional Modules

The other functional modules implemented in LigBuilder V3 are directly inherited from LigBuilder V2, including: (1) drug-like and privileged building blocks; (2) Toxic fragments; (3) Drug-like rules (for example, Lipinski rule); (4) Ligand-binding site detection module; (5) Synthesis analysis modules; (6) Scoring function; (7) LogP module; (8) GA fitness function (composed of scoring function, MLE, toxic fragments filter, and drug-like rules). It should be noted that the binding affinity predicted by scoring function for multi-target ligand is calculated by its average of binding affinity predicted for each target.

Design HIV-PR/HIV-RT Dual-Functional Inhibitor

Structural Preparation

The crystal structures of PR and RT used in this study are downloaded from the RCSB Protein Data Bank (Berman et al., 2003) (PDB code: 3A2O, Hidaka et al., 2009, and 4G1Q, Kuroda et al., 2013, respectively), and both structures are complexes with potent inhibitors solved at high resolution (0.88 and 1.51 Å, respectively). The inhibitor binding sites of PR and RT were defined by binding site detection program Cavity (Yuan et al., 2013; Zhang et al., 2015; Xu et al., 2018), which provides the detailed definition for boundary of "design space." The drug-like and privileged building blocks used in this study were inherited from LigBuilder V2. Then three different design strategies were used to design dual-function inhibitors for PR and RT.

De novo Design Approach

LigBuilder V3 inherited the seed generation and optimization algorithm from LigBuilder V2, that is, LigBuilder V3 could iteratively extract seed structures from designed compounds and use the extract seed structures for design new compounds. The GA parameters were set as follows: GA population size of 1,000, GA parent ratio of 10%, GA generation number of 12. Total 1 million candidate dual-functional compounds were generated by LigBuilder V3 with *de novo* design mode.

Growing Approach

Growing approach is for optimization of prepared seed structures. We collected all protein-ligand binding complex of PR or RT from the RCSB Protein Data Bank (Berman et al., 2003), including 323 PR-ligand complexes and 141 RT-ligand complexes (Listed in **Table S1**). All the PR-ligand complexes were aligned to the PR structure complexed with KNI-1689 (PDB code: 3A2O, Hidaka et al., 2009), and all the RT-ligand complexes were aligned to the RT structure complexed with Rilpivirine (PDB code: 4G1Q, Kuroda et al., 2013) using Pymol (Schrodinger, 2010), which could ensure that all ligands in the complexes are also aligned according to the receptor alignment. With the fragment-extraction function of LigBuilder V3, fragments with no more than 20 heavy atoms were extract from these known PR or RT ligands. As the ligand efficiency is important for the seed structure, fragments with SLE index <0.1 were removed, and a total of 2,386 fragments for PR and 1,442 fragments for RT were obtained at this stage. Then fragments for PR and RT with the same 2D structure were paired with "hydrogen mapping" algorithm mentioned above, and a total of 3,506 paired fragments were prepared. The GA parameters were set as follows: GA population size of 1,000, GA parent ratio of 10%, GA generation number of 12. Total 100 K candidate dual-functional compounds were generated by LigBuilder V3 with growing design mode based on the prepared fragments. As the binding affinity is usually related to the size of molecule, large seed fragments are more competitive than small fragments especially for the genetic algorithm used in LigBuilder. So, each fragment was independently used as the seed structure with multiple runs of LigBuilder to avoid bias to large seed fragments.

Linking Approach

Linking approach is for integrating key fragments into new compounds. The paired fragments used in this approach were prepared in the same way of growing approach. However, all paired fragments were used together in linking approach to maximize the possibility of finding ways for linking fragments.

It would be intuitive that the more fragments provided, the better performance would be expected. The GA parameters were set as follows: GA population size of 10,000, GA parent ratio of 10%, ensemble population number of 10, GA generation number of 12. Total 10 K candidate dual-functional compounds were generated by LigBuilder V3 with linking design mode based on the 3,506 prepared fragments.

Post-processing

As multi-target ligand should bind to different proteins with the same chemical structure, ideally, each moiety of the ligand could contribute to its binding to all targets, so ligand efficiency would be important in evaluating a multi-target ligand. In this study, predicted pKd of all the output compounds are larger than 5.0, so only MSLE index were used to rank and select top 1,000 results with best ligand efficiency from multi-target design procedure for the three approaches. Because LigBuilder V3 only uses a fast empirical scoring function for estimating protein ligand binding affinity, in order to improve the accuracy of calculation, the total 3,000 selected compounds were further subjected to energy minimization and 100 ps short time molecular dynamic (MD) simulation by using the Amber package (Case et al., 2012) for

TABLE 1 | Binding free energy predicted by MM/GBSA method.

	***De novo* design**		**Growing approach**		**Linking approach**		**HIV-PR potent inhibitor[b]**	**HIV-PR weak inhibitor[c]**	**HIV-RT potent inhibitor[d]**	**HIV-RT weak inhibitor[e]**
	Top 10[a]	**Top 1**	**Top 10**	**Top 1**	**Top 10**	**Top 1**				
PR	−28.2	−35	−28.4	−32	−34.9	−35.0	−68.2	−21.9	–	–
RT	−33.6	−34.9	−34.5	−40.3	−38.0	−38.8	–	–	−41.5	−22.6
Ave.	−30.9	−35.0	−31.5	−36.2	**−36.5**	**−36.9**	–	–	–	–

Energy unit is kcal/mol.

[a]Top 10 indicates the average of best 10 compounds; [b]FDA approved drug Darunavir with IC_{50} of 0.15 nM (Shen et al., 2013). Complex structure was from PDB code 4LL3 (Kožíšek et al., 2014). [c]PR weak inhibitor with IC_{50} of 2.3 μM (Jhoti et al., 1994). Complex structure was from PDB code 1HTE (Jhoti et al., 1994). [d]FDA approved drug Efavirenz with IC_{50} of 41 nM (King et al., 2002). Complex structure was from PDB code 1FK9 (Ren et al., 2000). [e]RT weak inhibitor with IC_{50} of 1.2 μM (Chan et al., 2017). Complex structure was from PDB code 5VQS (Chan et al., 2017). Bold values indicate the best average energy among results from three approaches.

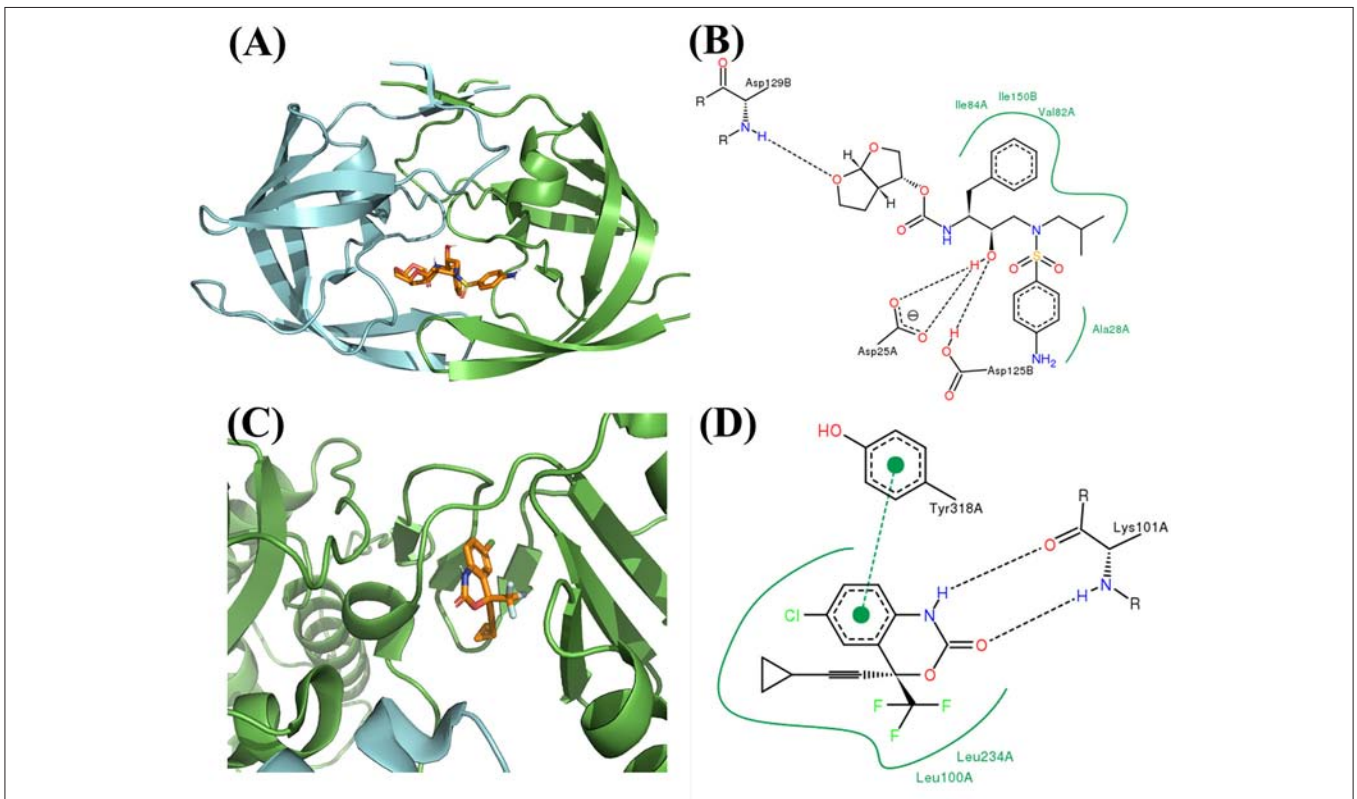


FIGURE 9 | (A) Binding mode of FDA approved PR inhibitor Darunavir. The dimerized PR is showed in green and cyan cartoon style, and the Darunavir is showed in golden stick style (figure generated by Pymol, Schrodinger, 2010). **(B)** The 2D interaction figure for Darunavir binding with PR (figure generated by PoseView, Stierand and Rarey, 2010). **(C)** Binding mode of FDA approved RT inhibitor Efavirenz. **(D)** The 2D interaction figure for Efavirenz binding with RT.

estimating the binding affinity with MM/GBSA method (Rastelli et al., 2010).

Application of LigBuiler V3 in Multi-Target Ligand Design

The concept prototype for growing mode algorithm of LigBuilder V3 has been experimental validated by designing COX2/LTA_4H dual-functional inhibitor, which resulted in a single ligand that binding to COX2 and LTA_4H with IC_{50} of 7.1 and 7.0 μM, respectively (Shang et al., 2014). Although this work is based on a developing version of LigBuilder V3, and many manual interventions were involved due to immature of the algorithm, the success of this case suggests the feasibility of using LigBuilder V3 to design multi-target ligand. Moreover, LigBuilder V3 were further developed based on the knowledge learned from this case. Besides the improvement of multi-target growing algorithm, both multi-target *de novo* design approach and multi-target linking approach are realized in this version of LigBuilder V3. In this study, we have tested the LigBuilder V3 by designing dual-functional inhibitor targeting two well-characterized virus enzymes, HIV protease (PR) and HIV reverse transcriptase (RT) with all three design modes. As both PR and RT are important drug targets of clinical antiretroviral therapy, the multi-target strategy such as combination of nucleoside reverse transcriptase inhibitors (NRTI) and protease inhibitor (PI) shows significant advantage over each single component and has been broadly used for HIV treatment (Lu et al., 2018). Consequently, researchers have been interesting in developing cocktail drug combinations, and pursue multi-target anti-HIV inhibitors for improving patient compliance. Matsumoto *et al.* have reported the strategy of linking PR and RT inhibitor by spontaneously cleavable linker (Matsumoto et al., 2000). Furthermore, scaffold merging strategy is successfully applied in designing multi-target anti-HIV inhibitors in recent years (Song et al., 2015; Sun et al., 2016). However, both the dependency of known inhibitors and specific requirement of molecular structure limit the practical applications of structure merging strategy. Therefore, we present a more universal solution of multi-target design with the example of designing dual-functional inhibitors for PR and RT by LigBuilder V3. The detailed methods and parameters are described in the Method and Algorithm section.

The top 1,000 compounds from each design modes were selected and subjected to 100 ps short time molecular dynamic simulation, then the binding affinity of each compounds were estimated by MM/GBSA method. The average binding affinity of the top 10 compounds and top 1 compound for each design modes were collected in **Table 1**. Although the designed multi-target compounds could not compare with the super potent PR and RT inhibitors with sub-nanomolar level activity, the designed compounds are predicted to be more potent than micromolar level inhibitor of both PR and RT, that is, these compounds are expected to be dual-functional inhibitor for PR and RT at sub-micromolar level activity for both targets.

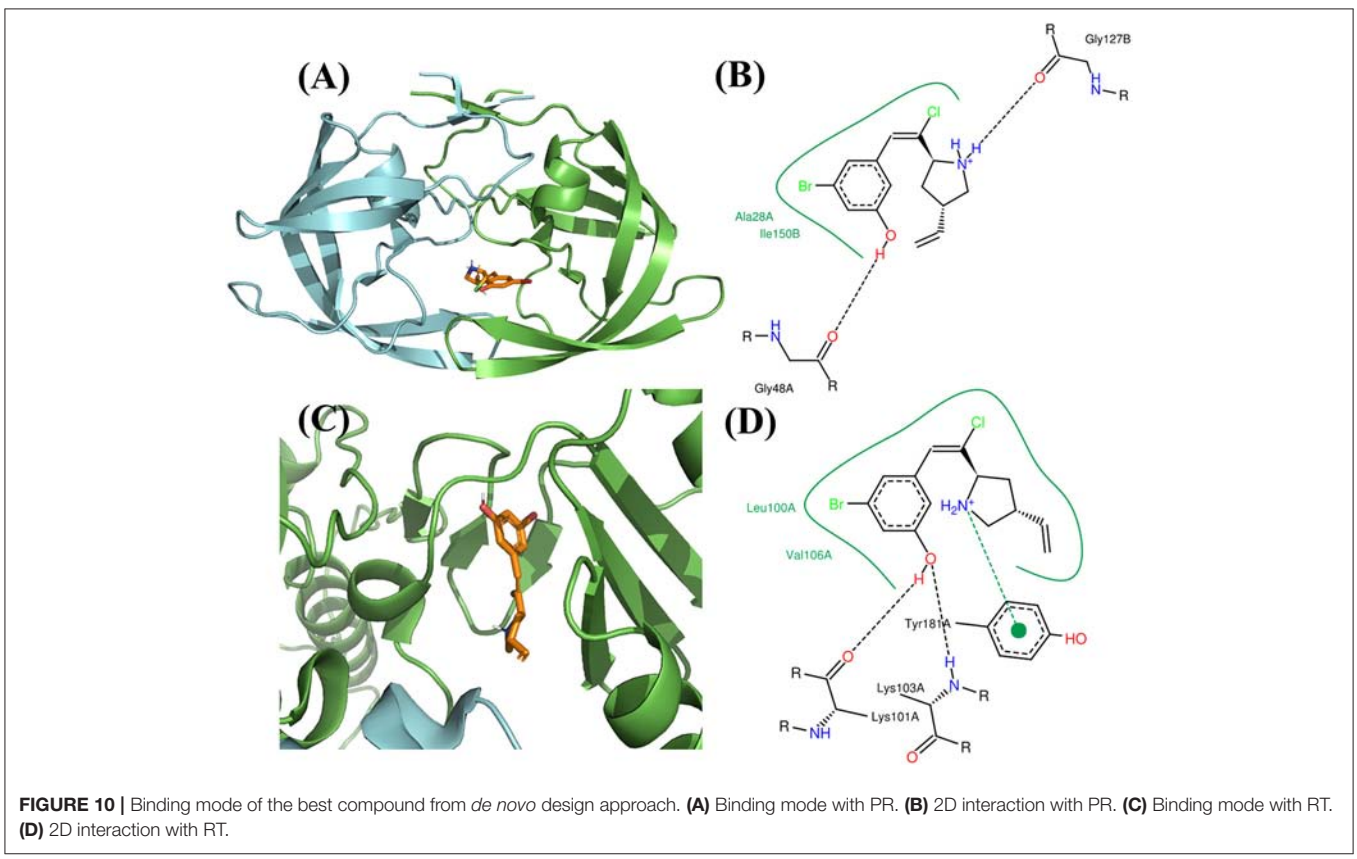


FIGURE 10 | Binding mode of the best compound from *de novo* design approach. **(A)** Binding mode with PR. **(B)** 2D interaction with PR. **(C)** Binding mode with RT. **(D)** 2D interaction with RT.

The binding mode of FDA approved PR inhibitor and RT inhibitor are depicted in **Figure 9**. Obviously, they adopt very distinct protein-ligand interaction modes. The best compound from *de novo* design approach is depicted in **Figure 10**. This compound forms distinct interaction comparing with known PR or RT inhibitors. It is highly compact and fully utilized its polar groups and hydrophobic groups to form interaction with PR and RT in different manner. The best compound from fragments growing approach is depicted in **Figure 11**. This compound is growing from a benzene ring which is one of the most common fragments in PR and RT inhibitors. The best compound from fragments linking approach is depicted in **Figure 12**. This compound is much bigger than compounds from *de novo* approach and growing approach, which indicates its relatively lower ligand efficiency. As linking approach is intensively pursing possible ways for linking provided fragments, the success of linking is more important than ligand efficiency, so the algorithm is preferable to allow generating derivates with much lower ligand efficiency which may enhance the possibility of linking. Overall, all of these compounds are relatively small, and groups in these compounds usually contribute to the binding with different protein in different manner, which is the most need feature for designing highly compact multi-target ligand.

Essentially, the design process of LigBuilder is a kind of "random evolution" process implemented by genetic algorithm. So the quality of design result is expected to be improved along with the total computational time. As the output result could be unlimited, it is not realistic to achieve the maximal quality. For a fair comparison among the results from three strategies, we used different output number to ensure they consumed roughly similar computational time. So we designed 1 million, 100 K, and 10 K compounds for *de novo* approach, growing approach, and linking approach, respectively, which is roughly corresponding to the compound generating efficiency of three approaches in

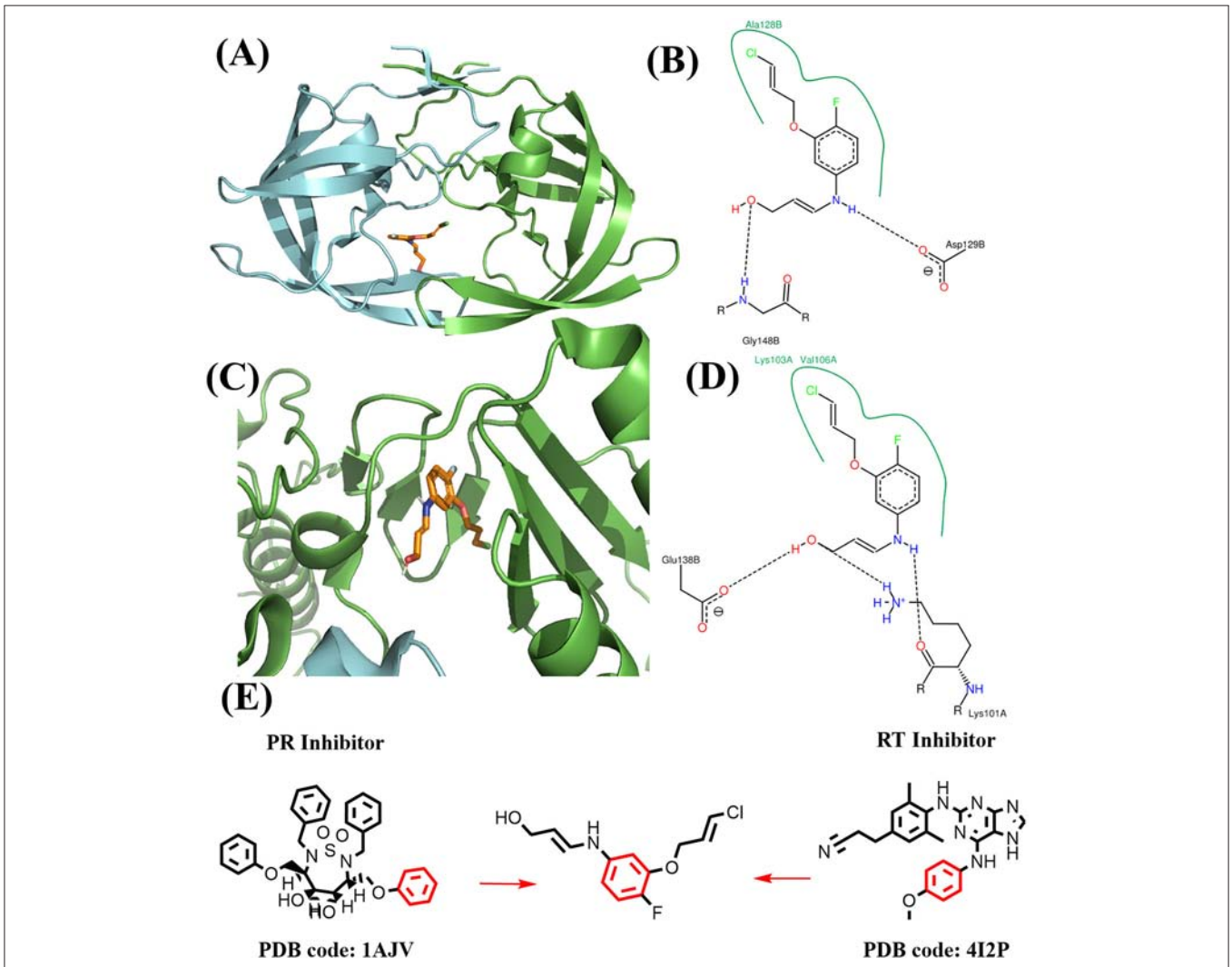


FIGURE 11 | Binding mode of the best compound from growing approach. **(A)** Binding mode with PR. **(B)** 2D interaction with PR. **(C)** Binding mode with RT. **(D)** 2D interaction with RT. **(E)** Source of fragments in PR inhibitor and RT inhibitor for growing are colored in red.

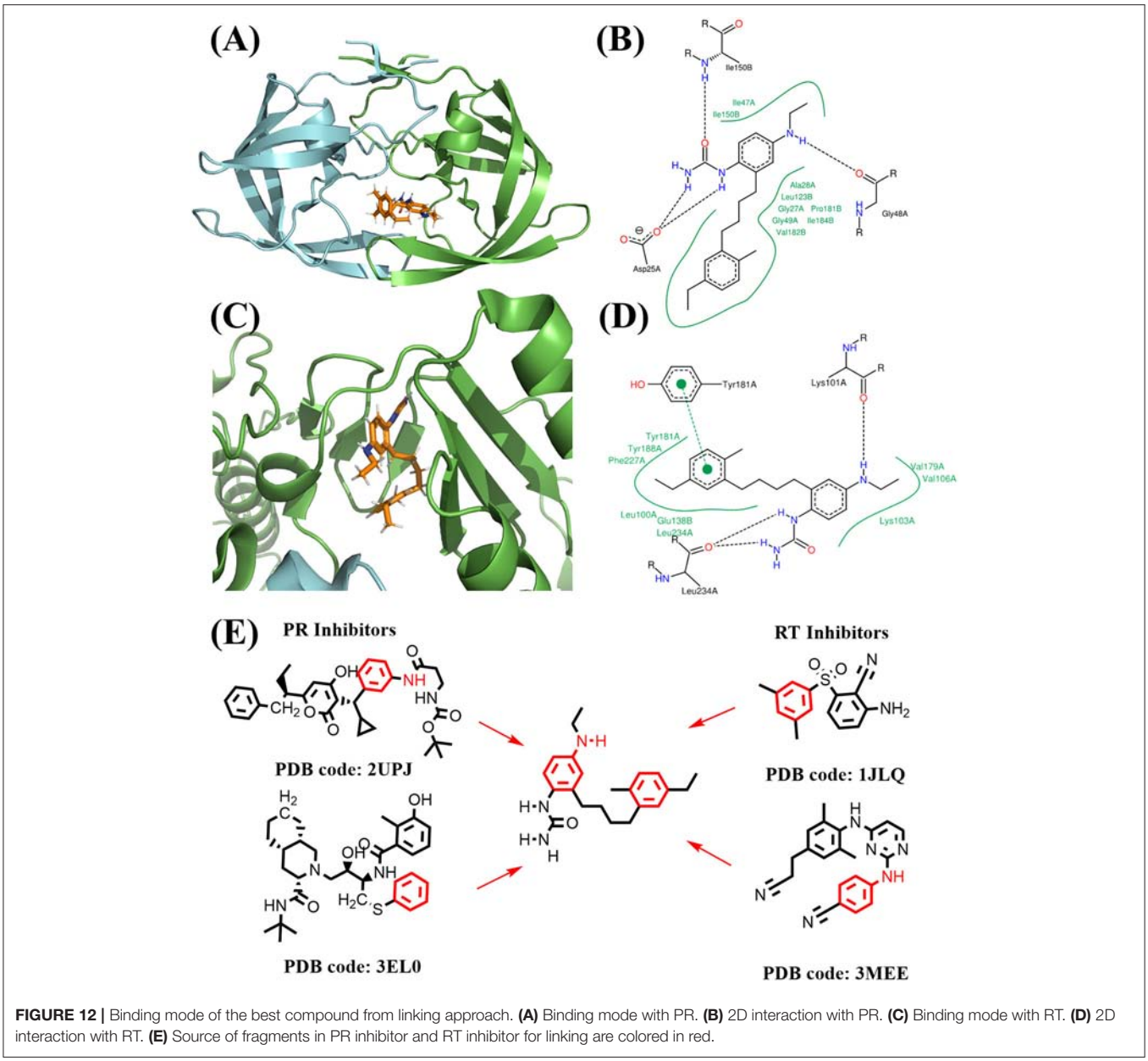


FIGURE 12 | Binding mode of the best compound from linking approach. **(A)** Binding mode with PR. **(B)** 2D interaction with PR. **(C)** Binding mode with RT. **(D)** 2D interaction with RT. **(E)** Source of fragments in PR inhibitor and RT inhibitor for linking are colored in red.

this project. Base on the data in **Table 1**, linking approach is most effective way of design high affinity ligands, and growing approach is also more effective than *de novo* approach. This is not surprising because known fragments would provide good starting points for derivation and significantly reduce searching space. The linking approach uses more known fragments which further improve its efficiency comparing with growing approach using only 1 fragment. However, the results from *de novo* approach demonstrated that this approach could achieve similar design performance to growing approach or linking approach if more computational resource is provided. As the *de novo* approach does not reply on known fragments, it would be very useful for design ligands for new targets or discover novel ligands for known targets. On the other hand, growing approach and linking approach also have their unique advantages comparing with *de novo* approach. Since compounds designed by growing or linking approach contain "validated active fragments," it would reduce the risk of "false positive," which is very common in computer-aided drug discovery. So the three strategies could be complementary in practical drug discovery projects.

CONCLUSION

In this paper, we present the first *de novo* multi-target drug design program LigBuilder V3. In addition, building ligands from scratch, LigBuilder V3 also provides the feasibility of multi-target lead optimization and multi-target fragments linking. This

program is generally applicable in rational and elegant multi-target drug design and optimization, especially for the design of concise ligands for proteins targets with large difference in binding sites. The developing version of LigBuilder V3 was successfully applied in designing COX2/LTA_4H dual-functional inhibitors with micromolar level activity. In this study, we further demonstrated the three design strategies of LigBuilder V3 with computational evaluation of designing HIV-PR and HIV-RT dual functional inhibitors. We hope the concept and LigBuilder V3 can be validated by applications from the users in the future.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

JP and LL conceived the project. YY, JP, and LL designed the experiments, analyzed the results, and wrote the manuscript. YY performed the experiments.

FUNDING

This work was supported in part by the Ministry of Science and Technology of China (2016YFA0502300 and 2015CB910300) and the National Natural Science Foundation of China (21633001 and 21673010).

ACKNOWLEDGMENTS

We thank Dr. Erchang Shang for feedback of experiment data in developing of LigBuilder V3, thank Dr. Daqi Yu, Dr. Yifei Qi, Dr. Changsheng Zhang, Dr. Fangjin Chen, Dr. Weilin Zhang, and Dr. Youjun Xu for providing supports in algorithm.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2020.00142/full#supplementary-material

REFERENCES

Abdolmaleki, A., Ghasemi, J. B., and Ghasemi, F. (2017). Computer aided drug design for multi-target drug design: SAR/QSAR, molecular docking and pharmacophore methods. *Curr. Drug Targets* 18, 556–575. doi: 10.2174/1389450117666160101120822

Alcaro, S., Bolognesi, M. L., García-Sosa, A. T., and Rapposelli, S. (2019). Multi-target-directed ligands (MTDL) as challenging research tools in drug discovery: from design to pharmacological evaluation. *Front. Chem.* 7:71. doi: 10.3389/fchem.2019.00071

Berman, H. M., Bourne, P. E., Westbrook, J., and Zardecki, C. (2003). *The Protein Data Bank in Protein Structure*. Boca Raton, FL: CRC Press. 394–410.

Bhatti, A. B., Usman, M., and Kandi, V. (2016). Current scenario of HIV/AIDS, treatment options, and major challenges with compliance to antiretroviral therapy. *Cureus* 8:e515. doi: 10.7759/cureus.515

Blonde, L., Jendle, J., Gross, J., Woo, V., Jiang, H., Fahrbach, J. L., et al. (2015). Once-weekly dulaglutide versus bedtime insulin glargine, both in combination with prandial insulin lispro, in patients with type 2 diabetes (AWARD-4): a randomised, open-label, phase 3, non-inferiority study. *Lancet* 385, 2057–2066. doi: 10.1016/S0140-6736(15)60936-9

Bottegoni, G., Veronesi, M., Bisignano, P., Kacker, P., Favia, A. D., and Cavalli, A. (2016). Development and application of a virtual screening protocol for the identification of multitarget fragments. *ChemMedChem* 11, 1259–1263. doi: 10.1002/cmdc.201500521

Bremermann, H. J. (1958). *The Evolution of Intelligence: The Nervous System as a Model of its Environment*. Seattle, WA: University of Washington, Department of Mathematics.

Brown, D., and Superti-Furga, G. (2003). Rediscovering the sweet spot in drug discovery. *Drug Discov. Today* 8, 1067–1077. doi: 10.1016/S1359-6446(03)02902-7

Case, D. A., Darden, T. A., Cheatham T. E. III, Simmerling, C. L., Wang, J., Duke, R. E., et al. (2012). *AMBER 12*. San Francisco, CA: University of California.

Cavalli, A., Bolognesi, M. L., Minarini, A., Rosini, M., Tumiatti, V., Recanatini, M., et al. (2008). Multi-target-directed ligands to combat neurodegenerative diseases. *J. Med. Chem.* 51, 347–372. doi: 10.1021/jm7009364

Chan, A. H., Lee, W.-G., Spasov, K. A., Cisneros, J. A., Kudalkar, S. N., Petrova, Z. O., et al. (2017). Covalent inhibitors for eradication of drug-resistant HIV-1 reverse transcriptase: from design to protein crystallography. *Proc. Natl. Acad. Sci. U.S.A.* 114, 9725–9730. doi: 10.1073/pnas.1711463114

Chen, Z., Wu, Y., Liu, Y., Yang, S., Chen, Y., and Lai, L. (2011). Discovery of dual target inhibitors against cyclooxygenases and leukotriene A4 hydrolyase. *J. Med. Chem.* 54, 3650–3660. doi: 10.1021/jm200063s

Durrant, J. D., Friedman, A. J., and Mccammon, J. A. (2011). CrystalDock: a novel approach to fragment-based drug design. *J. Chem. Inf. Model.* 51, 2573–2580. doi: 10.1021/ci200357y

Eaton, B. E., Gold, L., and Zichi, D. A. (1995). Let's get specific: the relationship between specificity and affinity. *Chem. Biol.* 2, 633–638. doi: 10.1016/1074-5521(95)90023-3

Eisen, S. A., Miller, D. K., Woodward, R. S., Spitznagel, E., and Przybeck, T. R. (1990). The effect of prescribed daily dose frequency on patient medication compliance. *Arch. Intern. Med.* 150, 1881–1884. doi: 10.1001/archinte.1990.00390200073014

Fraser, A. S. (1957). Simulation of genetic systems by automatic digital computers ii. Effects of linkage on rates of advance under selection. *Aust. J. Biol. Sci.* 10, 492–500. doi: 10.1071/BI9570492

Geppert, H., Vogt, M., and Bajorath, J. (2010). Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* 50, 205–216. doi: 10.1021/ci900419k

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91–100. doi: 10.1038/nature11245

Giles, T. D., Weber, M. A., Basile, J., Gradman, A. H., Bharucha, D. B., Chen, W., et al. (2014). Efficacy and safety of nebivolol and valsartan as fixed-dose combination in hypertension: a randomised, multicentre study. *Lancet* 383, 1889–1898. doi: 10.1016/S0140-6736(14)60614-0

González-Díaz, H., Prado-Prado, F. J., Santana, L., and Uriarte, E. (2006). Unify QSAR approach to antimicrobials. part 1: predicting antifungal activity against different species. *Bioorg Med Chem.* 14, 5973–5980. doi: 10.1016/j.bmc.2006.05.018

He, B., Lu, C., Zheng, G., He, X., Wang, M., Chen, G., et al. (2016). Combination therapeutics in complex diseases. *J. Cell. Mol. Med.* 20, 2231–2240. doi: 10.1111/jcmm.12930

Hidaka, K., Kimura, T., Abdel-Rahman, H. M., Nguyen, J.-T., Mcdaniel, K. F., Kohlbrenner, W. E., et al. (2009). Small-sized human immunodeficiency virus type-1 protease inhibitors containing allophenylnorstatine to explore the $S2'$ pocket. *J. Med. Chem.* 52, 7604–7617. doi: 10.1021/jm9005115

Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press.

Hopkins, A. L., Groom, C. R., and Alex, A. (2004). Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* 9, 430–431. doi: 10.1016/S1359-6446(04)03069-7

Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. (2011). Principles of early drug discovery. *Br. J. Pharmacol.* 162, 1239–1249. doi: 10.1111/j.1476-5381.2010.01127.x

Jhoti, H., Singh, O. M., Weir, M. P., Cooke, R., Murray-Rust, P., and Wonacott, A. (1994). X-ray crystallographic studies of a series of penicillin-derived asymmetric inhibitors of HIV-1 protease. *Biochemistry* 33, 8417–8427. doi: 10.1021/bi00194a005

Kamb, A., Wee, S., and Lengauer, C. (2007). Opinion - why is cancer drug discovery so difficult? *Nat. Rev. Drug Discov.* 6, 115–120. doi: 10.1038/nrd2155

King, R. W., Klabe, R. M., Reid, C. D., and Erickson-Viitanen, S. K. (2002). Potency of nonnucleoside reverse transcriptase inhibitors (NNRTIs) used in combination with other human immunodeficiency virus NNRTIs, NRTIs, or protease inhibitors. *Antimicrob. Agents Chemother.* 46, 1640–1646. doi: 10.1128/AAC.46.6.1640-1646.2002

Korcsmaros, T., Szalay, M. S., Csaba, B., Kovacs, I. A., and Csermely, P. (2007). How to design multi-target drugs: target search options in cellular networks. *Expert Opin. Drug Discov.* 2, 799–808. doi: 10.1517/17460441.2.6.799

KoŽíšek, M., Lepšík, M., Grantz Šašková, K., Brynda, J., Konvalinka, J., and Rezáčová, P. (2014). Thermodynamic and structural analysis of HIV protease resistance to darunavir–analysis of heavily mutated patient-derived HIV-1 proteases. *FEBS J.* 281, 1834–1847. doi: 10.1111/febs.12743

Kumar, A., and Sharma, A. (2018). "Computational modeling of multi-target-directed inhibitors against Alzheimer's disease," in *Computational Modelling Drugs Against Alzheimer's Disease*, ed K. Roy (New York, NY: Springer), 533–571.

Kuroda, D. G., Bauman, J. D., Challa, J. R., Patel, D., Troxler, T., Das, K., et al. (2013). Snapshot of the equilibrium dynamics of a drug bound to HIV-1 reverse transcriptase. *Nat. Chem.* 5:174. doi: 10.1038/nchem.1559

Lepailleur, A., Freret, T., Lemaitre, S., Boulouard, M., Dauphin, F. O., Hinschberger, A., et al. (2014). Dual histamine H3R/serotonin 5-HT4R ligands with antiamnesic properties: pharmacophore-based virtual screening and polypharmacology. *J. Chem. Inf. Model.* 54, 1773–1784. doi: 10.1021/ci500157n

Lu, D.-Y., Wu, H.-Y., Yarla, N. S., Xu, B., Ding, J., and Lu, T.-R. (2018). HAART in HIV/AIDS treatments: future trends. *Infect. Disord. Drug Targets* 18, 15–22. doi: 10.2174/1871526517666170505122800

Mashalidis, E. H., Sled,z, P., Lang, S., and Abell, C. (2013). A three-stage biophysical screening cascade for fragment-based drug discovery. *Nat. Protoc.* 8, 2309–2324. doi: 10.1038/nprot.2013.130

Matsumoto, H., Hamawaki, T., Ota, H., Kimura, T., Goto, T., Sano, K., et al. (2000). 'Double-drugs'—A new class of prodrug form of an HIV protease inhibitor conjugated with a reverse transcriptase inhibitor by a spontaneously cleavable linker. *Bioorg. Med. Chem. Lett.* 10, 1227–1231. doi: 10.1016/S0960-894X(00)00202-X

Morphy, J. R., and Harris, C. J. (2012). Designing multi-target drugs. *R. Soc. Chem.* doi: 10.1039/9781849734912

Morphy, R., Kay, C., and Rankovic, Z. (2004). From magic bullets to designed multiple ligands. *Drug Discov. Today* 9, 641–651. doi: 10.1016/S1359-6446(04)03163-0

Morphy, R., and Rankovic, Z. (2005). Designed multiple ligands. An emerging drug discovery paradigm. *J. Med. Chem.* 48, 6523–6543. doi: 10.1021/jm058225d

Morphy, R., and Rankovic, Z. (2006). The physicochemical challenges of designing multiple ligands. *J. Med. Chem.* 49, 4961–4970. doi: 10.1021/jm060315

Morphy, R., and Rankovic, Z. (2007). Fragments, network biology and designing multiple ligands. *Drug Discov. Today* 12, 156–160. doi: 10.1016/j.drudis.2006.12.006

Morphy, R., and Rankovic, Z. (2009). Designing multiple ligands-medicinal chemistry strategies and challenges. *Curr. Pharm. Des.* 15, 587–600. doi: 10.2174/138161209787315594

Morphy, R., and Rankovic, Z. (2010). "Design of multitarget ligands," in *Lead Generation Approaches in Drug Discovery*, eds Z. Rankovic and R. Morphy (Hoboken, NJ: John Wiley & Sons), 141–164.

Ramsay, R. R., Popovic-Nikolic, M. R., Nikolic, K., Uliassi, E., and Bolognesi, M. L. (2018). A perspective on multi-target drug discovery and design for complex diseases. *Clin. Transl. Med.* 7:3. doi: 10.1186/s40169-017-0181-2

Rastelli, G., Del Rio, A., Degliesposti, G., and Sgobba, M. (2010). Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *J. Comput. Chem.* 31, 797–810. doi: 10.1002/jcc.21372

Reddy, A. S., and Zhang, S. (2013). Polypharmacology: drug discovery for the future. *Expert Rev. Clin. Pharmacol.* 6, 41–47. doi: 10.1586/ecp.12.74

Ren, J., Milton, J., Weaver, K. L., Short, S. A., Stuart, D. I., and Stammers, D. K. (2000). Structural basis for the resilience of efavirenz (DMP-266) to drug resistance mutations in HIV-1 reverse transcriptase. *Structure* 8, 1089–1094. doi: 10.1016/S0969-2126(00)00513-X

Scannell, J. W., Blanckley, A., Boldon, H., and Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* 11, 191. doi: 10.1038/nrd3681

Schrodinger, Llc, (2010). *The PyMOL Molecular Graphics System*, Version 1.3r1.

Shang, E., Yuan, Y., Chen, X., Liu, Y., Pei, J., and Lai, L. (2014). *De novo* design of multitarget ligands with an iterative fragment-growing strategy. *J. Chem. Inf. Model.* 54, 1235–1241. doi: 10.1021/ci500021v

Shen, Y., Altman, M. D., Ali, A., Nalam, M. N., Cao, H., Rana, T. M., et al. (2013). Testing the substrate-envelope hypothesis with designed pairs of compounds. *ACS Chem. Biol.* 8, 2433–2441. doi: 10.1021/cb400468c

Song, A., Yu, H., Wang, C., Zhu, X., Liu, K., and Ma, X. (2015). Novel dual small-molecule HIV inhibitors: scaffolds and discovery strategies. *Curr. Pharm. Des.* 21, 950–962. doi: 10.2174/1381612820666140929095102

Stierand, K., and Rarey, M. (2010). PoseView–molecular interaction patterns at a glance. *J. Cheminform.* 2:P50. doi: 10.1186/1758-2946-2-S1-P50

Sun, X., Fan, N., Xu, W., Sun, Y., Xie, X., Guo, Y., et al. (2016). Design, synthesis and biological evaluation of caffeoyl benzanilides as dual inhibitors of HIV integrase and CCR5. *MedChemComm* 7, 2028–2032. doi: 10.1039/C6MD00311G

Von Hoff, D. D., Goldstein, D., and Renschler, M. F. (2014). Albumin-bound paclitaxel plus gemcitabine in pancreatic cancer. *N. Engl. J. Med.* 370, 479–480. doi: 10.1056/NEJMc1314761

Wang, T., Wu, M.-B., Chen, Z.-J., Chen, H., Lin, J.-P., and Yang, L.-R. (2015). Fragment-based drug discovery and molecular docking in drug design. *Curr. Pharm. Biotechnol.* 16, 11–25. doi: 10.2174/1389201015666141122204532

Wei, D., Jiang, X., Zhou, L., Chen, J., Chen, Z., He, C., et al. (2008). Discovery of multitarget inhibitors by combining molecular docking with common pharmacophore matching. *J. Med. Chem.* 51, 7882–7888. doi: 10.1021/jm8010096

Whitley, D. (1994). A genetic algorithm tutorial. *Stat. Comput.* 4, 65–85. doi: 10.1007/BF00175354

Wu, Y., He, C., Gao, Y., He, S., Liu, Y., and Lai, L. (2012). Dynamic modeling of human 5-lipoxygenase–inhibitor interactions helps to discover novel inhibitors. *J. Med. Chem.* 55, 2597–2605. doi: 10.1021/jm201497k

Xu, Y., Wang, S., Hu, Q., Gao, S., Ma, X., Zhang, W., et al. (2018). CavityPlus: a web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. *Nucleic Acids Res.* 46, W374–W379. doi: 10.1093/nar/gky380

Yuan, Y., Pei, J., and Lai, L. (2011). LigBuilder 2: a practical *de novo* drug design approach. *J. Chem. Inf. Model.* 51, 1083–1091. doi: 10.1021/ci100350u

Yuan, Y., Pei, J., and Lai, L. (2013). Binding site detection and druggability prediction of protein targets for structure-based drug design. *Curr. Pharm. Des.* 19, 2326–2333. doi: 10.2174/13816128113191320019

Zhan, P., and Liu, X. Y. (2009). Designed multiple ligands: an emerging Anti-HIV drug discovery paradigm. *Curr. Pharm. Des.* 15, 1893–1917. doi: 10.2174/138161209788453266

Zhang, W., Pei, J., and Lai, L. (2017). Computational multitarget drug design. *J. Chem. Inf. Model.* 57, 403–412. doi: 10.1021/acs.jcim.6b00491

Zhang, W., Yuan, Y., Pei, J., and Lai, L. (2015). "CAVITY: mapping the druggable binding site," in *Part of the Series Methods in Pharmacology and Toxicology*, ed W. Zhang (New York, NY: Humana Press), 111–132.

Zimmermann, G. R., Lehar, J., and Keith, C. T. (2007). Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug Discov. Today* 12, 34–42. doi: 10.1016/j.drudis.2006.11.008

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Yuan, Pei and Lai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 13 March 2020
doi: 10.3389/fchem.2020.00176

A Definition of "Multitargeticity": Identifying Potential Multitarget and Selective Ligands Through a Vector Analysis

Juan Francisco Sánchez-Tejeda[1], *Juan F. Sánchez-Ruiz*[2], *Juan Rodrigo Salazar*[1] *and Marco A. Loza-Mejía*[1]*

[1] *Facultad de Ciencias Químicas, Universidad La Salle, Mexico City, Mexico,* [2] *Ciencia y Estrategia S.A. de C.V., Mexico City, Mexico*

The design of multitarget drugs is an essential area of research in Medicinal Chemistry since they have been proposed as potential therapeutics for the management of complex diseases. However, defining a multitarget drug is not an easy task. In this work, we propose a vector analysis for measuring and defining "multitargeticity." We developed terms, such as order and force of a ligand, to finally reach two parameters: multitarget indexes 1 and 2. The combination of these two indexes allows discrimination of multitarget drugs. Several training sets were constructed to test the usefulness of the indexes: an experimental training set, with real affinities, a docking training set, within theoretical values, and an extensive database training set. The indexes proved to be useful, as they were used independently *in silico* and experimental data, identifying actual multitarget compounds and even selective ligands in most of the training sets. We then applied these indexes to evaluate a virtual library of potential ligands for targets related to multiple sclerosis, identifying 10 compounds that are likely leads for the development of multitarget drugs based on their *in silico* behavior. With this work, a new milestone is made in the way of defining multitargeticity and in drug design.

Keywords: multitarget drugs, drug discovery, drug-design, multitarget index, multiple sclerosis, polypharmacology

OPEN ACCESS

Edited by:
Simone Brogi,
University of Pisa, Italy

Reviewed by:
Andrea Armirotti,
Italian Institute of Technology, Italy
Giulio Rastelli,
University of Modena and Reggio Emilia, Italy
Orazio Nicolotti,
University of Bari Aldo Moro, Italy

***Correspondence:**
Marco A. Loza-Mejía
marcoantonio.loza@lasalle.mx

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *03 October 2019*
Accepted: *26 February 2020*
Published: *13 March 2020*

Citation:
Sánchez-Tejeda JF, Sánchez-Ruiz JF, Salazar JR and Loza-Mejía MA (2020) A Definition of "Multitargeticity": Identifying Potential Multitarget and Selective Ligands Through a Vector Analysis. Front. Chem. 8:176. doi: 10.3389/fchem.2020.00176

INTRODUCTION

In the field of polypharmacology, combinatorial therapies and multitarget drugs are the main alternatives for dealing with complex diseases. The first one consists of a combination of multiple single-targeted drugs. On the other hand, multitarget drugs are molecules with the ability to act on different targets at the same time. Designing multitarget drugs is a problematic task; however, it solves several concerns that are seen in combinatorial therapies, such as complex therapeutic regimens, difficulty in including numerous drugs in a single formulation and drug interactions at the different pharmacokinetics levels: absorption, distribution, metabolism, and elimination (Rosini, 2014). In the last two decades, the number of multitarget drugs on the market has been rising. From 2015 to 2017, 21% of the drugs approved by the Food and Drug Administration (FDA) were multitarget drugs (MTD), primarily antineoplastic agents (Ramsay et al., 2018). This trend may indicate that the number of multitarget drugs will continue to rise since they present advantages over single-target drugs. For example, MTDs have higher *in vivo* efficacy, and several

in silico methods and strategies for designing them are currently being developed (Zhang et al., 2017). A common strategy is to combine two pharmacophores in the same molecule or partially overlap them, allowing binding to two or more targets (Talevi, 2015).

Binding to two or more targets at the same time offers the possibility of treating multifactorial diseases. Neurodegenerative diseases are a potential field for multitarget drugs. For example, ladostigil is a dual cholinesterase–monoamine oxidase-B (MAO-B) inhibitor currently being researched for the treatment of Alzheimer's disease and other neurodegenerative diseases (van der Schyf, 2011). Cancer is another relatively emergent field for multitarget drugs, mainly as more druggable targets are being discovered. The use of multitarget drugs is promising as it lowers the possibility of the disease to evolve into a drug-resistant phenotype (Xie and Bourne, 2015). Currently, several anti-cancer drugs are considered multitarget drugs since they inhibit two or more kinases or receptors (Lu et al., 2012). Another example is in the field of microbiology, in which dual ligands can be used to treat tuberculosis. This dual mechanism of action is useful in treating multidrug-resistant *Mycobacterium tuberculosis* (Chiarelli et al., 2018).

One of the limitations that multitarget drug design faces is data analysis. In some cases, the number of targets or compounds being analyzed can be high. In PubChem, 71,303 molecules have been identified as ligands that have two or more biological targets, and more than 30,000 ligands were found to be active against more than 400 targets (Hu et al., 2014). Quantifying and defining "multitargeticity" may be useful for analyzing these datasets. Additionally, multitarget metrics could help multitarget drug design by providing comparable and workable parameters for drugs and ligands.

To the best of our knowledge, there is no current measurement of "multitargeticity," i.e., how multitargeted a ligand is. Construction of a multitarget parameter should not be based only on the simple average of the *in silico* or experimental data; for example, highly selective ligands to a single target would appear as multitarget drugs, since the average is a measure sensitive to extreme values. With this in mind, our research group suggested the use of a virtual multitarget parameter, which consisted of a weighted average of the docking scores of potential biopesticides (Loza-Mejía et al., 2018). This analysis proved useful for comparing a ligand's "multitargeticity." However, a more rigid index may help even further in multitarget drug design.

Originally, this project started with the purpose of designing dual ligands. We designed 211 ligands, and we wanted a parameter that could summarize or identify the ligands that had the most potency toward the two targets (the nature of the ligands and the targets will be explained later). To analyze the data, we plotted the docking score of the ligands of one target against the docking score of the second one. In this plot (**Figure 1**), a ligand can be described by the coordinates or docking scores of both targets. Since they are coordinates, the ligand describes an *arrow* or vector, starting from the origin. The angle described by the vector is the selectivity; in fact, the formula of the tangent is the formula for selectivity (Equation 1).

$$\tan \alpha = \frac{Target\ \#2\ \ ligand's\ affinity}{Target\ \#1\ \ ligand's\ affinity} \quad (1)$$

Moreover, the magnitude of the vector is likely related to how potent the ligand is. Greater affinities reflect greater magnitudes. With the graph (**Figure 1**), there is a sense of what a multitarget drug would be: one that equally distributes its magnitude among the two targets. In other words, a ligand that had the same affinity for both targets. Measuring the multitargeticity of a dual ligand can be as simple as obtaining the tangent of the angle (Equation 1). If the tangent equals 1 ($\alpha = 45°$), then mathematically, the ligand attacks both targets with the same "strength." However, this interpretation was meant for more targets. In these scenarios, a single parameter for defining multitargeticity would not suffice because more angles are involved.

A radial plot was considered to extend the analysis to further dimensions (**Figure 1**). The central idea was kept: **a** multitarget ligand would equally distribute its *strength* among all the targets. In the radial representation, a multitarget drug would appear as a regular polygon, and the area could be considered the magnitude or strength of the ligand. Multitargeticity could be defined by the similarity between the figure described by the ligand and a regular polygon. We calculated some parameters that could be used to describe this relationship and therefore give a quantitative definition of multitargeticity. However, the area and shape are sensible to the order in which the targets are analyzed. A different order would give a completely different value, as shown in **Figure 1**.

The solution was to treat ligands as vectors and extend the analysis to further dimensions, even if it cannot be visualized. With two targets, an ideal multitarget drug is a vector that makes a perfect square in a 2D plane. In 3 dimensions, a cube would be the shape of an ideal multitarget ligand. Therefore, a hypercube is the central analysis of this interpretation. In contrast, a different distribution of affinities would produce the shape of a rectangle, rectangular cuboid, or hyperrectangle, depending on the number of targets analyzed. Measuring the similarity between the hyperrectangle and a hypercube is, in fact, a measure of multitargeticity, which is only a mathematical definition of how much a ligand equally distributes its *strength* among all the targets. With this analysis, another step was made toward defining and measuring multitargeticity.

One of the several complex diseases that can be treated with a multitarget drug approach is multiple sclerosis (MS). MS is a disease that affects the central nervous system. Currently, MS is the neurological disease with the highest incidence; in 2013, ∼2.3 million people were estimated to have MS (Browne et al., 2014). The pathophysiology of the disease is based on the demyelination of axons, primarily due to the loss of oligodendrocytes, cells responsible for maintaining the myelin sheaths around them (Dobson and Giovannoni, 2019). Although the exact origin of MS is not known, it is well-established that it causes damage to the myelin sheath. Depending on the type of MS and the damage

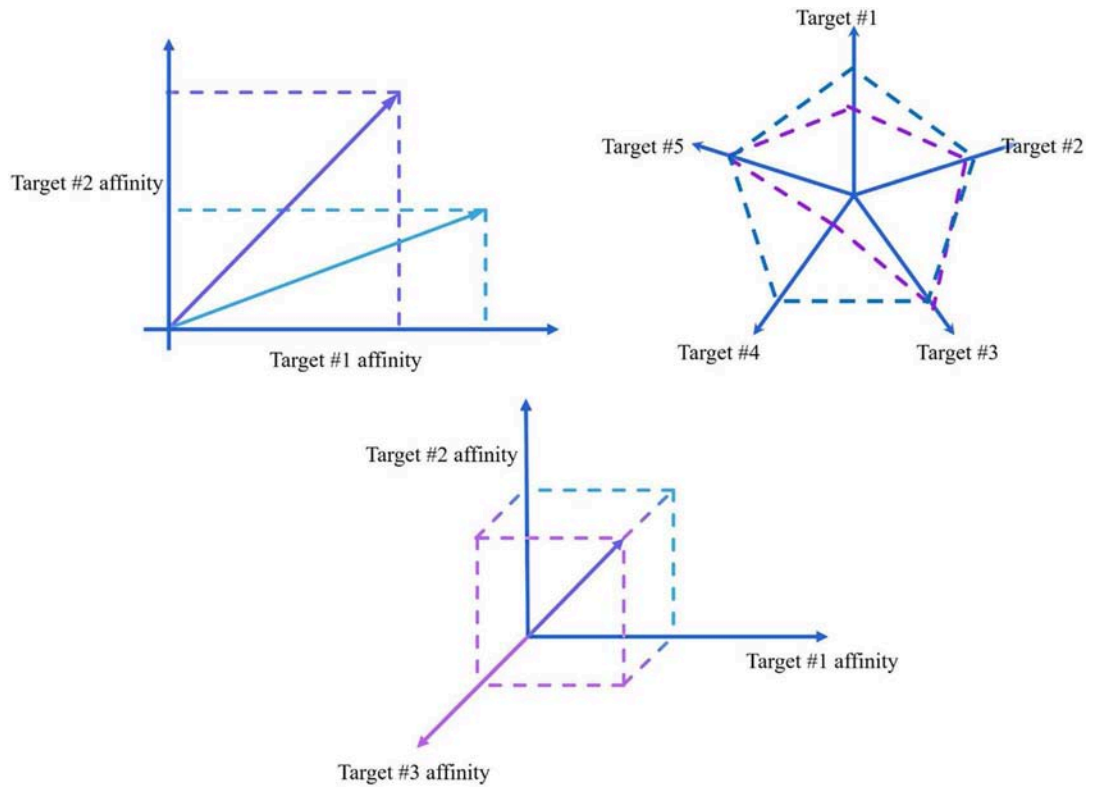


FIGURE 1 | Plotting the ligands' affinities in a dispersion plot describe vectors in a 2D space. For further targets, in a radial plot, an ideal multitarget would describe a regular polygon. Measuring the area of these polygons is influenced by the order in which the targets are plotted. A hypercube is not affected by ordering and therefore makes an ideal representation of what a multitarget ligand would be.

present, the slow transmission of electronic impulses may lead to axon loss, consequently damaging the optic nerve and leading to degeneration of vision, weakness, atrophy, and muscular rigidity, coordination and balance failures, recurrent fatigue, dysphagia, depression, and anxiety, among other symptoms (Huang et al., 2017). PAR-1 and KLK-6 are two molecules of biological interest as possible targets for MS treatment due to their role in oligodendrogliopathy and autoimmune response (Burda et al., 2013). PAR-1 is a protease-activated receptor involved in coagulation, angiogenesis, proinflammatory responses, oligodendrocyte death, and myelination (Macfarlane et al., 2001; Yoon et al., 2015; Pan et al., 2016; Lee et al., 2017). Antagonists of PAR-1 have been shown to reduce the symptoms of experimental autoimmune encephalomyelitis (EAE), which is the most studied animal model for MS (Kim et al., 2015). Kallikrein 6 (KLK-6) is the most abundant serine protease in the central nervous system (CNS) and has proteolytic activity against myelin basic protein (MBP) and amyloid precursor protein, which are part of the myelin sheath and are involved in myelination (Burda et al., 2013; Yoon and Scarisbrick, 2016). KLK-6 is also involved in T-cell survival and apoptotic signalizing (Scarisbrick et al., 2011).

Additionally, recently, a new drug for secondary progressive MS was approved by the FDA: siponimod, which is sold under the trade name Mayzent®. Siponimod is a dual drug itself: it binds to sphingosine-1-phosphate receptor 1 and 5 (S1PR1 and S1PR5) (O'Sullivan et al., 2016). Siponimod reduces oligodendrocyte death and demyelination, acting as an effective neuroprotective agent (Behrangi et al., 2019). S1PR1 is involved in regulating the inflammatory response and therefore is of interest as a third biological target (Chi and Nicol, 2010).

In this work, we present the construction of multitarget indexes as parameters that can define multitargeticity, their evaluation on several datasets, and their use in identifying potential multitarget ligands for PAR-1, KLK-6, and S1PR1.

MATERIALS AND METHODS

Construction of an Experimental Training Set: Multi-Kinase Ligands

We selected 10 known FDA-approved drugs labeled multikinase-directed drugs as models of multitarget drugs (Li et al., 2016). Additionally, two non-multikinase drugs were included as negative controls. The binding affinity (defined in terms of K_i) of each drug to its target was searched in the Binding DB (Gilson et al., 2016). The targets of each ligand were selected according to the FDA approved information. For the negative controls, tyrosine-protein kinase ABL1 was included in the analysis. The K_i was transformed into pM units and linearized. The objective

of this analysis was to test if the multitarget index could correctly classify drugs using experimental values. The analysis was done, as stated in section Construction of the Multitarget Index.

Construction of a Docking Training Set: Multi-Kinase Ligands

We selected 10 known FDA-approved drugs labeled multikinase-directed drugs as models of multitarget drugs (Li et al., 2016). The full list of the drugs we considered can be found in the **Supplementary Information**. Three non-multikinase drugs were also included as negative controls. The 13 ligands were docked on the tyrosine-protein kinase KIT (PDB id: 4HVS), vascular endothelial growth factor receptor 2 (PDB id: 3VO3) and platelet-derived growth factor receptor beta (PDB id: 1SHA). The docking studies were carried out in Molegro Virtual Docker version 6.0 using the standard protocol suggested by the manufacturer. All waters, cofactors, and non-active ligands were removed from the workspace. MolDock optimizer was used as the running algorithm with 25 runs per ligand. A sphere with a radius of 15 Å was constructed around the active sites of the three proteins and selected as the search site. The poses with the lowest MolDock score were used for further analysis. The objective of this analysis was to determine if the multitarget index could correctly classify drugs using theoretical values. The data were processed, as stated in section Construction of the Multitarget Index.

Construction of an Experimental Evaluation Set: DUD

We downloaded the database of DUD (Directory of Useful Decoys; Huang et al., 2006), containing the energy scores of nearly 98258 molecules against 40 targets; some targets had a smaller number of calculated energies but were included in the analysis, as this would challenge the indexes. The package was cleaned so that only the negative energies of each ligand were analyzed. The objective of this analysis was to determine if the multitarget index could filter an extensive database. The data were processed, as stated in section Construction of the Multitarget Index.

Virtual Library of Ligands for MS

Selection and Construction of the Virtual Library

For ligand construction, PAR-1 antagonists with demonstrated activity were searched. Vorapaxar is a commercially available platelet antiaggregant whose mechanism of action is PAR-1 antagonism; therefore, it was used as a reference ligand. F16357 and SCH79797 are molecules whose antagonism has been previously studied, and thus, they were used as starting points for the design of multitarget molecules (Manaenko et al., 2013; Readmond and Wu, 2017). Four possible scaffolds were selected for PAR-1 (**Figure 2**), of which scaffolds W and X were obtained by scaffold hopping from F16357 and vorapaxar, respectively, with the help of Mcule (Kiss et al., 2012). F16357 was used as scaffold Y, and scaffold Z was an annular modification of SCH79797.

For the selection of KLK-6 ligands, benzamidine isosteres were designed, because this compound is known to be a serine protease inhibitor (Silva et al., 2017). The selected benzamidine isosteres were aminopyridine (A), aminopyridine with carboxylic acid (B), aminopyridine with alcohol (C), 2-aminopirimidine aminoquinoline (E), aminoisoquinoline (F), aminoquinazoline (G), and benzylamine (H) (**Figure 3**). Currently, there are no

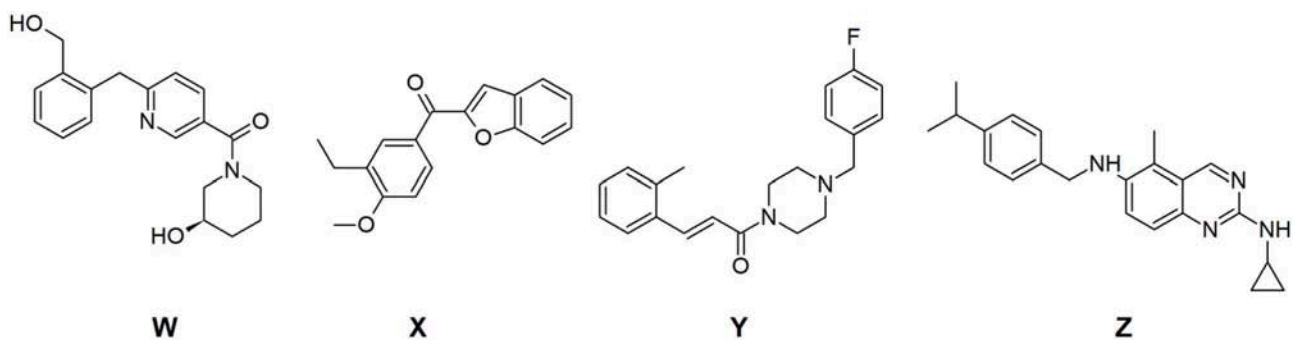


FIGURE 2 | The four scaffolds used for ligand construction.

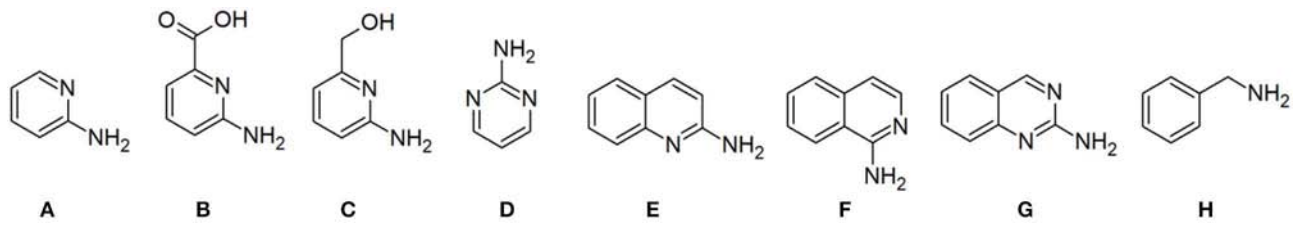


FIGURE 3 | Proposed isosteres that may interact in the active site of KLK-6.

commercially available drugs whose mechanism of action is selective inhibition of KLK-6. However, it was found that 0HM, a benzylamine derivative, was previously determined as a compound with high binding energy to KLK-6 and thus was used as a reference ligand for this enzyme (Liang et al., 2012).

Finally, 211 compounds were constructed from a combination of both types of ligands with the help of Marvin Sketch 16.2.22.0 and saved in *.smiles format. The three-dimensional geometry was optimized with Spartan '14 (1.1.4) using MMFF and HF 6-31 G*.

Molecular Docking

The crystalline structures of PAR-1 complexed with vorapaxar (PDB id: 3VW7), human S1PR1(PDB id: 3V2Y) and human KLK-6 with 0HM (PDB id: 4D8N) were downloaded (Bernett et al., 2002; Hanson et al., 2012; Liang et al., 2012; Zhang et al., 2012). The water molecules and co-crystallized ligands were removed from the work area. The docking procedure was carried out in the same manner used for multikinase ligands.

Cheminformatics Analysis

The smiles codes of the 211 ligands were placed in admetSAR in order to predict some of their pharmacokinetic properties (Cheng et al., 2012). The following probability scores were obtained: permeability of the blood-brain barrier (BBB), human intestinal absorption (HIA), glycoprotein P substrate (PGP-substrate), carcinogen, acute oral toxicity (AOT) and inhibition of hERG (human Ether-à-go-go-Related Gene). A coefficient of +1 was assigned to all values that fulfilled the following conditions: BBB +, HAI +, PGP-non-substrate, non-carcinogen, AOT III, or IV and weak hERG inhibitor. Otherwise, a negative coefficient was assigned in such a way that the desirable properties were considered positive. With these coefficients, an average of the chemoinformatic properties was calculated, which was called Chemoinformatic Score (CIS).

Construction of the Multitarget Index

Vector analysis, mentioned in the introduction, was used as the mathematical basis for the index construction. Besides, vector analysis allowed new interpretations of ligands, concepts, and parameters that may have a significant impact on multitarget drug design.

Order of a Ligand

The core idea of the index is to interpret ligands as vectors. The theoretical affinity or score for a target may be interpreted as a coordinate within this vector. This interpretation treats targets as independent variables that are orthogonal to each other. The ligand (L) is then defined as follows:

$$\vec{L} = (a_1, a_2, \ldots, a_i) \tag{2}$$

where a_i is the affinity for each target. The usefulness is that the number of targets is now coded as the number of coordinates or dimensions. Therefore, the order of a ligand (n) relates to the number of targets being tested: a multitarget of order n.

Force of a Ligand

This parameter corresponds to the norm of the vector, which is a metric that combines all the affinities of the ligands into a single value. It is generally understood as magnitude, meaning that greater values correspond to ligands whose particular affinities are large. It is a useful parameter when trying to compare combined affinities. However, this metric is also sensible to extreme values. The force of each ligand (F) was calculated in the following way:

$$F = ||\vec{L}|| = \sqrt{{a_1}^2 + {a_2}^2 + \ldots + {a_n}^2} \tag{3}$$

Plotting the ligand can enhance the interpretation, as seen in **Figure 1**. However, for more than three coordinates, representations must be truncated into a radial web for better visualization.

Binding Capacity and Total Multitarget Capacity

Because each coordinate is a vector, the cross product of all the targets will give a new vector. This new vector is an indirect measure of the binding capacity of a ligand, which geometrically corresponds to an nth volume (Equation 4). This metric is more sensible than the force because it is a multiplication of affinities, and considerable differences between affinities have greater repercussions. This operation is the same as calculating the geometric mean. We interpreted the metric as the binding capacity, a measure of a ligand's tendency to bind to more targets. Higher binding capacity means it can bind efficiently to more targets.

$$Bc = \text{geometric mean} = \sqrt[n]{\prod_{i=1}^{n} a_n} \tag{4}$$

The average is interpreted as a ligand's affinity to all the targets. The average simulates a drug that has equal affinities for all the proteins. Therefore, the average (μ) is defined as the total multitarget capacity of the ligand (MTc) (Equation 5). It is only a capacity since it is an idealized value.

$$MTc = \mu = \frac{1}{n}\sum_{i=n}^{n} a_i \tag{5}$$

Finally, the quotient of the binding capacity and the total multitarget capacity gives a proportion of how much of that multitarget capacity is being used. If the binding capacity equals the total multitarget capacity, then the ligand is a true multitarget ligand. By itself, this quotient is an index of "multitargeticity" that can be expressed in percentage for easier reading and read as "used multitarget capacity" (U_{MTc}).

$$U_{MTc} = \frac{Bc}{MTc} \tag{6}$$

Index Standardization, Definition, and Interpretation

Since the idea behind the index is to compare different ligands, it is necessary to standardize the index. The following scheme

was proposed: the ligand should be its own reference for standardization. This can be achieved through the following formula, in which the individual contribution of each affinity to the force is calculated.

$$\hat{L_s} = n\left(\frac{\vec{L}}{F}\right)^2 = \left(n\cdot\left[\frac{a_1}{F}\right]^2, n\cdot\left[\frac{a_2}{F}\right]^2, \ldots, n\cdot\left[\frac{a_n}{F}\right]^2\right) \quad (7)$$

where n is the number of targets, F is the norm or force of the ligand, and a is the affinity. This formula also standardizes the mean, or multitarget capacity (MTc) to 1, independently of the number of targets or the type of input used. For simplicity, the new coordinates were renamed "standardized affinities" ($\hat{a}$).

$$\hat{L_s} = \left(\hat{a_a}, \hat{a_2}, \ldots, \hat{a_n}\right) \quad (8)$$

This simplifies the standardized, used multitarget capacity (U_{MTc}) to a simple geometric mean ranging from 0 to 1, effectively making it an index or measurement of "multitargeticity." As in linear regression, a quadratic estimator exacerbates the value, making it ideal for a multitarget index (Equation 9).

$${}^{1st}\text{MTi} = \left(\sqrt[n]{\prod_{i=1}^{n}\hat{a}_n}\right)^2 = \left(\prod_{i=1}^{n}\hat{a}_n\right)^{2/n} \quad (9)$$

The interpretation is the one originally described in the introduction: how similar the hyperrectangle described by the ligand is similar to a hypercube. Alternatively, in a less abstract way, it is an efficiency measurement: how much "multitarget capacity" is being used.

A Second Multitarget Index

A second parameter was calculated; the standard deviation (σ). With this, another index was constructed that could measure the dispersion of the affinities: the bigger the value, the less variation among the targets. Since the standardized affinities' mean equals one, the second multitarget index is defined as follows:

$${}^{2nd}\text{MTi} = 1 - \sigma \quad (10)$$

As in ^{1st}MTi, the value can be expressed as a percentage (%). This is a more sensitive parameter that ranges from 1 to negative values. This index also encodes selectivity: smaller values, even negative, indicate more selectivity.

Defining of a Multitarget Ligand

With the two indexes, we propose the following values for classifying a ligand as a multitarget:

The ligand is multitarget if ${}^{1st}\text{MTi} \geq 0.84$ and ${}^{2nd}\text{MTi} \geq 0.60$

These values correspond to an ∼20% deviation from the mean affinity. It is worth mentioning that, although this gives a quantitative definition of multitargeticity, ligands that do not fulfill the criteria should not be discarded. These indexes quantify the dispersion and variation of the affinities and do not indicate in any way the potency.

Multitarget Potency and Selectivity

Equally low affinities will give high MT indexes values. For this reason, a final critical parameter was introduced: the multitarget potency. This value is the product of the force times, both multitarget indexes (Equation 11), which is the equivalent of calculating how much of that force is due to the multitargeticity of the ligand.

$$P_{MT} = \frac{F}{\sqrt{n}} \cdot {}^{1st}\text{MTi} \cdot {}^{2nd}\text{MTi} \quad (11)$$

We propose the multitarget potency as a metric for drug design since the highest values of potency represent a possible multitarget hit or even multitarget lead. In the same line of thought, the next parameter that we propose is an attempt to identify selective ligands. The selectivity is calculated as follows:

$$S = \frac{F}{\sqrt{n}} \cdot (1 - {}^{1st}\text{MTi})(1 - {}^{2nd}\text{MTi}) \quad (12)$$

Both parameters maintain the desired properties (higher values indicate higher potency and selectivity) and are useful for identifying possible multitarget and selective ligands. These metrics, as with any other, have their benefits and drawbacks and will be discussed further on.

RESULTS AND DISCUSSIONS

Performance of the MT Indexes in the Experimental Training Set

With the criteria set on point 2.5.6, of the 10 approved multikinase drugs analyzed (imatinib, sunitinib, dasatinib, afatinib, bosutinib, lapatinib, nintedanib, pazopanib, sorafenib, vandetanib), only afatanib was not classified as a multitarget ligand. Although it is biologically active in both of its targets, the epidermal growth factor receptor (EGFR) and the receptor tyrosine-protein kinase erbB-2 (HER2 or erbB2), it has a considerable preference over EGFR ($K_i = 0.1$ vs. $K_i = 5$ [nM]).

The two negative controls did not fulfill the criteria to be cataloged as multitarget drugs. In fact, the epidermal growth factor receptor (EGFR) is the main and only target for gefitinib and erlotinib (Wishart et al., 2018). The indexes can reliably classify and discriminate multitarget molecules in experimental values, giving strength to the analysis.

The multitarget potency, the critical parameter proposed, supports the findings, making dasatinib the most potent multitarget drug of the analyzed set ($P_{MT} = 16.4$). Sunitinib was de 2nd most potent multitarget ($P_{MT} = 14.6$); although it had better indexes, the total strength was reduced since it had more targets tested, and the affinities were not as strong as dasatinib (**Figure 4**). In contrast, gefitinib was the least potent ($P_{MT} = 14.6$), but it was also the 2nd most selective ligand of the dataset ($S_{gefitinib} = 2.7$ vs. $S_{dasatinin} = 1.3$ and $S_{sunitinib} = 0.2$), being afatinib the first ($S_{afatinib} = 2.9$).

*All P_{MT} and S values are dimensionless.

Performance of the MT Indexes in the Docking Training Set

With the criteria set on point 2.5.6, of the ten known approved multikinase drugs tested, only sorafenib was not classified as a multitarget ligand. By contrast, the three negative controls were classified as multitarget ligands according to the index. The apparent discrepancy between these results and the experimental ones is explained by considering that all the 13 ligands were docked in the same three targets, indistinctively if they were active or not, while in the experimental analysis, the preferred targets were analyzed according to each ligand. Sorafenib had larger calculated affinities than most of the ligands but was further apart from each other, which lead to it nor being classified as multitarget. The performance of sunitinib and dasatinib is observed in **Figure 5**, and the results agree with the experimental set. The only difference is that sunitinib is, in this case, the most potent multitarget ligand.

The analysis still proves useful once the ligands are arranged in order of highest to lowest multitarget potency, or the force of each ligand is compared. It is important to emphasize that the purpose of the index is not to reclassify drugs but instead to provide useful metrics for analyzing data and aiding in the drug design process, especially in the design of multitarget drugs. In this case, seven multitarget ligands would be discovered or tested

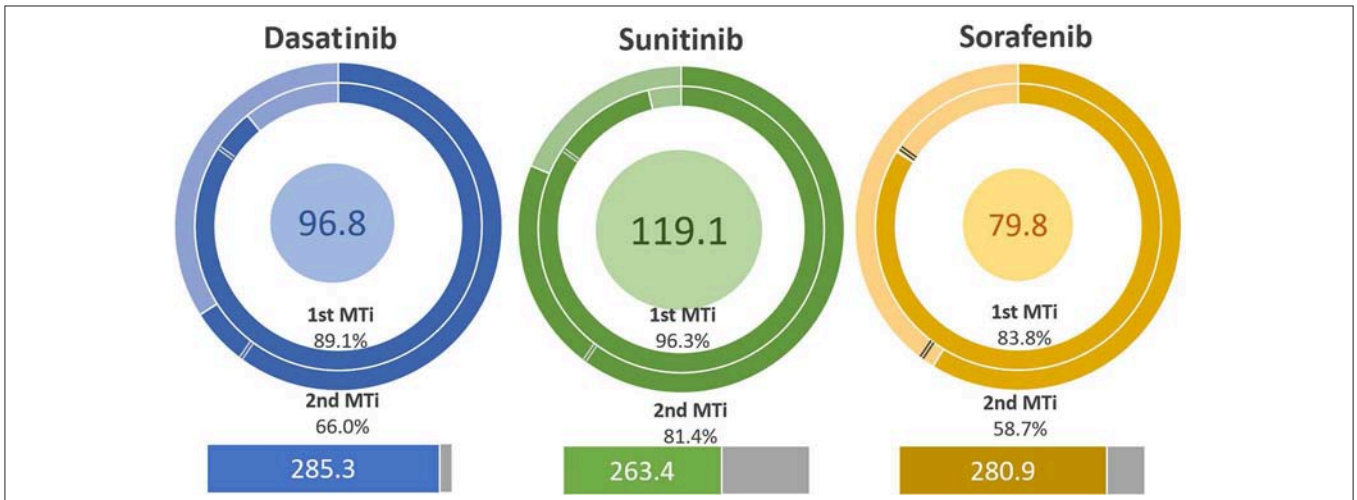


FIGURE 4 | Dasatinib had 3 targets ($n = 3$) tested and greater affinities toward those, while sunitinib had order $n = 5$ and gefitinib $n = 4$. That is why, although sunitinib has better MTi values, dasatinib has more multitarget potency (the number inside the circle). All the values are presented as % (or times 100). The threshold for considering a ligand multitarget is viewed as a cut in the circles. The inner and outer rings are the 1st and 2nd MT indexes, respectively.

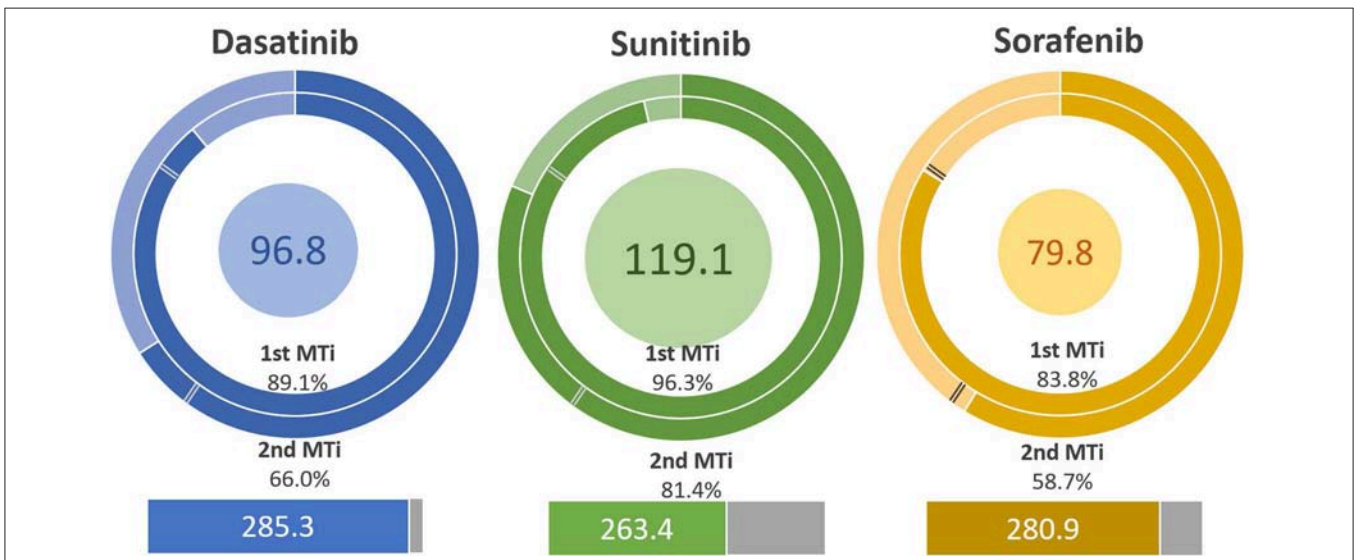


FIGURE 5 | Comparing with **Figure 4**, the multitarget tendency of sunitinib and dasatinib remains similar. The difference in the calculated affinities gives the change in the potency. Sorafenib is an example of how the MT indexes affect the strength of the ligand, since that strength is unevenly distributed. The inner circle is the MT potency. The inner and outer rings are the 1st and 2nd MT index, respectively. The inferior rectangle is the force of the ligand.

before encountering a non-multitarget drug previously classified as an MT drug. The multitarget indexes are useful when they are used with the force of the ligand. The top 7 ligands are indeed classified as multitarget ligands and are approved by the FDA as multikinase drugs, and a summary of the performance can be reviewed in **Table 1**.

Since this is an *in silico* evaluation, the scoring is affected by the computational limitations of the docking procedures. These limitations should be taken into consideration when applying the metrics previously described. These are virtual metrics and are sensible to the *in silico* scoring functions, which themselves do not reflect the *in vivo* effect. Furthermore, the indexes and metrics should be tested with experimental values to prove the strength, robustness, and validity of this classification. As mentioned above, the criteria for classifying a ligand as a multitarget can be modified, but the measure of multitargeticity persists.

TABLE 1 | Sunitinib, imatinib, and sorafenib are approved multitarget drugs by the FDA.

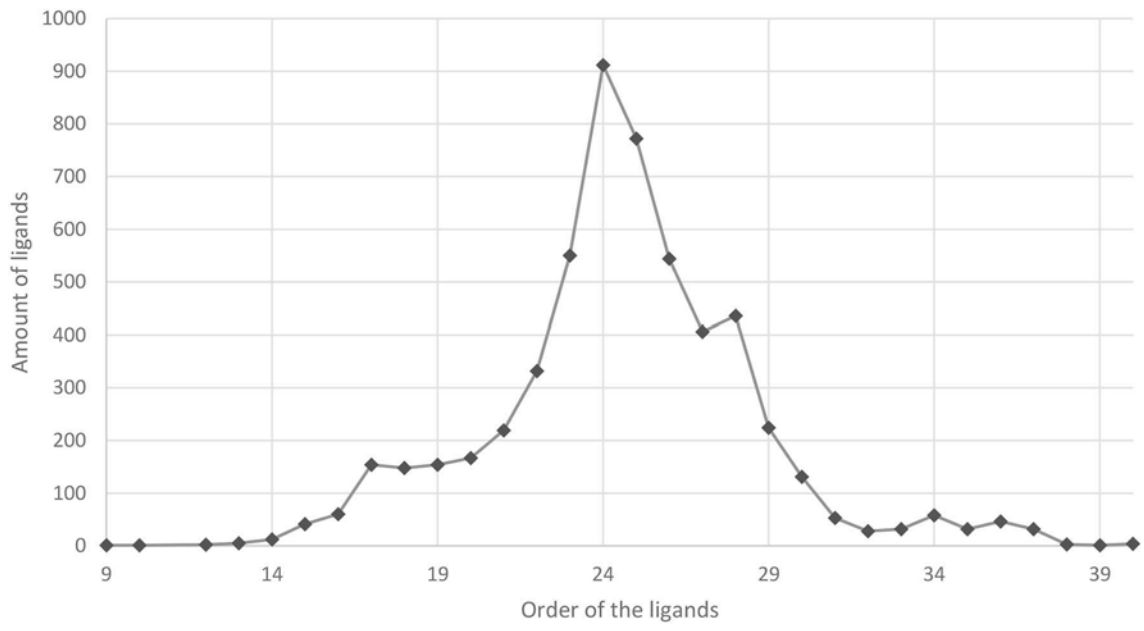
Name	4HVS MolDock score	3VO3 MolDock score	1SHA MolDock score	F	^{1st}MT index (%)	^{2nd}MT index (%)	P_{MT}
Sunitinib	−166.0	−131.5	−156.6	263.4	96.3	81.4	**119.1**
Imatinib	−193.9	−130.7	−163.2	285.1	90.4	69.1	**102.8**
Idelalisib	−138.7	−104.8	−129.5	216.8	94.7	77.9	**92.4**
Sorafenib	−201.6	−119.8	−154.8	280.9	83.8	58.7	**79.8**
Letrozole	−137.9	−98.4	−114.8	204.6	92.7	72.4	**79.3**

Sorafenib did classify as a multitarget drug. However, the multitarget potency of sunitinib and imatinib compared to letrozole and idelasib (negative controls in purple) is greater. The rest of the molecules can be found in the ***Supplementary Information****. The bold values indicate the main or critical parameter(s) being evaluated.*

DUD Database

From the 98258 ligands analyzed, 5561 molecules were found to be multitarget. This corresponds to about 5.7% of all the analyzed ligands. The orders of the ligands varied widely, ranging from 9 to 40. A total of 912 ligands were found to be multitarget of order 24. The distribution can be seen in **Graph 1**. The multitarget index filtered the ligands, meaning it does not classify every ligand as multitarget, and only a subset is eligible according to the criteria. It is important to notice that the chemical structures in this dataset were diverse. In these cases, the MT indexes gain strength as they "clean" the database and facilitating further research.

The most potent multitarget ligands were an unnamed compound of the ZINC database, which could be further tested to determine if its potential multitargeticity is only theoretical. This is the purpose of the indexes, to be useful in drug design and in identifying potential multitarget ligands. The selectivity was also used to identify the most selective ligands. In the DUD database, the third most selective ligand corresponded to hepsulfam, with the following MT indexes: 6.6% (1^{st}) and −272.1% (2^{nd}). A comparison is made in **Figure 6**. The selectivity was toward catechol-O-methyltransferase (COMT). Hepsulfam is an alkylsulfonate alkylating drug-like busulfan used in cancer therapy. COMT is a modulator of the dopaminergic and adrenergic response, and it can influence nausea and vomiting (Gan and Habib, 2016). Hepsulfam binding to COMT could explain mild nausea and vomiting seen in clinical trials, contrasting serious nausea present in other types of anticancer drugs (Ravdin et al., 1991).



Graph 1 | Distribution of how many orders of ligands were among the 5561 ligands identifies as multitarget.

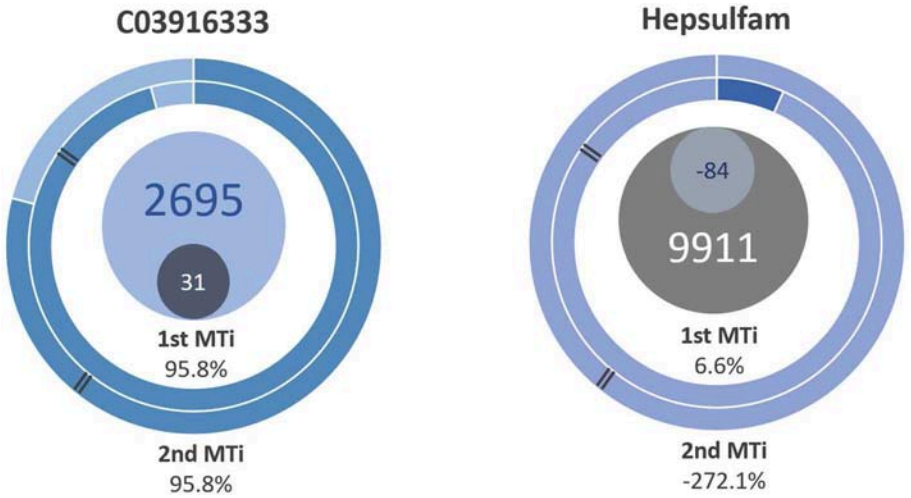


FIGURE 6 | Multitarget parameters of the most potent multitarget ligand and hepsulfam, the 3rd most selective. The darker circle represents the selectivity, while the lighter represents de potency. Hepsulfam had a negative MT potency.

Advantages and Limitations of the Multitarget Indexes

The usefulness of using a multitarget index varies highly according to the necessities of each research group. As a first instance, the multitarget indexes give an initial quantitative and workable definition for what a multitarget drug is. They define and measure multitargeticity. The primary purpose of the analysis is in drug design for identification of *in silico* and potential *in vivo* multitarget drugs. In a sense, these indexes could be useful in identifying multitarget hits and leads in the drug discovery process. Second, because it is an index, it can be useful in data analysis when comparing several ligands or targets at the same time. Moreover, the analysis assumes that the targets are independent of each other, which provides freedom regarding the number of studied targets.

The index is also modifiable and perfectible in several ways. For example, if highlighting a particular target is desired, then coefficients can be introduced so that the affinities are weighted. More calculations may be performed on the affinities in previous steps without changing the procedure or interpretation of the index, such as introducing ligand efficiency metrics. The multitarget indexes do not only identify multitarget ligands but are useful when selectivity is desired, making them not only applicable in multitarget drug design but also in designing selective single-target drugs and in the drug discovery process in general. Finally, the analysis can be further perfected with more statistical rigor, more meaningful parameters, and an *in vitro* and *in vivo* extensions.

Like all other metrics, it has limitations that skew or simplify the underlying mechanisms. For example, equal affinities may not necessarily imply multitarget *in vivo* effectiveness, since there are more variables to consider.

Various suppositions are needed in order to treat ligands as vectors. The most obvious one is that indexes do not consider how the ligand binds the target or the mechanism of action. Second, it is assumed the affinities are calculated or measured under the same conditions. Third, although the coordinates can be any type of input, the final MT index value changes if the units of the affinities introduced are different; therefore, MT index values can only be compared if the data is processed the same way and is in the same units.

The performance of the MT indexes in all the training sets shows that they correctly classify drugs using experimental values, identify potential multitarget ligands *in silico*, and can filter an extensive database, making them valuable for the intended purposes.

MT Indexes in the Virtual Library of Candidates for MS

With the known limitations and advantages, the MT indexes were used to analyze the experimental set. The 214 docked ligands were submitted to the analysis previously described and ranked in descending order of multitarget potency. Of the 211 designed compounds, 45 were derivatives from scaffold W, 34 from scaffold X, 29 from scaffold Y, and 103 from scaffold Z. Scaffold Z was present in 9 out of the top 10 most potent multitarget ligands. Benzamidine isostere A was present in 5 of them, hinting that an aminopyridine fragment may be ideal for a multitarget effect.

In most cases, a linker of 1 carbon atom and an ester group were found in the most potent ligands. More linkers diminished the MT index value, increasing the selectivity toward KLK-6, the only exception to this rule was the top, most potent molecule with four linkers. The reference molecules and their multitarget parameters can be seen in **Table 2**. In **Table 3**, the top 5 most potent multitarget ligands from the experimental set are presented.

In multitarget drug design, the multitarget potency combines the two indexes and the force. For choosing candidates, the force against the 2nd MT index can be plotted, and the regions divided into quadrants. The most favorable zone would be the upper left since it would group the most potent and most specific multitarget ligands (**Graph 2**).

Chemoinformatic Analysis

With the help of the chemoinformatic score (CIS), the ligands were classified into three arbitrary categories: preferred (CIS> 0.75), sufficient (0.75> CIS> 0.5) and risky (CIS <0.5). In total, nine ligands (4.23%) entered the preferred classification. Of these, eight belonged to scaffold X, and 1 to scaffold Z. Overall, 165 ligands (77.46%) fit into the sufficient category, and 39 (18.31%) were classified as risky (see **Table 4**).

From the 45 scaffold W ligands, 14 (31.11%) were considered risky, while 31 ligands (68.89%) were considered sufficient; none reached a CIS> 0.75. From scaffold X, 11 ligands (32.35%) obtained a risky score, 15 ligands (44.11%) were sufficient, and 8 (23.23%) were preferred. Scaffold Y had seven ligands (24.14%) classified as risky, and 22 ligands (75.86%) classified as sufficient. Finally, seven ligands (6.80%) of scaffold Z were considered risky, while 95 ligands (92.22%) were considered sufficient, and only 1 (∼1%) was preferred. These results can be seen in **Table 5**.

TABLE 2 | Summary of the multitarget metrics of the reference ligands.

Name	KLK-6 MolDock Score	PAR-1 MolDock Score	S1RP1 MolDock Score	F	^{1st}MT index (%)	^{2nd}MT index (%)	P_{MT}
Vorapaxar	−183.0	−147.7	−128.2	267.9	91.7	70.2	**99.6**
Siponimod	−143.9	−184.2	−155.8	280.9	95.7	78.8	**122.3**
0HM	−185.5	−140.5	−148.2	275.9	94.1	74.5	**111.8**

The bold values indicate the main or critical parameter(s) being evaluated.

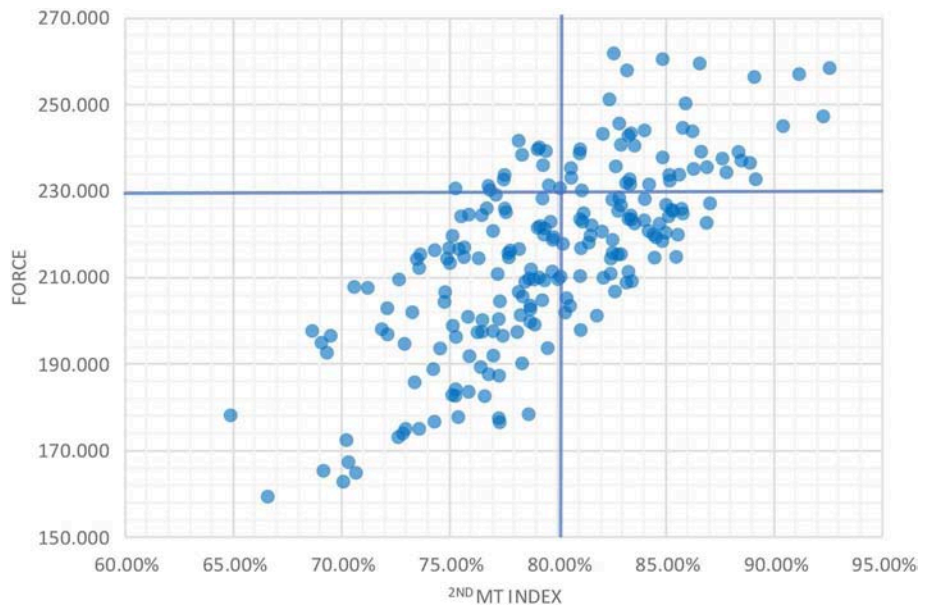
TABLE 3 | Summary of the top 5 ligands which had the highest potency.

Name	KLK-6 MolDock score	PAR-1 MolDock score	S1RP1 MolDock score	F	^{1st}MT index (%)	^{2nd}MT index (%)	P_{MT}
$Z_{b0}A_{d4}$	−210.2	−174.8	−177.9	326.2	97.2	82.6	261.9
$Z_{b0}C_{0l1}$	−193.4	−161.3	−188.9	314.8	97.5	84.8	260.5
$Z_{c0}B_{0l1}$	−190.1	−160.9	−176.8	305.4	98.2	86.6	259.5
$Z_{d0}A_{p2}$	−167.1	−153.4	−165.5	280.8	99.4	92.6	258.5
$Z_{b0}A_{d1}$	−202.5	−164.5	−183.8	319.1	97.2	83.2	257.9

TABLE 4 | Distribution of the chemoinformatic score among the ligands with the 4 scaffolds.

	Preferred (CIS>0.75)	Sufficient (0.75>CIS>0.5)	Risky (CIS<0.5)	Total
Scaffold W	-	31 (68.9%)	14 (31.1%)	45
Scaffold X	8 (23.2%)	15 (44.1%)	11 (32.3%)	34
Scaffold Y	-	22 (75.8%)	7 (24.1%)	29
Scaffold Z	1 (∼1%)	95 (92.2%)	7 (6.8%)	103
Total	9	163	39	211

Ligands with scaffold X were the most preferred in terms of pharmacokinetics theoretical properties.



Graph 2 | When plotted, the upper left quadrant is the most valuable or with the most potentially multitarget ligands.

TABLE 5 | Distribution of the chemoinformatic score among benzamidine isosteres.

	Preferred (CIS>0.75)	Sufficient (0.75>CIS>0.5)	Risky (CIS<0.5)	Total
Isostere A	8 (5.9%)	114 (83.8%)	14 (10.3%)	136
Isostere B	-	2 (15.4%)	11(84.6%)	13
Isostere C	-	7 (53.8%)	6 (46.1%)	13
Isostere D	-	6 (66.6%)	3 (33.3%)	9
Isostere E	-	10 (100%)	-	10
Isostere F	-	9 (100%)	-	9
Isostere G	1 (10%)	7 (70%)	2 (20%)	10
Isostere H	-	9 (81.8%)	2 (18.2%)	11

A considerable proportion of ligands with isosteres B and C were classified as risky, while ligands with isostere A had more preferred CIS.

TABLE 6 | Summary of the top 5 ligands which had the highest combined values of Potency and CIS.

Name	F	1st MTi (%)	2nd MTi (%)	P	CIS	Final value
$X_{b0}G_0$	**294.1**	96.9	82.7	235.8	0.81	**10.5**
$Z_{b0}A_{s0}$	**314.3**	95.8	79.4	239.3	0.78	**10.4**
$X_{b0}A_{s0}$	**269.4**	97.8	85.7	225.9	0.82	**10.4**
$X_{b0}A_{m1}$	**273.4**	97.2	84.0	223.3	0.82	**10.3**
$X_{b0}A_{m0}$	**261.1**	98.2	86.9	222.7	0.80	**10.1**

The bold values indicate the main or critical parameter(s) being evaluated.

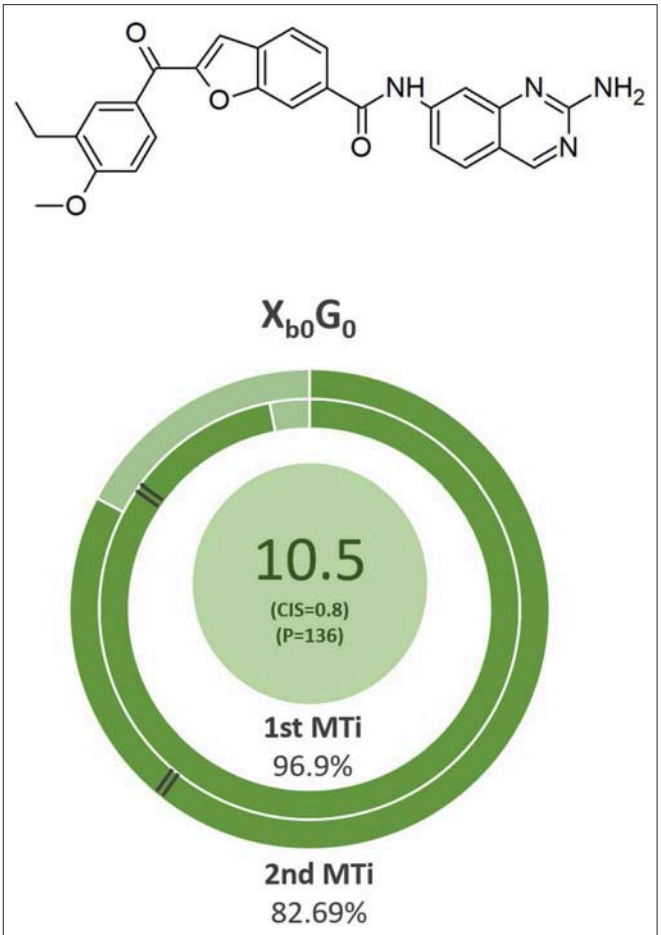


FIGURE 7 | Ligand Xb0G0 had the highest combined multitarget theoretical affinity and *in silico* ADME profile. The parameters of the MT indexes analysis. The final score is the combination of the MT potency and CIS (in parenthesis).

These results show that scaffold Z has the most balanced theoretical pharmacological properties. However, scaffold X also presented desirable properties in the CIS; scaffold X is pharmacokinetically desired. It is also worth mentioning that the aminopyridine derivatives with carboxylic acids (B) and alcohols (C) presented a risky CIS in the chemoinformatic analysis. Therefore, ligands with these isosteres are not considered candidates for therapeutic applications.

Pharmacokinetic and Pharmacodynamic Viable Candidates

For determining possible final candidates, the CIS score and multitarget potency were combined with the geometric mean. The final table (**Table 6**) groups the ligands that combine the highest potency and CIS values, meaning they are the most likely to have a biological effect while remaining relatively safe. This is a theoretical approach; therefore, the top molecules are potential multitarget alternative candidates to treat MS. The ligands with the highest combined score shared scaffold X in most of the cases (10 out of the 20 top ligands), with scaffold Z being the second most shared among the top 20. In 9 out of the top 10 ligands, isostere A was present in the ligand structure. It is also noted that a small linker is optimal for joining these two fragments. From these results, it is assumed that isostere A, as well as scaffolds X and Z, contribute to theoretical multitarget effects. However, scaffold X ligands do not have as much multitargeticity nor force as ligands with scaffold Z but remain the safest and more pharmacokinetically favorable. The applied multitarget metrics simplified the analysis and criteria for determining viable candidates. In **Figure 7**, the most viable candidate is presented.

CONCLUSIONS

As multitarget drugs are designed and tested, methods for effectively comparing and optimizing ligands are required. We present a new interpretation of ligands as vectors; new multitarget definitions and metrics, such as order of a ligand, the force of a ligand, binding capacity and multitarget capacity; and two multitarget indexes representing multitarget potency and selectivity, all of which might prove useful in drug design. The training sets allowed the identification of the advantages and disadvantages of using these metrics in multitarget drug discovery. The data analyzed through the MT indexes served to identify pharmacokinetically and pharmacodynamically viable multitarget therapeutic candidates for MS. The indexes were also

useful for identifying selective ligands. The definitions, metrics, and analysis proposed here may provide a guide toward the definition of "multitargeticity."

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

All authors have contributed significantly to this work, in the design of the model, performance of the computational experiments, analysis of the generated data, and writing the paper. All authors read, edited, and approved the final manuscript.

ACKNOWLEDGMENTS

The authors wish to thank Universidad La Salle for funding through project NEC-07/18.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2020.00176/full#supplementary-material

REFERENCES

Behrangi, N., Fischbach, F., and Kipp, M. (2019). Mechanism of siponimod: anti-inflammatory and neuroprotective mode of action. *Cells* 8:24. doi: 10.3390/cells8010024

Bernett, M. J., Blaber, S. I., Scarisbrick, I. A., Dhanarajan, P., Thompson, S. M., and Blaber, M. (2002). Crystal structure and biochemical characterization of human kallikrein 6 reveal that a trypsin-like kallikrein is expressed in the central nervous system. *J. Biol. Chem.* 277, 24562–24570. doi: 10.1074/jbc.M202392200

Browne, P., Chandraratna, D., Angood, C., Tremlett, H., Baker, C., Taylor, B. V., et al. (2014). Atlas of multiple sclerosis 2013: a growing global problem with widespread inequity. *Neurology* 83, 1022–1024. doi: 10.1212/WNL.0000000000000768

Burda, J. E., Radulovic, M., Yoon, H., and Scarisbrick, I. A. (2013). Critical role for PAR1 in kallikrein 6-mediated oligodendrogliopathy. *Glia* 61, 1456–1470. doi: 10.1002/glia.22534

Cheng, F., Li, W., Zhou, Y., Shen, J., Wu, Z., Liu, G., et al. (2012). admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. *J. Chem. Inf. Model.* 52, 3099–3105. doi: 10.1021/ci300367a

Chi, X. X., and Nicol, G. D. (2010). The sphingosine 1-phosphate receptor, S1PR1, plays a prominent but not exclusive role in enhancing the excitability of sensory neurons. *J. Neurophysiol.* 104, 2741–2748. doi: 10.1152/jn.00709.2010

Chiarelli, L. R., Mori, G., Orena, B. S., Esposito, M., Lane, T., de Jesus Lopes Ribeiro, A. L., et al. (2018). A multitarget approach to drug discovery inhibiting *Mycobacterium tuberculosis* PyrG and PanK. *Sci. Rep.* 8:3187. doi: 10.1038/s41598-018-21614-4

Dobson, R., and Giovannoni, G. (2019). Multiple sclerosis - a review. *Eur. J. Neurol.* 26, 27–40. doi: 10.1111/ene.13819

Gan, T. J., and Habib, A. S. (2016). *Postoperative Nausea and Vomiting: A Practical Guide*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9781316135853

Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J. (2016). BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* 44, D1045–D1053. doi: 10.1093/nar/gkv1072

Hanson, M. A., Roth, C. B., Jo, E., Griffith, M. T., Scott, F. L., Reinhart, G., et al. (2012). Crystal structure of a lipid G protein–coupled receptor. *Science* 335, 851–855. doi: 10.1126/science.1215904

Hu, Y., Gupta-Ostermann, D., and Bajorath, J. (2014). Exploring compound promiscuity patterns and multi-target activity spaces. *Comput. Struct. Biotechnol. J.* 9:e201401003. doi: 10.5936/csbj.201401003

Huang, N., Shoichet, B. K., and Irwin, J. J. (2006). Benchmarking sets for molecular docking. *J. Med. Chem.* 49, 6789–6801. doi: 10.1021/jm0608356

Huang, W.-J., Chen, W.-W., and Zhang, X. (2017). Multiple sclerosis: pathology, diagnosis and treatments. *Exp. Ther. Med.* 13, 3163–3166. doi: 10.3892/etm.2017.4410

Kim, H. N., Kim, Y. R., Ahn, S. M., Lee, S. K., Shin, H. K., and Choi, B. T. (2015). Protease activated receptor-1 antagonist ameliorates the clinical symptoms of experimental autoimmune encephalomyelitis via inhibiting breakdown of blood-brain barrier. *J. Neurochem.* 135, 577–588. doi: 10.1111/jnc.13285

Kiss, R., Sandor, M., and Szalai, F. A. (2012). http://Mcule.com: a public web service for drug discovery. *J. Cheminform.* 4:P17. doi: 10.1186/1758-2946-4-S1-P17

Lee, P. R., Johnson, T. P., Gnanapavan, S., Giovannoni, G., Wang, T., Steiner, J. P., et al. (2017). Protease-activated receptor-1 activation by granzyme B causes neurotoxicity that is augmented by interleukin-1β. *J. Neuroinflamm.* 14:131. doi: 10.1186/s12974-017-0901-y

Li, Y. H., Wang, P. P., Li, X. X., Yu, C. Y., Yang, H., Zhou, J., et al. (2016). The human kinome targeted by fda approved multi-target drugs and combination products: a comparative study from the drug-target interaction network perspective. *PLoS ONE* 11:e0165737. doi: 10.1371/journal.pone.0165737

Liang, G., Chen, X., Aldous, S., Pu, S.-F., Mehdi, S., Powers, E., et al. (2012). Human kallikrein 6 inhibitors with a para-amidobenzylanmine P1 group identified through virtual screening. *Bioorg. Med. Chem. Lett.* 22, 2450–2455. doi: 10.1016/j.bmcl.2012.02.014

Loza-Mejía, M., Salazar, J., and Sánchez-Tejeda, J. (2018). *In Silico* studies on compounds derived from calceolaria: phenylethanoid glycosides as potential multitarget inhibitors for the development of pesticides. *Biomolecules* 8:121. doi: 10.3390/biom8040121

Lu, J.-J., Pan, W., Hu, Y.-J., and Wang, Y.-T. (2012). Multi-target drugs: the trend of drug research and development. *PLoS ONE* 7:e40262. doi: 10.1371/journal.pone.0040262

Macfarlane, S. R., Seatter, M. J., Kanke, T., Hunter, G. D., and Plevin, R. (2001). Proteinase-activated receptors. *Pharmacol. Rev.* 53, 245–282.

Manaenko, A., Sun, X., Kim, C. H., Yan, J., Ma, Q., and Zhang, J. H. (2013). PAR-1 antagonist SCH79797 ameliorates apoptosis following surgical brain injury through inhibition of ASK1-JNK in rats. *Neurobiol. Dis.* 50, 13–20. doi: 10.1016/j.nbd.2012.09.004

O'Sullivan, C., Schubart, A., Mir, A. K., and Dev, K. K. (2016). The dual S1PR1/S1PR5 drug BAF312 (Siponimod) attenuates demyelination in organotypic slice cultures. *J. Neuroinflamm.* 13:31. doi: 10.1186/s12974-016-0494-x

Pan, H., Boucher, M., and Kaunelis, D. (2016). *PAR-1 Antagonists: An Emerging Antiplatelet Drug Class. CADTH Issues in Emerging Health Technologies*. Canadian Agency for Drugs and Technologies in Health. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27809429

Ramsay, R. R., Popovic-Nikolic, M. R., Nikolic, K., Uliassi, E., and Bolognesi, M. L. (2018). A perspective on multi-target drug discovery and design for complex diseases. *Clin. Transl. Med.* 7:3. doi: 10.1186/s40169-017-0181-2

Ravdin, P. M., Havlin, K. A., Marshall, M. V., Brown, T. D., Koeller, J. M., Kuhn, J. G., et al. (1991). A phase I clinical and pharmacokinetic trial of hepsulfam. *Cancer Res.* 51(23 Pt 1), 6268–6272.

Readmond, C., and Wu, C. (2017). Investigating detailed interactions between novel PAR1 antagonist F16357 and the receptor using docking and molecular dynamic simulations. *J. Mol. Graph. Model.* 77, 205–217. doi: 10.1016/j.jmgm.2017.08.019

Rosini, M. (2014). Polypharmacology: the rise of multitarget drugs over combination therapies. *Future Med. Chem.* 6, 485–487. doi: 10.4155/fmc.14.25

Scarisbrick, I. A., Epstein, B., Cloud, B. A., Yoon, H., Wu, J., Renner, D. N., et al. (2011). Functional role of kallikrein 6 in regulating immune cell survival. *PLoS ONE* 6:e18376. doi: 10.1371/journal.pone.0018376

Silva, R. N., Oliveira, L. C. G., Parise, C. B., Oliveira, J. R., Severino, B., Corvino, A., et al. (2017). Activity of human kallikrein-related peptidase 6 (KLK6) on substrates containing sequences of basic amino acids. Is it a processing protease? *Biochim. Biophys. Acta Proteins Proteom.* 1865, 558–564. doi: 10.1016/j.bbapap.2017.02.017

Talevi, A. (2015). Multi-target pharmacology: possibilities and limitations of the "skeleton key approach" from a medicinal chemist perspective. *Front. Pharmacol.* 6:205. doi: 10.3389/fphar.2015.00205

van der Schyf, C. J. (2011). The use of multi-target drugs in the treatment of neurodegenerative diseases. *Expert Rev. Clin. Pharmacol.* 4, 293–298. doi: 10.1586/ecp.11.13

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Xie, L., and Bourne, P. E. (2015). Developing multi-target therapeutics to fine-tune the evolutionary dynamics of the cancer ecosystem. *Front. Pharmacol.* 6:209. doi: 10.3389/fphar.2015.00209

Yoon, H., Radulovic, M., Drucker, K. L., Wu, J., and Scarisbrick, I. A. (2015). The thrombin receptor is a critical extracellular switch controlling myelination. *Glia* 63, 846–859. doi: 10.1002/glia.22788

Yoon, H., and Scarisbrick, I. A. (2016). Kallikrein-related peptidase 6 exacerbates disease in an autoimmune model of multiple sclerosis. *Biol. Chem.* 397, 1277–1286. doi: 10.1515/hsz-2016-0239

Zhang, C., Srinivasan, Y., Arlow, D. H., Fung, J. J., Palmer, D., Zheng, Y., et al. (2012). High-resolution crystal structure of human protease-activated receptor 1. *Nature* 492, 387–392. doi: 10.1038/nature 11701

Zhang, W., Pei, J., and Lai, L. (2017). Computational multitarget drug design. *J. Chem. Inf. Model.* 57, 403–412. doi: 10.1021/acs.jcim. 6b00491

Conflict of Interest: JS-R was employed by company Ciencia y Estrategia S.A. de C.V.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Sánchez-Tejeda, Sánchez-Ruiz, Salazar and Loza-Mejía. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 24 March 2020
doi: 10.3389/fchem.2020.00164

In silico Design of Novel HIV-1 NNRTIs Based on Combined Modeling Studies of Dihydrofuro[3,4-d]pyrimidines

Yanming Chen[1], *Yafeng Tian*[1], *Ya Gao*[1], *Fengshou Wu*[1], *Xiaogang Luo*[1,2], *Xiulian Ju*[1] *and Genyan Liu*[1]*

[1] Key Laboratory for Green Chemical Process of Ministry of Education, Hubei Key Laboratory of Novel Reactor and Green Chemical Technology, School of Chemical Engineering and Pharmacy, Wuhan Institute of Technology, Wuhan, China,
[2] School of Materials Science and Engineering, Zhengzhou University, Zhengzhou, China

A novel series of dihydrofuro[3,4-d]pyrimidine (DHPY) analogs have recently been recognized as promising HIV-1 non-nucleoside reverse transcriptase (RT) inhibitors (NNRTIs) with potent antiviral activity. To better understand the pharmacological essentiality of these DHPYs and design novel NNRTI leads, in this work, a systematic *in silico* study was performed on 52 DHPYs using three-dimensional quantitative structure–activity relationship (3D-QSAR), molecular docking, virtual screening, absorption-distribution-metabolism-excretion (ADME) prediction, and molecular dynamics (MD) methods. The generated 3D-QSAR models exhibited satisfactory parameters of internal validation and well-externally predictive capacity, for instance, the q^2, R^2, and r_{pred}^2 of the optimal comparative molecular similarity indices analysis model were 0.647, 0.970, and 0.751, respectively. The docking results indicated that residues Lys101, Tyr181, Tyr188, Trp229, and Phe227 played important roles for the DHPY binding. Nine lead compounds were obtained by the virtual screening based on the docking and pharmacophore model, and three new compounds with higher docking scores and better ADME properties were subsequently designed based on the screening and 3D-QSAR results. The MD simulation studies further demonstrated that the newly designed compounds could stably bind with the HIV-1 RT. These hit compounds were supposed to be novel potential anti-HIV-1 inhibitors, and these findings could provide significant information for designing and developing novel HIV-1 NNRTIs.

Keywords: HIV-1 non-nucleoside reverse transcriptase inhibitors (NNRTIs), dihydrofuro[3,4-d]pyrimidines, virtual screening, molecular docking, rational drug design

OPEN ACCESS

Edited by:
Kamil Kuca,
University of Hradec Králové, Czechia

Reviewed by:
Marco Tutone,
University of Palermo, Italy
Margherita Brindisi,
University of Naples Federico II, Italy

***Correspondence:**
Genyan Liu
liugenyan@yahoo.com;
liugenyan@wit.edu.cn

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *11 October 2019*
Accepted: *25 February 2020*
Published: *24 March 2020*

Citation:
Chen Y, Tian Y, Gao Y, Wu F, Luo X, Ju X and Liu G (2020) In silico Design of Novel HIV-1 NNRTIs Based on Combined Modeling Studies of Dihydrofuro[3,4-d]pyrimidines. Front. Chem. 8:164. doi: 10.3389/fchem.2020.00164

INTRODUCTION

Acquired immune deficiency syndrome (AIDS) caused by human immunodeficiency virus (HIV) is one of the most widely spread infectious diseases worldwide. There is no effective drug or vaccine that could cure AIDS absolutely at present. According to the report from the Joint United Nations Program on HIV/AIDS, there were approximately 36.9 million people living with HIV worldwide in 2018, and neighboring 1.8 million new cases and 0.94 million AIDS-related deaths

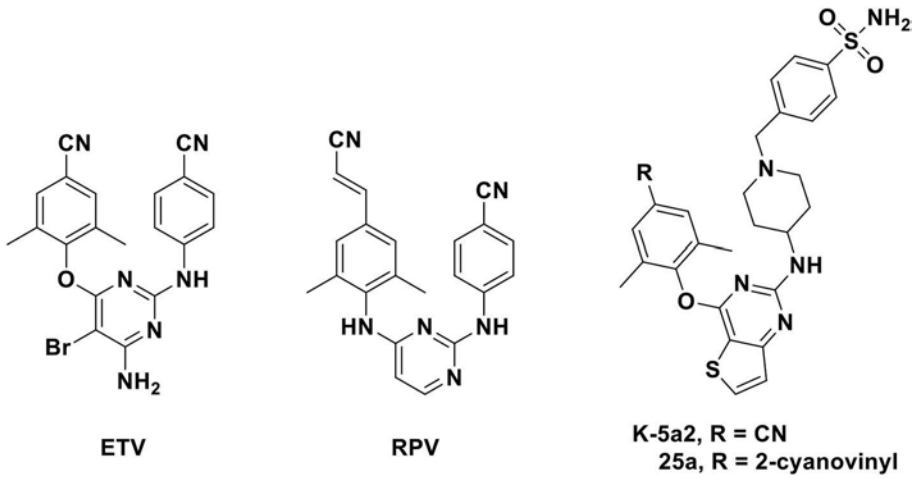


FIGURE 1 | Chemical structures of diarylpyrimidines (DAPYs).

in 2017[1]. Two main types of HIV (HIV-1 and HIV-2) have been identified currently. HIV-1 is widely spread throughout the world, whereas HIV-2 has correspondingly poor transmission (Vasavi et al., 2019; Wang et al., 2019). In the fight against HIV-1, highly active antiretroviral therapy (HAART) has been considered to be a relatively successful and effective therapy in controlling HIV-1 epidemics (Chen et al., 2011; Wang et al., 2018).

HIV-1 reverse transcriptase (RT), as one of the most important enzymes that convert the single-stranded RNAs into double-stranded DNAs, is vital to restrain HIV-1 replication and a prime target for antiviral research (Esposito et al., 2012). Inhibitors of the HIV-1 RT are divided into nucleoside RT inhibitors (NRTIs) and non-nucleoside RT inhibitors (NNRTIs), and the latter binds to an allosteric site that is located about 10 Å distance from the polymerizing processing site (Zhan et al., 2009). NNRTIs have become an indispensable portion of HAART regimen due to its potent antiviral activity, high specificity, and low cytotoxicity. However, single mutations such as K103N, Y181C, V106A, and L100I in the binding site of the HIV-1 RT might result in decreased inhibitory potencies of NNRTIs, and a double mutation (K103N+Y181C) was more frequently discovered in the process of treating with NNRTIs (Das et al., 2008).

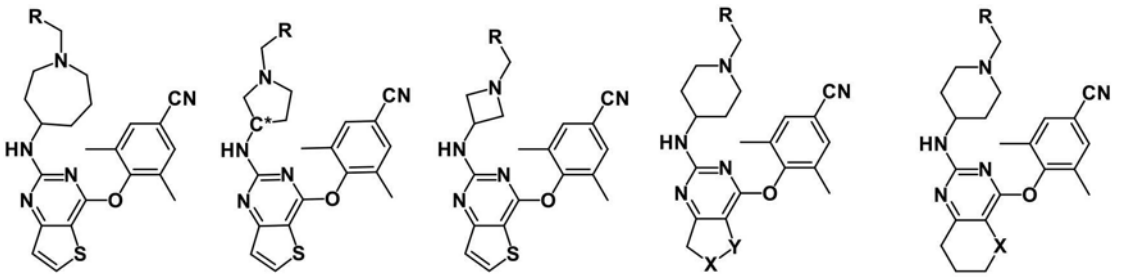
Six HIV-1 NNRTIs including nevirapine, delavirdine, efavirenz, etravirine (ETV), rilpivirine (RPV), and doravirine have been approved by US Food and Drug Administration for clinical use to date (Namasivayam et al., 2019). ETV and RPV (**Figure 1**), which belong to diarylpyrimidine (DAPY) derivatives that were recognized as one of the most effective families of NNRTIs, have attracted considerable attention due to their excellent potency against HIV-1 wild-type and mutant strains. However, the low solubility and unsatisfactory oral bioavailability of these analogs restrict their clinical usage in some respects (Gu et al., 2019). Thus, novel NNRTIs with improved pharmacokinetic profiles have been urged to design and discover.

Recently, Kang et al. (2016, 2017) have designed and synthesized a series of thiophene[3,2-d]pyrimidine derivatives, among which compounds **K-5a2** and **25a** (**Figure 1**) were two representative HIV-1 NNRTIs, exhibiting more drug-like pharmacokinetic properties and greater inhibitory activities compared to nevirapine and efavirenz. Compound **25a** also exhibited better inhibition against HIV-1 mutant strains than ETV and RPV. However, compound **K-5a2** did not display excellent activity against K103N+Y181C mutant HIV-1 strains (Kang et al., 2017; Yang et al., 2018). Further structural modification on **K-5a2** and **25a** using six alicyclic-fused pyrimidine rings led to a series of dihydrofuro[3,4-d]pyrimidine (DHPY) derivatives with potent anti-HIV activity (**Table 1**) (Kang et al., 2019).

To date, there are many computer-aided drug design methods applied in designing and developing novel HIV-1 inhibitors (Almerico et al., 2007). For example, the three-dimensional quantitative structure–activity relationship (3D-QSAR) and pharmacophore models were utilized to learn about structural characteristics of HIV-1 NNRTIs in our previous studies (Liu et al., 2018; Wan et al., 2018). The multivariate statistical procedures, containing principal component and discriminant analysis, could be as credible methods to predict the activities of HIV-1 inhibitors by taking advantage of the vast anti-HIV data (Almerico et al., 2003, 2006). The molecular docking and molecular dynamics (MD) simulation were often used to understand the binding conformations of ligands in the active sites of HIV-1-related proteins. Furthermore, a comparative analysis with the combination of docking and multivariate methods was used to study the drug resistance of HIV-1 inhibitors and to further design new compounds with appropriate structural features (Almerico et al., 2008).

[1]UNAIDS. *UNAIDS Data 2018*. Available online at: http://www.unaids.org/en/resources/documents/2018/unaids-data-2018

TABLE 1 | Chemical structures of DHPYs and their actual and predicted activities as HIV-1 NNRTIs.



compounds 01-04 | compounds 05-09: *C= (R) compounds 10-13: *C= (S) | compounds 14-18 | compounds 19-24: X=C Y=S compounds 25-29: X=S Y=C compounds 30-35: X=O Y=C compounds 36-41: X=C Y=C | compounds 42-47: X=C compounds 48-52: X=S

				CoMFA		CoMSIA	
No.	***R***	**EC_{50} (nM)**	**Actual pEC_{50}**	**Predicted pEC_{50}**	**Residual**	**Predicted pEC_{50}**	**Residual**
01[b]	4-SO_2NH_2-Ph	2.20	8.658	8.728	0.070	8.672	0.014
02	3-$CONH_2$-Ph	10.3	7.987	8.043	0.056	8.014	0.027
03	4-SO_2CH_3-Ph	9.98	8.001	7.919	−0.082	7.992	−0.009
04[a]	4-pyridinyl	31.7	7.499	7.528	0.029	7.392	−0.107
05[b]	4-SO_2NH_2-Ph	8.69	8.061	8.046	−0.015	8.076	0.015
06	4-$CONH_2$-Ph	10.4	7.983	8.042	0.059	8.025	0.042
07	3-$CONH_2$-Ph	41.3	7.384	7.360	−0.024	7.365	−0.019
08	4-SO_2CH_3-Ph	13.9	7.857	7.923	0.066	7.860	0.003
09[a]	4-pyridinyl	16.0	7.796	7.500	−0.296	7.830	0.034
10	4-SO_2NH_2-Ph	104	6.983	7.083	0.100	7.004	0.021
11	4-$CONH_2$-Ph	55.7	7.254	7.164	−0.090	7.242	−0.012
12	4-SO_2CH_3-Ph	50.7	7.295	7.214	−0.081	7.275	−0.020
13[b]	4-pyridinyl	16.7	7.777	7.843	0.066	7.781	0.004
14	4-SO_2NH_2-Ph	4.53	8.344	8.251	−0.093	8.335	−0.009
15	4-$CONH_2$-Ph	4.76	8.322	8.317	−0.005	8.304	−0.018
16	3-$CONH_2$-Ph	8.95	8.048	7.954	−0.094	8.053	0.005
17[a]	4-SO_2CH_3-Ph	207	6.684	6.860	0.176	7.329	0.645
18[b]	4-pyridinyl-Ph	2.21	8.656	8.690	0.034	8.680	0.024
19	4-SO_2NH_2-Ph	4.3	8.367	8.468	0.101	8.467	0.100
20[a]	4-$CONH_2$-Ph	4.8	8.319	8.490	0.171	8.328	0.009
21	4-SO_2CH_3-Ph	5.9	8.229	8.197	−0.032	8.238	0.009
22[b]	4-pyridinyl	2.6	8.585	8.613	0.028	8.637	0.052
23[a]	4-NO_2-Ph	8.0	8.097	8.103	0.006	8.226	0.129
24	3-$CONH_2$-Ph	27.7	7.558	7.500	−0.058	7.547	−0.011
25[a]	4-SO_2NH_2-Ph	37.2	7.429	7.494	0.065	7.263	−0.166
26[a,b]	4-SO_2Me-Ph	3.8	8.420	8.406	−0.014	7.714	−0.706
27	4-NO_2-Ph	11.5	7.939	7.902	−0.037	7.898	−0.041
28	4-NH_2-Ph	8.4	8.076	7.889	−0.187	8.067	−0.009
29	4-$NHSO_2$Me-Ph	11.2	7.951	8.043	0.092	7.987	0.036
30	4-SO_2NH_2-Ph	2.8	8.553	8.562	0.009	8.543	−0.010
31[b]	4-$CONH_2$-Ph	1.6	8.796	8.768	−0.028	8.812	0.016
32	4-SO_2CH_3-Ph	1.9	8.721	8.677	−0.044	8.629	−0.092
33	4-pyridinyl	2.3	8.638	8.701	0.063	8.736	0.098
34	4-NO_2-Ph	7.4	8.131	8.075	−0.056	8.178	0.047
35[a]	3-$CONH_2$-Ph	7.8	8.108	7.971	−0.137	8.238	0.130
36[b]	4-SO_2NH_2-Ph	1.1	8.959	8.867	−0.092	8.941	−0.018
37	4-$CONH_2$-Ph	6.1	8.215	8.387	0.172	8.270	0.055

(Continued)

TABLE 1 | Continued

No.	R	EC_{50} (nM)	Actual pEC_{50}	CoMFA		CoMSIA	
				Predicted pEC_{50}	Residual	Predicted pEC_{50}	Residual
38[a]	4-SO_2CH_3-Ph	4.9	8.310	8.603	0.293	8.479	0.169
39	4-pyridinyl	1.8	8.745	8.642	−0.103	8.452	−0.293
40	4-NO_2-Ph	14.5	7.839	7.953	0.114	7.858	0.019
41[a]	3-$CONH_2$-Ph	2.0	8.699	8.060	−0.639	8.437	−0.262
42	4-SO_2NH_2-Ph	6.0	8.222	8.145	−0.077	8.169	−0.053
43[b]	4-$CONH_2$-Ph	6.0	8.222	8.182	−0.040	8.218	−0.004
44[a]	4-SO_2CH_3-Ph	8.0	8.097	8.278	0.181	8.308	0.211
45	4-pyridinyl	8.6	8.066	8.033	−0.033	8.188	0.122
46	4-NO_2-Ph	77.4	7.111	7.330	0.219	7.106	−0.005
47[a]	3-$CONH_2$-Ph	6.5	8.187	7.792	−0.395	8.369	0.182
48[b]	4-SO_2NH_2-Ph	2.7	8.569	8.588	0.019	8.540	−0.029
49	4-$CONH_2$-Ph	3.0	8.523	8.523	0.000	8.434	−0.089
50	4-SO_2CH_3-Ph	3.9	8.409	8.453	0.044	8.451	0.042
51	4-NO_2-Ph	8.6	8.066	8.024	−0.042	8.056	−0.010
52[a]	3-$CONH_2$-Ph	5.1	8.292	8.087	−0.205	8.429	0.137

[a] *Test set compounds used for 3D-QSAR models.*

[b] *The compounds used for pharmacophore models.*

CoMFA, comparative molecular field analysis; CoMSIA, comparative molecular similarity indices analysis; DHPY, dihydrofuro[3,4-d]pyrimidine; NNRTIs, non-nucleoside reverse transcriptase inhibitors.

To further explore the essential structural and pharmacological features of the novel DHPYs as HIV-1 NNRTIs in this study, the combination of 3D-QSAR models, molecular docking, and MD simulation was applied to analyze the 3D-QSARs of these DHPYs and their binding modes in the HIV-1 RT. We also utilized the pharmacophore- and docking-based virtual screening to obtain some hit compounds from ZINC database and subsequently designed new potential NNRTIs according to the screening and 3D-QSAR results. Molecular docking and MD simulations were utilized to identify the binding of these new NNRTIs and the stabilization of the protein–ligand complexes.

MATERIALS AND METHODS

Preparation of Small Molecules

A total of 52 DHPY derivatives were collected from the published literature (Kang et al., 2019) for performing the molecular modeling study. Their structures, EC_{50}, and corresponding pEC_{50} ($-\log EC_{50}$) values were listed in **Table 1**. All compounds were stretched by SYBYL-X 2.1 (Tripos Inc., St. Louis, USA) running on Windows 7 workstation and minimized with Gasteiger–Hückel charges, the termination of 0.005 kcal/(mol·Å) and max iterations of 1,000 by Powell method. Other parameters were set to default values.

Three-Dimensional Quantitative Structure–Activity Relationship Model

The 3D-QSAR model could help to find a significant correlation between the biological activities of drug molecules and their structures (Borisa and Bhatt, 2015). In this study, comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) methods were used to construct 3D-QSAR models. All compounds were randomly divided into a training set (39 compounds) to generate CoMFA and CoMSIA models and a test set (13 compounds) to confirm the reliability of the generated models (**Table 1**). The number of test set compounds should be kept in the range from 1/4 to 1/3 of the total compounds. Compound **36** with the highest activity was used as a template, and all training set compounds were superimposed on it by the common skeleton alignment (**Figure 2A**).

For generating a reasonable model, the internal predictive ability was evaluated by partial least squares (PLS) regression method using the SAMPLS. The leave-one-out (LOO) cross-validation procedure was applied to determine the optimum number of components (ONC) and the highest cross-validation correlation coefficient (Q^2) (Bush and Nachbar, 1993), and non-cross-validated analysis was applied to compute the non-cross-validated correlation coefficient (R^2), standard error of estimate (SEE), and the Fisher test values (F) (Li et al., 2014). External validation parameters were also essential for further assessing the predictive capability of 3D-QSAR models, such as ${r_0}^2$, k, ${r_0}'^2$, and k'. ${r_0}^2$, and k were the corresponding correlation coefficient and the slope value of linear regression equation, respectively, for predicted vs. actual activities when the intercept was set to zero, and ${r_0}'^2$ and k' were for actual vs. predicted activities, respectively. In addition, ${r_m}^2$, ${r_m}'^2$, $\triangle {r_m}^2$, $\overline{r_m^2}$, and the root mean square error (RMSE) as traditional data were also calculated to appraise the predictive ability. A model, which met the requirements of $[(r^2-{r_0}^2)/r^2]$ or $[(r^2-{r_0}'^2)/r^2] < 0.1, 0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$, $\triangle {r_m}^2 < 0.2$ and $\overline{r_m^2} > 0.5$, especially

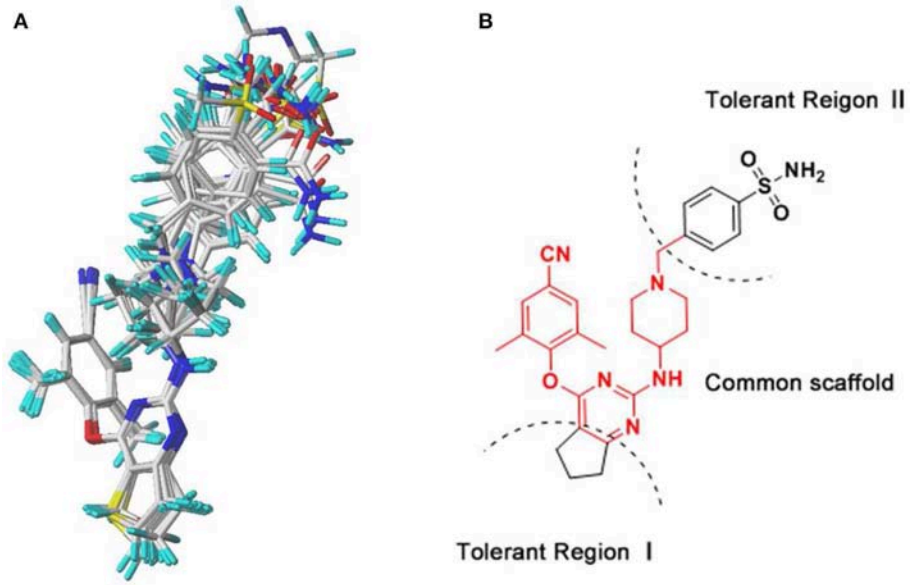


FIGURE 2 | The molecular alignment using compound **36** as a template. **(A)** The alignment results of the training set compounds used for the three-dimensional quantitative structure–activity relationship (3D-QSAR) models. **(B)** The chemical structure of compound **36**, and the red region represents the common scaffold.

the predictive correlation $r_{pred}^2 > 0.6$, would be deemed to possess well-predictive capability and reliability (Caballero, 2010; Ojha et al., 2011; Roy et al., 2016). The parameters were calculated according to our previous studies (Wang et al., 2018; Gao et al., 2019; Liu et al., 2019).

Pharmacophore Model

Ten compounds (**Table 1**) with high activities and diverse structures were selected to generate pharmacophore model using Genetic Algorithm with Linear Assignment of Hypermolecular Alignment of Database (GALAHAD) module in SYBYL-X 2.1. GALAHAD method mainly contained two steps. The ligands are neatly aligned to each other in internal coordinate space, and then the produced conformations as rigid bodies are aligned in Cartesian space. In the process of running GALAHAD, the parameters of population size, max generation, and molecules required to hit were automatically set according to the experiment activity data. Finally, 20 models with diverse parameters including SPECIFICITY, N_HITS, STERICS, HBOND, and Mol_Qry were generated.

In order to further validate the ability of the pharmacophore model, a decoy set method was used for evaluating the generated model. The decoy set database was comprised of 6,234 inactive compounds downloaded from the DUD-E database (http://dud.docking.org/) (Mysinger et al., 2012) and 42 active compounds from **Table 1** except the compounds used for constructing the pharmacophore model. The enrichment factor (EF) and Güner–Henry (GH) scores were considered as metrics to assess the reliability of the pharmacophore models. The GH score took the percent yield of actives in a hit list (%Y, recall) and the percent ratio of actives in a database (%A, precision) into account. While the GH score is ranging 0.6–1, the pharmacophore model would be regarded as a rational model (Kalva et al., 2014).

$$\%Y = H_a/H_t \times 100\% \quad (1)$$

$$\%A = H_a/A \times 100\% \quad (2)$$

$$EF = (H_a/H_t)/(A/D) \quad (3)$$

$$GH = (H_a\,(3A + H_t))/(4AH_t) \times (1 - (H_t - H_a)/(D - A)) \quad (4)$$

where H_a is the number of active molecules in the hit list, H_t is the hit compounds from the decoy set database, A is the total number of active compounds in the database, and D is the sum of the database.

Molecular Docking

The crystal structures of wild-type HIV-1 RT (PDB ID: 6C0J) and K103N/Y181C mutant RT (PDB ID: 6C0R) were downloaded from the Protein Data Bank and were used for the docking study. While preparing the two proteins, hydrogen atoms were added after the crystallographic ligands were extracted and all water molecules except for W936 were removed. In order to verify the rationality and reliability of the docking method, the extracted ligands (**K-5a2** and **25a**) were first redocked into the corresponding active site using the Surflex-Dock Geom module of SYBYL-X 2.1 with default parameters. All compounds were then docked into the binding pocket as the same pattern. Twenty conformations with different scores were produced for each docked compound, and the highest-score conformation of each compound was chosen for further study.

Virtual Screening

The selected GALAHAD model was converted into a UNITY query for virtual screening from ZINC database, and the "Flex search" was employed to serve as query type. Lipinski's rule of five as the primary filter was utilized to further decrease screened compounds. The QFIT score, whose value was between 0 and 100, reflected how closely the hit compounds matched with query. In consideration of the time and accuracy of screening, two ways of molecular docking including Surflex-Dock and Surflex-Dock Geom were implemented to verify the potential hit compounds obtained from the pharmacophore-based screening.

ADME Analysis

ADME properties are essential for selecting and evaluating lead candidates. The online tool Swiss ADME (http://www.swissadme.ch/index.php) was applied to calculate the pharmacokinetic properties of new NNRTI candidates, such as lipophilicity, water solubility, and blood–brain barrier (BBB) permeability (Daina et al., 2017). The synthetic accessibility was also predicted by the score from 1 to 10, in which a score of 1 suggested the synthetic route was relatively easy and a score closer to 10 indicated the compound had complex structure and was tough to be synthesized.

Molecular Dynamics Simulation

To further explore the dynamics protein–ligand interactions, 10 ns MD simulations were performed on compound **36** and newly designed inhibitors using GROMACS2016.5 software with AMBER 99SB force field. Before starting MD simulation, several important procedures should be performed to generate a steady environment. First, it was very momentous to generate the topology file of ligand by a acpype tool, which was on the basis of Python. Second, a 12 Å × 12 Å × 12 Å cubic box full of water models (transferable intermolecular potential with 3 points) was set to create the aqueous environment for the whole system. Nine chloride ions were added into the box for the sake of keeping the state of charge neutrality. In order to satisfy a tolerance of 10 kJ/mol, the steepest descent method for 50,000 steps was carried out for minimization without constraint to avoid possible collision between atoms. NVT at 300 K using V-rescale for 100 ps and NPT at 1 atm pressure using Parrine–Rahman for 100 ps were successively equilibrated to maintain proper temperature and pressure for the system. At last, the 10 ns MD simulation was run and the simulation step length was defined as 2 fs.

RESULTS AND DISCUSSION

Statistical Analysis of the Comparative Molecular Field Analysis and Comparative Molecular Similarity Indices Analysis Models

The classical parameters of the CoMFA and CoMSIA models were summarized in **Table S1**. In general, the q^2 and R^2 should be more than 0.5 and 0.9, respectively, and the *SEE* and *F*-value should be rational. As for the CoMSIA models, there were different combinations of five fields as shown in **Table S1**. The model generated by the combination of the steric (S), electrostatic (E), hydrogen-bond acceptor (A), hydrogen-bond donor (D), and hydrophobic (H) fields was the optimal CoMSIA model because of its satisfactory q^2, R^2, *SEE*, F, and r_{pred}^2 values. The

TABLE 2 | External validation results of the CoMFA and CoMSIA models.

Validation parameters	RMSE	MAE	r^2	r_0^2	$r_0'^2$	$\frac{r^2-r_0^2}{r^2}$	k	k′	r_m^2	$r_m'^2$	$\triangle r_m^2$	$\overline{r_m^2}$
CoMFA	0.263	1.608	0.750	0.746	0.709	0.006	1.007	0.992	0.700	0.597	0.103	0.648
CoMSIA	0.302	0.549	0.655	0.653	0.533	0.004	0.996	1.003	0.622	0.426	0.196	0.524

CoMFA, comparative molecular field analysis; CoMSIA, comparative molecular similarity indices analysis; RMSE, root mean square error; MAE, mean absolute error.

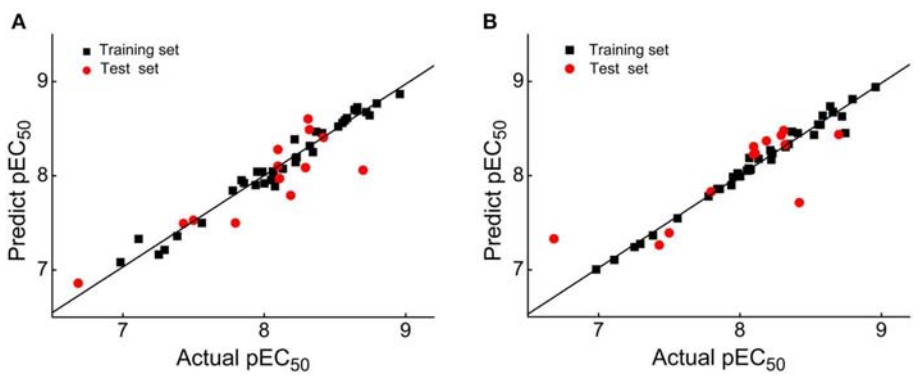


FIGURE 3 | Plots of actual vs. predicted pEC_{50} values of all dihydrofuro[3,4-d]pyrimidines (DHPYs) based on the comparative molecular field analysis (CoMFA) **(A)** and comparative molecular similarity indices analysis (CoMSIA) **(B)** models.

contributions of S, E, A, D, and H fields were 4.1, 19.7, 29, 33.4, and 13.8%, respectively, indicating that A and D fields played more important roles. The q^2 of the CoMFA and CoMSIA models were 0.647 and 0.735, respectively, which indicated that both models were rational. The R^2 values of the CoMFA and CoMSIA models were 0.970 and 0.982, respectively, and the r_{pred}^2 values were 0.751 and 0.672, respectively, suggesting that both models had excellent predictive abilities. In addition, it was common for the CoMFA and CoMSIA models that the E field contribution was more than the S field contribution, which illustrated that the E field could be more significant than the S field in the effect on compound activity.

External validation parameters could further confirm the reasonability of the constructed CoMFA and CoMSIA models. As shown in **Table 2**, all external validation results of the CoMFA and CoMSIA models were in the rational range, for example, the $\overline{r_m^2}$ values of the CoMFA and CoMSIA model were 0.648 and 0.524, respectively. The statistical results of **Table S1** and **Table 2** proved that the generated 3D-QSAR models were reliable and possessed excellent predictive capacity. **Figure 3** showed the plots of actual vs. predicted pEC_{50} values for all compounds based on the CoMFA and CoMSIA models. All compounds were evenly distributed in the two sides of the trend lines, which indicated that the 3D-QSAR models had excellent abilities to predict the activities of DHPYs. The predictive capacity of the CoMFA model seems to be better than that of the CoMSIA model.

Contour Maps of the Comparative Molecular Field Analysis and Comparative Molecular Similarity Indices Analysis Models

The contour maps of the CoMFA and CoMSIA models could visually provide significant information for the QSARs of

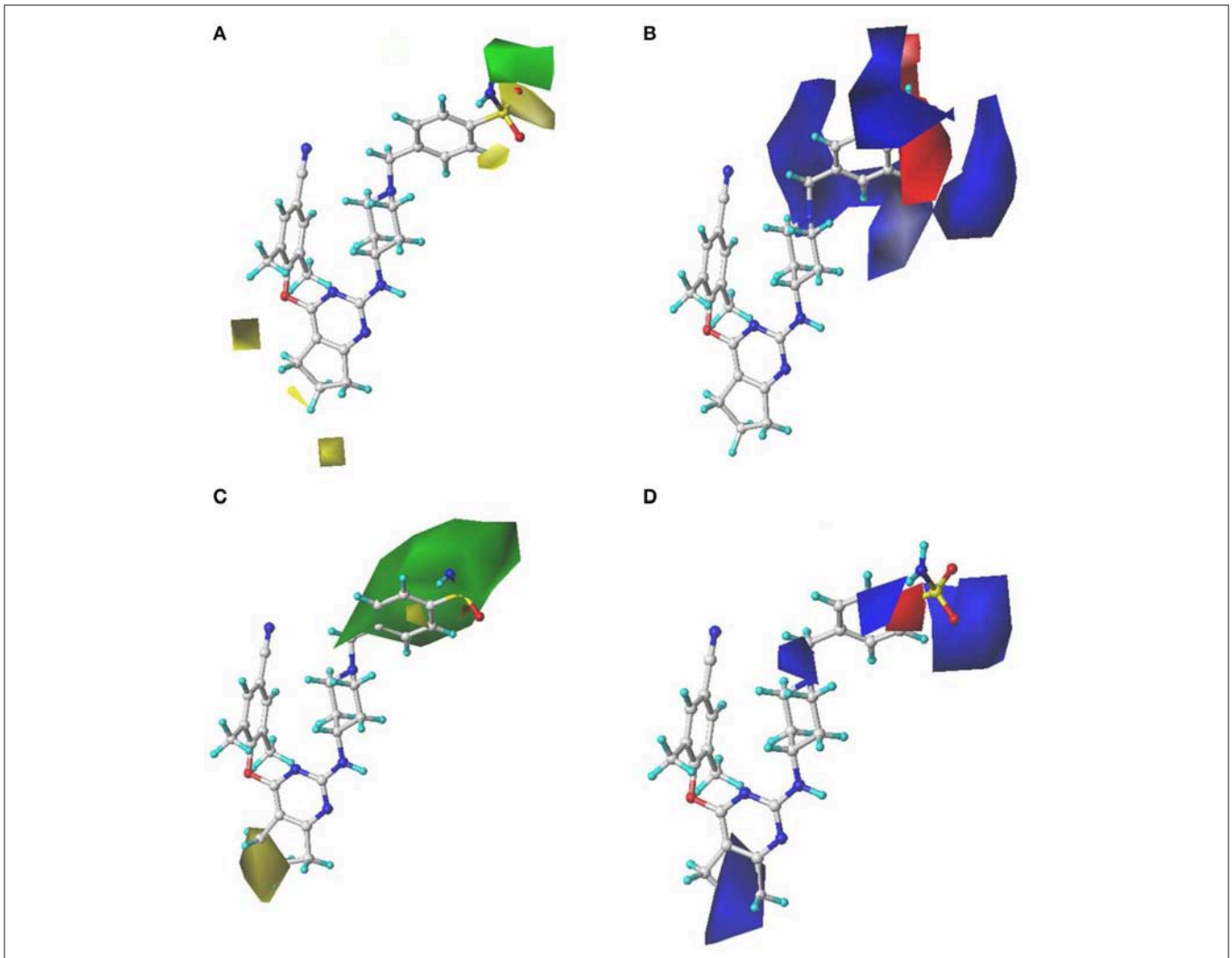


FIGURE 4 | Contour maps of steric and electrostatic fields with compound **36** as a reference in the comparative molecular field analysis (CoMFA) **(A,B)** and comparative molecular similarity indices analysis (CoMSIA) **(C,D)** models.

DHPYs. Compound **36** with the highest activity was utilized as a reference molecule to analyze the contour maps of both models. As shown in **Figure 2B**, the structure of compound **36** consisted of the common scaffold, Tolerant Regions I and II.

Figure 4 showed the S and E field contour maps of the CoMFA and CoMSIA models. In the S field, the green contour indicates that a bulky substituent at this position is beneficial for the activity, whereas a yellow block corresponds to a region where a small group is favorable for the activity. For the E field, a blue contour means that introduction of electropositive groups in this region may improve the biological activity, whereas the red contour indicates that electronegative groups may be beneficial for the activity. As can be seen from **Figures 4A,C**, the S field contours of the CoMFA model were consistent with those of the CoMSIA model. The yellow contour in the Tolerant Region I indicated that a relatively small group at this region would be beneficial for enhancing the activity, which might explain why the actual activities of compounds **30**–**41** were greater than those of compounds **25**–**29**. On the other hand, in the Tolerant Region II, there was a green contour at the terminal, suggesting that introduction of a bulky group was more favorable, which was in agreement with the activity orders: **18** (pyridine-4-yl-Ph) > **14** (4-SO_2NH_2-Ph) > **17** (4-SO_2CH_3-Ph), **19** (4-SO_2NH_2-Ph) > **21** (4-SO_2CH_3-Ph) > **23** (4-NO_2-Ph), and **42** (4-SO_2NH_2-Ph) > **44** (4-SO_2CH_3-Ph) > **46** (4-NO_2-Ph). At the *para-position* of the benzene ring of Tolerant Region II, two yellow contours indicated that small substituents here might be favorable for the activity, for instance, **3** (4-SO_2CH_3-Ph) > **2** (3-$CONH_2$-Ph) > **4** (pyridine-4-yl), **8** (4-SO_2CH_3-Ph) > **9** (pyridine-4-yl), **31** (4-$CONH_2$-Ph) > **33** (pyridine-4-yl). In **Figures 4B,D**, it can be clearly observed that a big blue contour was located at the terminal of Tolerant Region II, indicating that the positively charged group might be beneficial for the activity, such as **1** (4-SO_2NH_2-Ph) > **3** (4-SO_2CH_3-Ph), **15** (4-$CONH_2$-Ph) > **17** (4-SO_2CH_3-Ph), and **19** (4-SO_2NH_2-Ph) > **21** (4-SO_2CH_3-Ph). In addition, a red contour was located at the *para-position* of the benzene ring of Tolerant Region II, indicating that electronegative groups were beneficial for the antiviral activity at this position.

The H, D, and A field contour maps of the CoMSIA models were shown in **Figure 5**. In the H field, yellow contours represent the favorable zone of hydrophobic groups, whereas white contours show the unfavorable zone of hydrophobic groups. As shown in **Figure 5A**, a huge white near Tolerant Region I indicated that this place was appropriate to introduce hydrophobic groups. In addition, there was a white contour at the benzene ring of Tolerant Region II, which illustrated that hydrophobic substituents here were beneficial. The H field results were in good consistency with those of the previous study (Kang et al., 2019) that DHPYs with hydrophobic groups at corresponding positions exhibited promising activities. As for the D field, cyan suggests hydrogen-bond donor groups are useful for enhancing the activity, whereas purple is opposite. In **Figure 5B**, a cyan contour close to the linker atom of the pyrimidine ring and the right wing showed that the hydrogen-bond donor might be helpful for the activity at this position. There was also a cyan contour at the terminal of Tolerant Region II, indicating that hydrogen-bond donor groups were beneficial here, for example, **28** (4-NH_2-Ph) > **27** (4-NO_2-Ph). A purple contour near the para-position of the benzene ring of Tolerant Region II manifested that the place might not be suitable for hydrogen-bond donor groups, such as **1** (4-SO_2NH_2-Ph) > **2** (3-$CONH_2$-Ph). In the A field, beneficial and unbeneficial contour of hydrogen-bond acceptors are colored in magenta and red, respectively. In **Figure 5C**, a red contour at the terminal of Tolerant Region II signified that the hydrogen-bond acceptors at this position were disadvantageous for the activity, and two magenta contours at the para-position of the benzene ring of Tolerant Region II illustrated that the hydrogen-bond acceptor was advantageous. In short, introduction of hydrogen-bond acceptors at the para-position of the benzene ring of Tolerant Region II and hydrogen-bond donors at the terminal of Tolerant Region II might be advantageous for the inhibitory activity.

In a word, the contour maps of 3D-QSAR models presented that a small and/or hydrophobic group in Tolerant Region I;

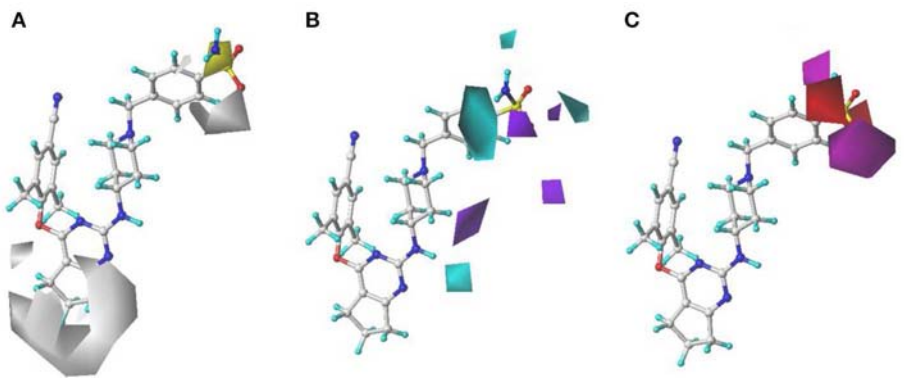


FIGURE 5 | Contour maps of hydrophobic **(A)**, hydrogen-bond donor **(B)**, and hydrogen-bond acceptor **(C)** fields in the optimal comparative molecular similarity indices analysis (CoMSIA) model.

a small, electronegative and/or hydrogen-bond accepter group at the *para-position* of the benzene ring of Tolerant Region II; and/or a bulky, electropositive and/or hydrogen-bond donor group at the terminal of Tolerant Region II would be favorable for increasing the activity, respectively.

Pharmacophore Model

The statistical parameters of 20 pharmacophore models generated by GALAHAD were listed in **Table S2**. As for pharmacophore models, it could be served as the query for a UNITY flex search only if the SPECIFITY value was more than 5. The identical value of the PARETO column indicated that all models were statistically equivalent. In general, a good pharmacophore model should have small ENERGY and high SPECIFITY, N_HITS, STERICS, and MOL_QRY (Caballero, 2010). Among 20 models, model_20 was regarded as the optimal model by the comprehensive consideration of the abovementioned parameters.

The pharmacophore features of the best GALAHAD Model_20 were displayed in **Figure 6**, including three hydrophobic centers (HYs, cyan), four hydrogen-bond acceptor atoms (AAs, green), and one hydrogen-bond donor atom (DAs, magenta). All features were located in the left and middle structures of DHPYs. One of the hydrogen-bond acceptor atom at the connecting atom of the left ring indicated that hydrogen-bond acceptor groups might increase the inhibitory activities at this position, which was consistent with our previous study (Wan et al., 2018). The other hydrogen-bond acceptor atoms were located at the nitrogen atoms of the pyrimidine ring and the cyano group of the left benzene ring. Moreover, the hydrophobic center of the left phenyl ring was located at the hydrophobic pocket of the HIV-1 RT, which was also in good consistency with our previous studies (Wan et al., 2018). The right linker atom was the hydrogen-bond donor atom, which suggested that the hydrogen-bond donor atom at this position was likely to improve the anti-HIV-1 activities, which was in good agreement with the 3D-QSAR results.

For the optimal pharmacophore, there were 70 compounds screened from the decoy database, and 42 of them were active molecules. In addition, the calculated values of %Y, %A, EF, and GH were 60%, 100%, 89.66, and 0.70, respectively, which met the requirements that the EF value should be more than 1 and the GH value should be in the range from 0.6 and 1. These statistical results indicated that model_20 had excellent abilities of recognizing the false positives and distinguishing the

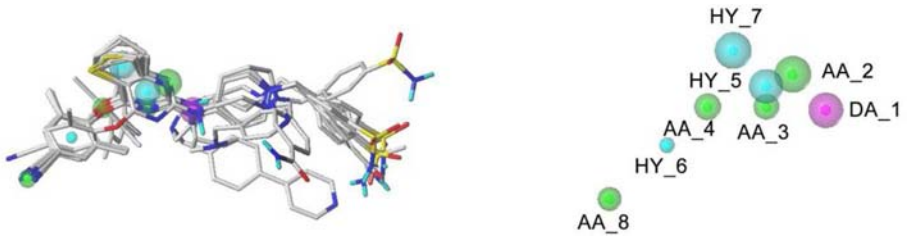


FIGURE 6 | The best pharmacophore model with the alignment of 10 training set compounds. The model includes four hydrogen-bond acceptor atoms (green), three hydrophobic centers (cyan), and one hydrogen-bond donor atom (magenta).

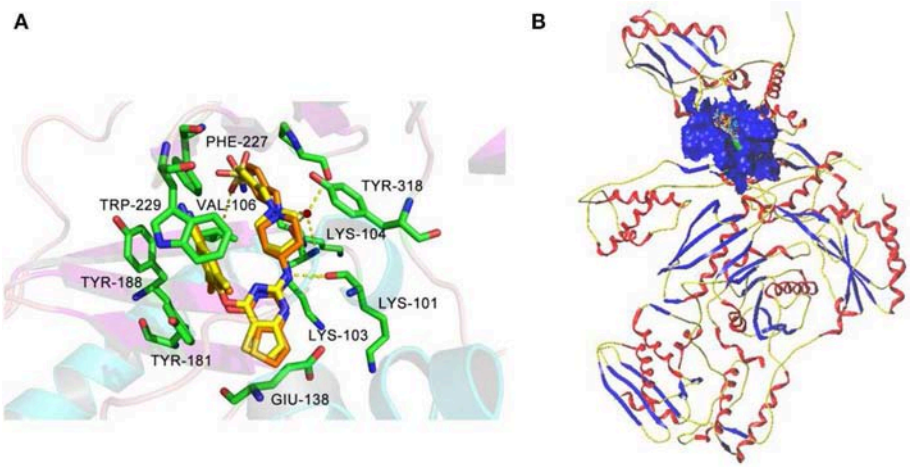


FIGURE 7 | The redocked results of **K-5a2** in the binding pocket of the HIV-1 reverse transcriptase (RT) (PDB: 6C0J). **(A)** The superimposition of the cognate **K-5a2** (yellow) and the redocked **K-5a2** (orange). **(B)** The blue region represents the surface of the binding pocket.

similar structures of active and inactive compounds from the database. Thus, model_20 could be used for the next virtual screening studies.

Molecular Docking

Molecular docking was performed to investigate the binding modes of DHPYs at the active site of the HIV-1 RT. To validate the reliability of the molecular docking method, the cognate ligand (**K-5a2**) was redocked into the binding pocket of the HIV-1 RT (PDB: 6C0J), and the result was shown in **Figure 7A**. The original crystallographic and redocked conformations were almost superposition, and the root mean square deviation (RMSD) value between them for all atoms was 0.38 Å, which suggested that the docking method and used parameters were reasonable (Khan et al., 2010). As seen from **Figure 7A**, the two ligands adopted a similar binding pattern, in which the left benzene ring was located at the hydrophobic region consisting of residues Tyr181, Tyr188, Trp229, Phe227, and Val106 and could form π-π stacking interactions with the aromatic residues of them. In addition, it was noteworthy that the two ligands not only formed hydrogen-bond interactions with residues Lys101, Lys104, and Val106, respectively, but also interacted with residues Lys103 and Pro236 *via* a network of hydrogen bonds by a water molecule (W936). Those results were in good consistency with previous reports (Yang et al., 2018; Kang et al., 2019). At the same time, the hydrogen bond formed between the C = O of Lys101 and the NH of the right linker atom indicated that hydrogen-bond donor atoms were beneficial in the place, which was a good agreement with the results of pharmacophore and 3D-QSAR models. As shown in **Figure 7B**, 52 DHPYs embedded in the binging pocket by the similar *U*-shaped conformations, suggesting the accuracy of the docking method.

Virtual Screening

To discover and design novel HIV-1 NNRTI leads, a multistage-filtered virtual screening was performed based on the constructed pharmacophore model and the established molecular docking method (**Figure 8**). First, a total of 19,740 compounds were obtained from ZINC database by the pharmacophore-based virtual screening and the restriction with Lipinski's rule of five. Then, 3,451 compounds were selected on the basis of the QFIT score of more than 50. In order to enhance the efficiency and accuracy of docking screening, the preliminary docking by Surflex-Dock and the second round docking by Surflex-Dock Geom were performed. The results indicated that only 20 compounds met the requirements simultaneously. In view of the predicted ADME properties of the screened 20 compounds, nine compounds were selected to regard as NNRTI hits, whose structures and docking scores were shown in **Table 3**. Furthermore, the interactions between the screened compounds and the HIV-1 RT were shown in **Table S3**. Nine screened compounds formed hydrophobic interactions with residues Tyr181, Tyr188, Phe227, Trp229, and Val106 and π-π stacking interactions with the aromatic residues of them. Except for ZINC_91409938, which formed a hydrogen-bond network with the residues Pro236 and Lys103 by a water molecule (W936), the screened compounds also formed hydrogen bonds with the key residues Lys101 and Glu138. The docking results indicated that nine screened compounds might be potential NNRTIs.

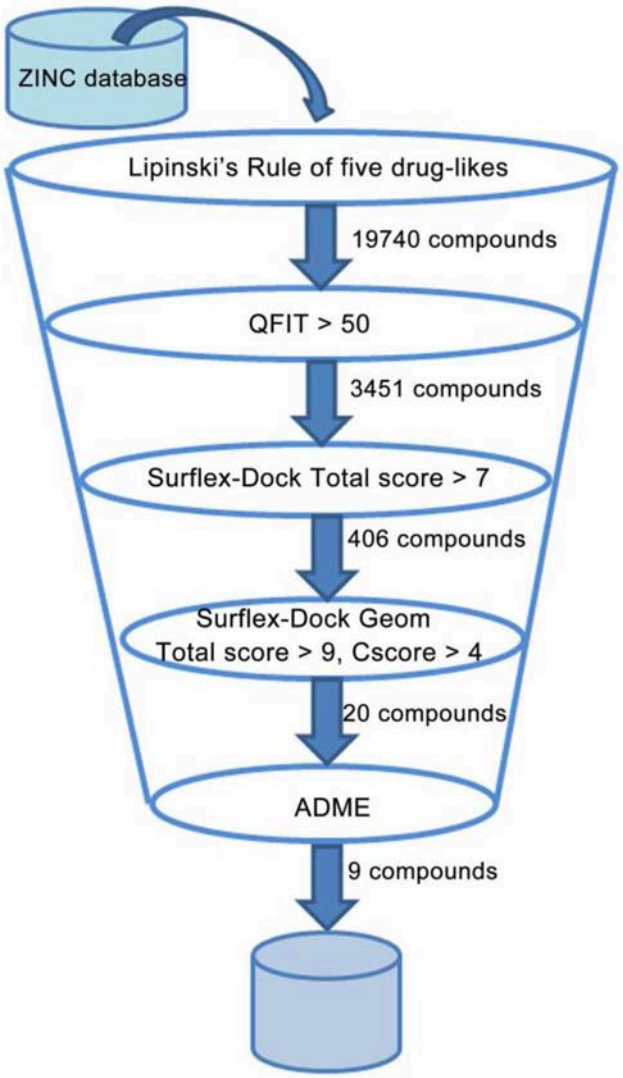
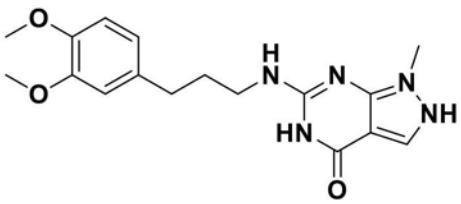
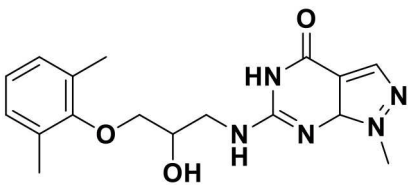
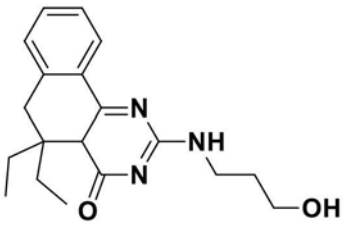
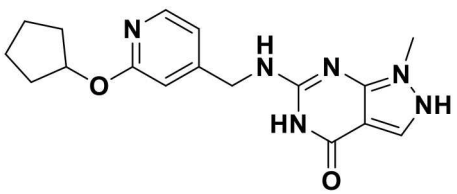
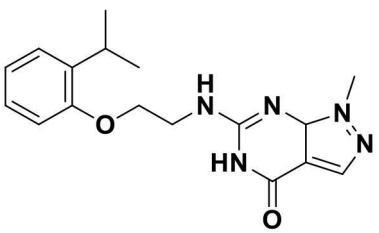
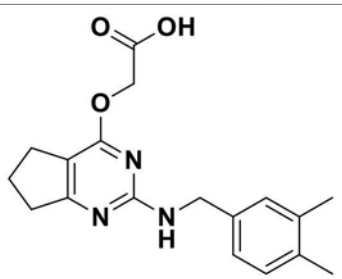
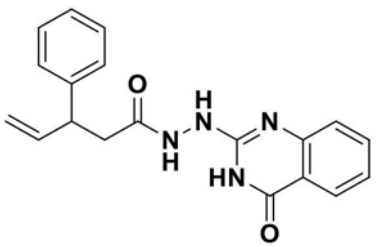


FIGURE 8 | Representation of the overall virtual screening process.

Newly Designed Non-nucleoside Reverse Transcriptase Inhibitors

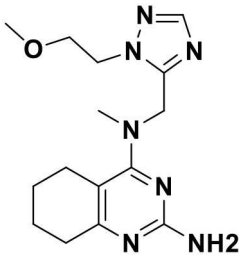
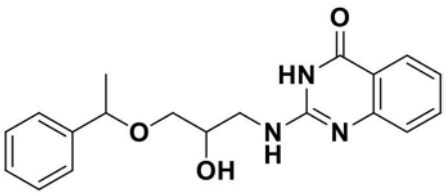
According to the structural characteristics of DHPYs and the results of the 3D-QSAR models and molecular docking, we further designed three new compounds (**N1**, **N2,** and **N3**; **Table 4**) using ZINC_73709240 as a lead compound. The 3D-QSAR contour maps indicated that the hydrogen-bond acceptor at the *para-position* of the benzene ring of Tolerant Region II and the hydrogen-bond donor at the terminal of Tolerant Region II were favorable to the inhibitory activity. Therefore, we designed compounds **N1**, **N2,** and **N3** by adding amide or

TABLE 3 | Chemical structures and docking scores of the screened hit compounds as novel HIV-1 NNRTIs from ZINC database.

Compound No.	Structure	Docking score
ZINC_57841658		9.43
ZINC_60381334		9.02
ZINC_63070905		9.56
ZINC_69532225		9.12
ZINC_71894576		9.00
ZINC_73709240		9.64
ZINC_89506228		9.29

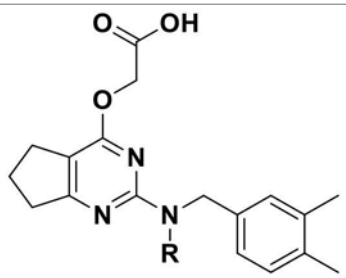
(Continued)

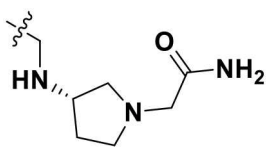
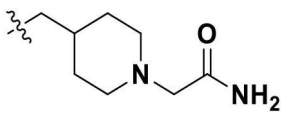
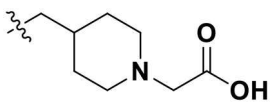
TABLE 3 | Continued

Compound No.	Structure	Docking score
ZINC_91409938		9.30
ZINC_97995063		9.22

NNRTIs, non-nucleoside reverse transcriptase inhibitors.

TABLE 4 | Chemical structures and docking scores of the newly designed HIV-1 NNRTIs.



Compound No.	*R*	Docking score	
		Wild-type HIV-1 RT	Mutant HIV-1 RT (K103N+Y181C)
N1		13.83	10.60
N2		12.59	12.03
N3		12.93	12.88

NNRTIs, non-nucleoside reverse transcriptase inhibitors; RT, reverse transcriptase.

carboxyl groups as hydrogen-bond donors or acceptors at these positions (**Table 4**).

All designed compounds were then docked into the binding site of HIV-1 RT by Surflex-Dock Geom method. The docking scores of compounds **N1**, **N2,** and **N3** were 13.83, 12.59, and 12.93, respectively, and higher than that of compound **36** (11.86), suggesting that the interactions between the newly designed compounds and the protein might be more stable. As shown

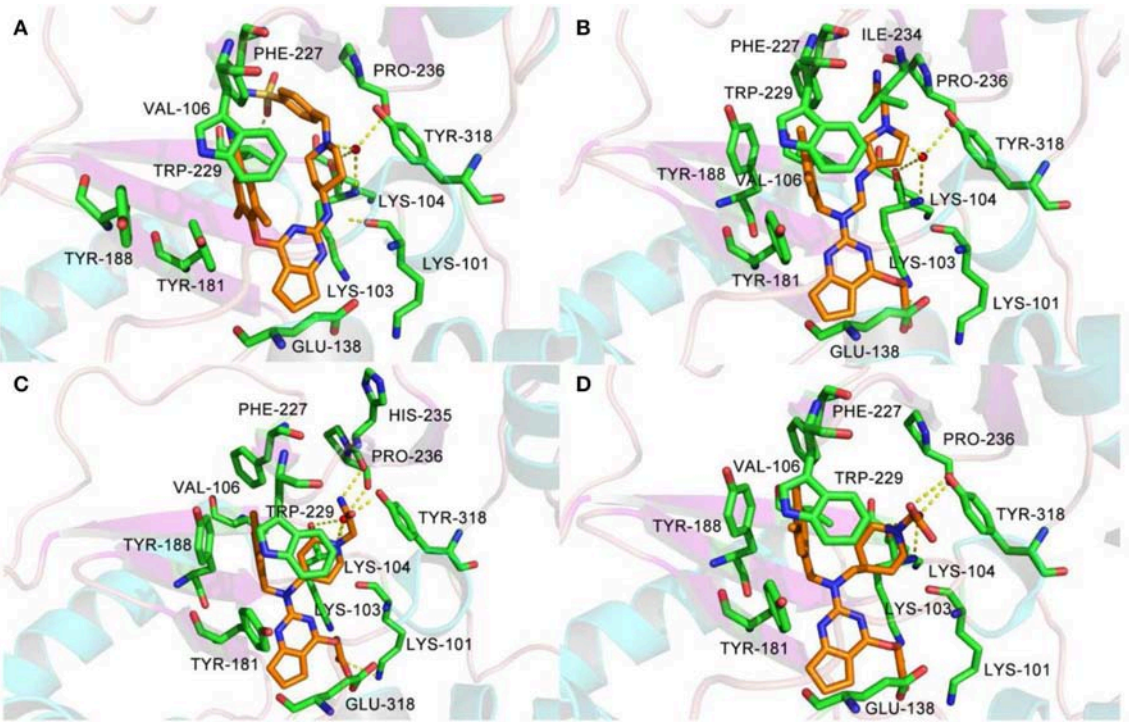


FIGURE 9 | The docked results of compounds **36 (A)**, **N1 (B)**, **N2 (C)**, and **N3 (D)** in the binding pocket of wild-type HIV-1 reverse transcriptase (RT) (PDB: 6C0J).

in **Figure 9**, the binding modes of compounds **N1**, **N2,** and **N3** with the protein were basically similar to that of compound **36**. The left wings of four ligands were all located in the hydrophobic region and formed π-π stacking interactions with residues Tyr181, Tyr188, Trp229, and Phe227, and the positive nitrogen of their right wing formed hydrogen-bond networks with Lys103 and Pro236 through a water molecule (W936), which was in good agreement with the docking results. However, there were some differences for four compounds in terms of protein–ligand interactions. As for compound **36**, it formed three hydrogen bonds with Lys101 (Lys101-O…H-N-, 2.8 Å) and Val106 (Val106-NH…O=C, 3.0 Å; -O…H-N-, 2.7 Å), which was consistent with the redocked results of **K-5a2**. As can be seen from **Figures 9A,B**, compounds **N1** and **N3** not only formed hydrogen bonds with Glu138 and Lys103 but also had hydrogen bonds with Ile234 and/or Tyr318. Another finding was that five hydrogen bonds were formed between residues Glu138, Lys103, Lys101, Tyr318, and His235 with compound **N2**. The docking results revealed that the four compounds interacted with key amino acid residues (Lys101 and Glu138), and several new hydrogen bonds between three newly designed compounds and residues Lys103, Ile234, Tyr318, and His235 were found. These results suggested that compounds **N1**, **N2,** and **N3** might be the potential inhibitors with improving anti-HIV-1 activities.

To further explore whether the newly designed compounds could inhibit mutant HIV-1 RT, they were also docked into the mutant (K103N+Y181C) RT (PDB ID: 6C0R) (**Figure S1**; **Table 4**). The co-crystallizing ligand (**25a**) of 6C0R as a reference compound was also redocked into the binding site as displayed in **Figure S1A**. Kang et al. (2017) reported that the inhibitory activity of compound **25a** (EC_{50} = 5.5 ± 0.81 nM) against the K103N+Y181C mutant RT was better than that of RPV (EC_{50} = 11 ± 1.9 nM). Our docking results indicated that compound **25a** formed four hydrogen bonds with residues Lys101, Val106, Lys104, and Tyr188, respectively, and the π-π stacking and/or hydrophobic interactions were also found with residues Trp229, Phe227, and Tyr188. In addition, the residue Tyr183 played an important role in the binding site of the mutant RT and could offset the loss of π-stacking and hydrophobic interactions between inhibitors and residue Tyr181 as it was mutated to Cys181 (Das et al., 2008).

The docking scores of three hit compounds were relatively high (**Table 4**), especially compounds **N2** and **N3**, whose docking scores were higher than **25a** (12.02), indicating that the newly designed molecules might have better inhibitory activity against the mutant RT. It was observed that the hydrogen bond with residue Lys101, π-π stacking, and hydrophobic interactions still existed for three complexes (**Figure S1**). However, the difference was that compounds **N1**, **N2**, and **N3** could form a direct hydrogen bond with the mutated residue Asn103, which indicated that these designed molecules could bind well in the binding pocket with mutations. These docking results demonstrated that the three hit compounds might have the ability to inhibit the HIV-1 RT mutant. However, the actual anti-HIV activities of the three hits are necessary to be identified in future studies.

TABLE 5 | Predicted absorption-distribution-metabolism- excretion (ADME) parameters and drug-like properties of compound **36** and the newly designed inhibitors (**N1–3**).

Properties	Parameters	Compounds			
		36	**N1**	**N2**	**N3**
Physicochemical	MW[a] (g/mol)	532.66	482.58	481.59	482.57
Properties	Rotatable bonds	7	11	10	10
	HBA	8	8	7	8
	HBD	2	3	2	2
	TPSA[b]	142.61	133.91	121.88	116.09
Lipophilicity	iLOGP	3.82	3.29	3.38	3.60
	XLOGP3	4.17	0.04	0.92	1.57
	WLOGP	4.33	0.63	1.72	2.32
	MLOGP	2.06	0.44	1.42	1.82
	Silicos IT logP	3.51	2.08	3.09	3.33
	Consensus logP	3.58	1.30	2.11	2.53
Water Solubility	ESOL Class	MS[c]	S[d]	S	S
	Ali Class	PS[e]	S	S	S
	Silicos IT Class	PS	MS	MS	MS
Pharmacokinetics	GI[f] absorption	low	high	high	high
	BBB[g] permeat	No	No	No	No
	CYP1A2 inhibitor	No	No	No	No
	CYP2C19 inhibitor	Yes	No	No	No
	CYP2C9 inhibitor	Yes	No	No	No
	CYP2D6 inhibitor	Yes	No	Yes	Yes
	CYP3A4 inhibitor	Yes	No	Yes	Yes
Druglikeness	Lipinski violations	1	0	0	0
	Ghose violations	2	2	2	2
	Egan violations	1	1	0	0
	Muegge violations	1	1	0	0
	Bioavailability Score	0.55	0.55	0.55	0.56
Medicinal Chemistry	PAINS[h] alerts	0	0	0	0
	Brenk alerts	0	1	0	0
	Leadlikeness violations	2	2	1	2
	Synthetic accessibility	4.68	4.68	4.12	4.10

[a] *Molecular weight.*
[b] *Total polar surface area.*
[c] *Moderately soluble.*
[d] *Soluble.*
[e] *Poorly soluble.*
[f] *Gastrointestinal.*
[g] *Blood–brain barrier.*
[h] *Pan assay interference compounds.*

ADME Analysis

ADME prediction studies were carried out for compound **36** and three newly designed NNRTIs (**N1**, **N2,** and **N3**). The results were depicted in **Table 5**. In this program, five inhibitors of cytochrome P450 (CYP) enzymes were predicted. CYPs, which primarily mediated oxidation of various compounds and participated in physiological and pathophysiological processes, were the major phase I drug-metabolizing enzymes and responsible for metabolism of about 75% of all marketed drugs (Moroy et al., 2012). In the family of CYP enzymes, the CYP3A4 was the most important enzyme on account of metabolizing ∼50% of all drugs by itself, and the CYP2C9 enzyme mainly metabolizes several clinically used drugs such as celecoxib and diclofenac (Daly et al., 2017). As shown in **Table 5**, compounds **N1**, **N2,** and **N3** could be easier to be metabolized compared with compound **36**. In addition, three newly designed compounds showed high human gastrointestinal absorption (HIA), indicating that they might have a high chance of brain penetration (Li et al., 2019). The topological polar surface area (TPSA) values of compound **N1** and **N3** were in the range from 20 to 130 $Å^2$, which suggested that they possessed good transport properties *in vivo*. Notably, the synthetic accessibilities of designed compounds were lower than 5, suggesting that they were relatively easy to be synthesized. On the whole, the

ADME properties of the three newly designed compounds were superior to that of compound **36**, especially in pharmacokinetics, druglikeness, and medicinal chemistry properties. Thus, the three newly designed compounds might be supposed to have good pharmacokinetics properties.

Molecular Dynamics Simulation

As for newly designed molecules, their stability of protein–ligand interactions should be taken into account. Thus, 10 ns MD simulations were performed for four complex systems, 6C0J-36, 6C0J-N1, 6C0J-N2, and 6C0J-N3, respectively. The RMSD values of backbone atoms for the four complexes were displayed in **Figure 10A**. During the 10 ns MD simulations, the RMSD values of the four systems were relatively stable and were lower than 0.3 nm. **Figure 10B** showed the RMSD values of the four ligands during 10 ns MD simulations. The four ligands had similar fluctuations and reached equilibrium at approximately 0.5 ns. The root mean square fluctuation (RMSF) profiles of the four complexes (**Figures 10C,D**) also exhibited similar trends during the MD simulations. It should be pointed out that the key residues, Lys101 of chain A and Glu138 of chain B, had relatively lower RMSF values. As shown in **Figure 10E**, the radius of gyration (Rg) values, which could explain the compactness of the protein throughout simulation, basically maintained at about 3.5 nm, indicated that greater changes of the conformations of protein did not take place. In addition,

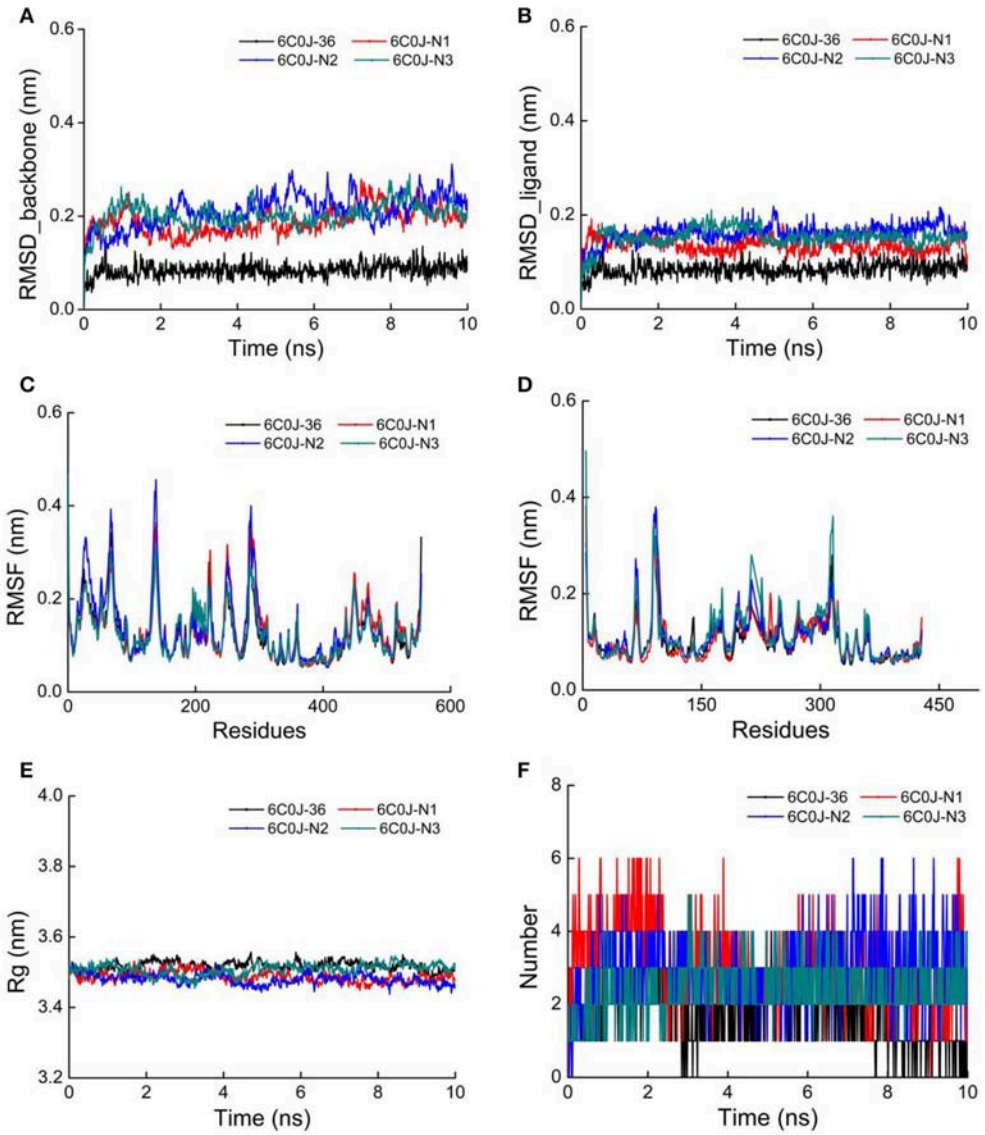


FIGURE 10 | The 10 ns molecular dynamics (MD) results of compounds **36**, **N1**, **N2**, and **N3** in wild-type HIV-1 reverse transcriptase (RT). **(A)** Root mean square deviation (RMSD) values of backbone atoms of the protein. **(B)** RMSD values of the ligands. **(C)** Root mean square fluctuation (RMSF) values of the chain A. **(D)** RMSF values of the chain B. **(E)** Radius of gyration (Rg) values of backbone atoms. **(F)** The total number of hydrogen bonds between the ligands and the protein.

the intermolecular hydrogen bonds could be used to analyze the protein–ligand interaction. As shown in **Figure 10F**, the hydrogen-bond numbers of 6C0J-36, 6C0J-N1, 6C0J-N2, and 6C0J-N3 complexes were 1-2, 1-6, 1-6, and 2-5 over the 10 ns simulations, respectively, which suggested that all newly designed compounds might be more stable than compound **36**. The MD simulation results revealed that four protein–ligand complexes could maintain a relative stability in the dynamic simulation and three newly designed compounds might have more interactions with the HIV-1 RT than compound **36**. These were in good consistency with the docking results.

In the same pattern, 10 ns MD simulations were also carried out for three protein–ligand complexes (6C0R-N1, 6C0R-N2, and 6C0R-N3) to further study whether they still could remain stable in the dynamic environment. The results were shown in **Figure S2**. The RMSD values of protein backbones of the three complexes were displayed in **Figure S2A**, and it can be clearly seen that they basically reached stability after 5 ns and were below 0.4 nm. The RMSD values of the three ligands were also stable (**Figure S2B**). **Figures S2C,D** were the RMSF plots of chains A and B, respectively, which showed that the residues of the three complexes fluctuated in the same trend, indicating that they had great stability. In addition, the Rg values just slightly floated within 3.5 nm from **Figure S2E**, indicating that the proteins had good compactness. The number of hydrogen bonds was also essential to verify the stability. As shown in **Figure S2F**, the hydrogen-bond numbers of compounds **N1**, **N2**, and **N3** were 2-4, 1-5, and 2-5 over the 10 ns MD, respectively, suggesting that the three compounds could tightly bind to the mutant RT. The abovementioned results revealed that the three complexes could keep stable during MD simulations and the three designed compounds could interact well with the mutant HIV-1 RT. However, the experimental activities of the three new hits against wild-type and mutant HIV-1 strains remain to be studied.

CONCLUSION

In conclusion, 52 DHPYs were collected to construct the CoMFA and CoMSIA models, which exhibited rationally statistical parameters and good predictive ability. These models well-explained the 3D-QSARs of these DHPY and provided useful information for designing new HIV-1 NNRTIs. The optimal pharmacophore model containing eight features was in agreement with the 3D-QSAR results. The docking results revealed that Lys101 was the key amino acid residue, and the hydrophobic and π-π stacking interactions with Tyr181, Tyr188, Trp229, and Phe227 also played key roles for the anti-HIV activity of DHPYs. Nine lead compounds were obtained by the pharmacophore-based and docking-based virtual screening as well as ADME prediction. Three novel inhibitors were designed by modifying the structure of the screened compound ZINC_73709240 according to the 3D-QSAR and docking results. Three newly designed inhibitors showed good stability and strong interactions not only in the wild-type RT but also in the K103N/Y181C RT mutant based on the docking and MD simulation results. The ADME prediction indicated that compounds **N1**, **N2,** and **N3** might possess desirable drug-like properties. However, further study on synthesis and anti-HIV activities of the three newly designed hits is necessary. We expect that the screened and designed compounds could be served as lead candidates of novel HIV-1 NNRTIs.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

GL and YC proposed the research idea and designed the experiment. YC performed the experiment. YC, FW, YT, and YG analyzed the data. YC, XL, XJ, and GL wrote the manuscript. All authors revised and approved the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 21807082), the Hubei Provincial Natural Science Foundation of China (No. 2017CFB121), and the Hubei Provincial Department of Education of China (No. Q20171503).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2020.00164/full#supplementary-material

REFERENCES

Almerico, A. M., Lauria, A., Tutone, M., Diana, P., Barraja, P., Montalbano, A., et al. (2003). A multivariate analysis on non-nucleoside HIV-1 reverse transcriptase inhibitors and resistance induced by mutation. *QSAR Comb. Sci.* 22, 984–996. doi: 10.1002/qsar.200330834

Almerico, A. M., Tutone, M., Ippolito, M., and Lauria, A. (2007). Molecular modelling and QSAR in the discovery of HIV-1 integrase inhibitors. *Curr. Comput. Aid. Drug Des.* 3, 214–233. doi: 10.2174/157340907781695468

Almerico, A. M., Tutone, M., and Lauria, A. (2008). Docking and multivariate methods to explore HIV-1 drug-resistance: a comparative analysis. *J. Comput. Aid. Mol. Des.* 22, 287–297. doi: 10.1007/s10822-008-9186-7

Almerico, A. M., Tutone, M., Lauria, A., Diana, P., Barraja, P., Montalbano, A., et al. (2006). A multivariate analysis of HIV-1 protease inhibitors and resistance induced by mutation. *J. Chem. Inf. Model.* 46, 168–179. doi: 10.1021/ci050139z

Borisa, A., and Bhatt, H. (2015). 3D-QSAR (CoMFA, CoMFA-RG, CoMSIA) and molecular docking study of thienopyrimidine and thienopyridine derivatives to explore structural requirements for aurora-B kinase inhibition. *Eur. J. Pharm. Sci.* 79, 1–12. doi: 10.1016/j.ejps.2015.08.017

Bush, B. L., and Nachbar, R. B. (1993). Sample-distance partial least squares: PLS optimized for many variables, with application to CoMFA. *J. Comput. Aid. Mol. Des.* 7, 587–619. doi: 10.1007/BF00124364

Caballero, J. (2010). 3D-QSAR (CoMFA and CoMSIA) and pharmacophore (GALAHAD) studies on the differential inhibition of aldose reductase by flavonoid compounds. *J. Mol. Graph. Model.* 29, 363–371. doi: 10.1016/j.jmgm.2010.08.005

Chen, X. W., Zhan, P., Li, D. Y., Clercq, E. D., and Liu, X. Y. (2011). Recent advances in DAPYs and related analogues as HIV-1 NNRTIs. *Curr. Med. Chem.* 18, 329–376. doi: 10.2174/092986711794839142

Daina, A., Michielin, O., and Zoete, V. (2017). SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* 7:42717. doi: 10.1038/srep42717

Daly, A. K., Rettie, A. E., Fowler, D. M., and Miners, J. O. (2017). Pharmacogenomics of CYP2C9: functional and clinical considerations. *J. Pers. Med.* 8:1. doi: 10.3390/jpm8010001

Das, K., Bauman, J. D., Clark, A. D., Frenkel, Y. V., Lewi, P. J., Shatkin, A. J., et al. (2008). High-resolution structures of HIV-1 reverse transcriptase/TMC278 complexes: strategic flexibility explains potency against resistance mutations. *PNAS.* 105, 1466–1471. doi: 10.1073/pnas.0711209105

Esposito, F., Corona, A., and Tramontano, E. (2012). HIV-1 reverse transcriptase still remains a new drug target: structure, function, classical inhibitors, and new inhibitors with innovative mechanisms of actions. *Mol. Biol. Int.* 2012:586401. doi: 10.1155/2012/586401

Gao, Y., Chen, Y. M., Tian, Y. F., Zhao, Y. L., Wu, F. S., Luo, X. G., et al. (2019). *In silico* study of 3-hydroxypyrimidine-2,4-diones as inhibitors of HIV RT-associated RNase H using molecular docking, molecular dynamics, 3D-QSAR, and pharmacophore models. *New J. Chem.* 43, 17004–170017. doi: 10.1039/C9NJ03353J

Gu, S. X., Xiao, T., Zhu, Y. Y., Liu, G. Y., and Chen, F. E. (2019). Recent progress in HIV-1 inhibitors targeting the entrance channel of HIV-1 non-nucleoside reverse transcriptase inhibitor binding pocket. *Eur. J. Med. Chem.* 174, 277–291. doi: 10.1016/j.ejmech.2019.04.054

Kalva, S., Azhagiya Singam, E. R., Rajapandian, V., Saleena, L. M., and Subramanian, V. (2014). Discovery of potent inhibitor for matrix metalloproteinase-9 by pharmacophore based modeling and dynamics simulation studies. *J. Mol. Graph. Model.* 49, 25–37. doi: 10.1016/j.jmgm.2013.12.008

Kang, D. W., Fang, Z. J., Huang, B. S., Lu, X. Y., Zhang, H., Xu, H., et al. (2017). Structure-based optimization of thiophene[3,2-d]pyrimidine derivatives as potent HIV-1 non-nucleoside reverse transcriptase inhibitors with improved potency against resistance-associated variants. *J. Med. Chem.* 60, 4424–4443. doi: 10.1021/acs.jmedchem.7b00332

Kang, D. W., Fang, Z. J., Li, Z. Y., Huang, B. S., Zhang, H., Lu, X. Y., et al. (2016). Design, synthesis, and evaluation of thiophene[3,2-d]pyrimidine derivatives as HIV-1 non-nucleoside reverse transcriptase inhibitors with significantly improved drug resistance profiles. *J. Med. Chem.* 59, 7991–8007. doi: 10.1021/acs.jmedchem.6b00738

Kang, D. W., Zhang, H., Wang, Z., Zhao, T., Ginex, T., Luque, F. J., et al. (2019). Identification of dihydrofuro[3,4-d]pyrimidine derivatives as Novel HIV-1 non-nucleoside reverse transcriptase inhibitors with promising antiviral activities and desirable physicochemical properties. *J. Med. Chem.* 62, 1484–1501. doi: 10.1021/acs.jmedchem.8b01656

Khan, K. M., Wadood, A., Ali, M., Zia, U., Ul-Haq, Z., Lodhi, M. A., et al. (2010). Identification of potent urease inhibitors via ligand- and structure-based virtual screening and *in vitro* assays. *J. Mol. Graph. Model.* 28, 792–798. doi: 10.1016/j.jmgm.2010.02.004

Li, M., Wei, D., Zhao, H., and Du, Y. (2014). Genotoxicity of quinolones: substituents contribution and transformation products QSAR evaluation using 2D and 3D models. *Chemosphere* 95, 220–226. doi: 10.1016/j.chemosphere.2013.09.002

Li, Q., Zhang, C., and Ren, Y. (2019). Molecular modeling technology studies of novel pyrazoylethylbenzamide derivatives as selective orexin receptor 1 antagonists. *J. Taiwan Inst. Chem. Eng.* 100, 1–17. doi: 10.1016/j.jtice.2019.03.018

Liu, G. Y., Wan, Y. L., Wang, W. J., Fang, S., Gu, S. X., and Ju, X. L. (2019). Docking-based 3D-QSAR and pharmacophore studies on diarylpyrimidines as non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Mol. Divers.* 23, 107–121. doi: 10.1007/s11030-018-9860-1

Liu, G. Y., Wang, W. J., Wan, Y. L., Ju, X. L., and Gu, S. X. (2018). Application of 3D-QSAR, pharmacophore, and molecular docking in the molecular design of diarylpyrimidine derivatives as HIV-1 nonnucleoside reverse transcriptase inhibitors. *Int. J. Mol. Sci.* 19:1436. doi: 10.3390/ijms19051436

Moroy, G., Martiny, V. Y., Vayer, P., Villoutreix, B. O., and Miteva, M. A. (2012). Toward *in silico* structure-based ADMET prediction in drug discovery. *Drug Discov. Today*. 17, 44–55. doi: 10.1016/j.drudis.2011.10.023

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55, 6582–6594. doi: 10.1021/jm300687e

Namasivayam, V., Vanangamudi, M., Kramer, V. G., Kurup, S., Zhan, P., Liu, X., et al. (2019). The journey of HIV-1 non-nucleoside reverse transcriptase inhibitors (NNRTIs) from lab to clinic. *J. Med. Chem.* 62, 4851–4883. doi: 10.1021/acs.jmedchem.8b00843

Ojha, P. K., Mitra, I., Das, R. N., and Roy, K. (2011). Further exploring r^2_m metrics for validation of QSPR models. *Chemom. Intell. Lab. Syst.* 107, 194–205. doi: 10.1016/j.chemolab.2011.03.011

Roy, K., Das, R. N., Ambure, P., and Aher, R. B. (2016). Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemom. Intell. Lab. Syst.* 152, 18–33. doi: 10.1016/j.chemolab.2016.01.008

Vasavi, C. S., Tamizhselvi, R., and Munusami, P. (2019). Exploring the drug resistance mechanism of active site, non-active site mutations and their cooperative effects in CRF01_AE HIV-1 protease: molecular dynamics simulations and free energy calculations. *J. Biomol. Struct. Dyn.* 37, 2608–2626. doi: 10.1080/07391102.2018.1492459

Wan, Y. L., Tian, Y. F., Wang, W. J., Gu, S. X., Ju, X. L., and Liu, G. Y. (2018). *In silico* studies of diarylpyridine derivatives as novel HIV-1 NNRTIs using docking-based 3D-QSAR, molecular dynamics, and pharmacophore modeling approaches. *RSC Adv.* 8, 40529–40543. doi: 10.1039/C8RA06475J

Wang, R., Xu, K., and Shi, W. (2019). Quinolone derivatives: potential anti-HIV agent-development and application. *Arch. Pharm.* 352:e1900045. doi: 10.1002/ardp.201900045

Wang, W. J., Tian, Y. F., Wan, Y. L., Gu, S. X., Ju, X. L., Luo, X. G., et al. (2018). Insights into the key structural features of N1-arybenzimidazols as HIV-1 NNRTIs using molecular docking, molecular dynamics, 3D-QSAR, and pharmacophore modeling. *Struct. Chem.* 30, 385–397. doi: 10.1007/s11224-018-1204-3

Yang, Y., Kang, D. W., Nguyen, L. A., Smithline, Z. B., Pannecouque, C., Zhan, P., et al. (2018). Structural basis for potent and broad inhibition of HIV-1 RT by thiophene[3,2-d]pyrimidine non-nucleoside inhibitors. *Elife* 7:e36340. doi: 10.7554/eLife.36340

Zhan, P., Liu, X. Y., and Li, Z. Y. (2009). Recent advances in the discovery and development of novel HIV-1 NNRTI platforms: 2006-2008 update. *Curr. Med. Chem.* 16, 2876–2889. doi: 10.2174/092986709788803231

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Chen, Tian, Gao, Wu, Luo, Ju and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 03 April 2020
doi: 10.3389/fchem.2020.00235

Competition Between Phenothiazines and BH3 Peptide for the Binding Site of the Antiapoptotic BCL-2 Protein

Aline Lagoeiro do Carmo [1], Fernanda Bettanin [2], Michell Oliveira Almeida [3], Simone Queiroz Pantaleão [1], Tiago Rodrigues [1], Paula Homem-de-Mello [1] and Kathia Maria Honorio [1,2]**

[1] Centro de Ciências Naturais e Humanas, Universidade Federal do ABC, Santo André, Brazil, [2] Escola de Artes, Ciências e Humanidades, Universidade de São Paulo (USP), São Paulo, Brazil, [3] Instituto de Química de São Carlos, Universidade de São Paulo (USP), São Paulo, Brazil

The study of proteins and mechanisms involved in the apoptosis and new knowledge about cancer's biology are essential for planning new drugs. Tumor cells develop several strategies to gain proliferative advantages, including molecular alterations to evade from apoptosis. Failures in apoptosis could contribute to cancer pathogenesis, since these defects can cause the accumulation of dividing cells and do not remove genetic variants that have malignant potential. The apoptosis mechanism is composed by proteins that are members of BCL-2 and cysteine-protease families. BH3-only peptides are the "natural" intracellular ligands of BCL-2 family proteins. On the other hand, studies have proved that phenothiazine compounds influence the induction of cellular death. To understand the characteristics of phenothiazines and their effects on tumoral cells and organelles involved in the apoptosis, as well as evaluating their pharmacologic potential, we have carried out computational simulation with the purpose of relating the structures of the phenothiazines with their biological activity. Since the tridimensional (3D) structure of the target protein is known, we have employed the molecular docking approach to study the interactions between compounds and the protein's active site. Hereafter, the molecular dynamics technique was used to verify the temporal evolution of the BCL-2 complexes with phenothiazinic compounds and the BH3 peptide, the stability and the mobility of these molecules in the BCL-2 binding site. From these results, the calculation of binding free energy between the compounds and the biological target was carried out. Thus, it was possible to verify that thioridazine and trifluoperazine tend to increase the stability of the BCL-2 protein and can compete for the binding site with the BH3 peptide.

Keywords: apoptosis, cancer, BCL-2, BH3, phenothiazines, docking, molecular dynamics, binding free energy calculations

OPEN ACCESS

Edited by:
Teodorico Castro Ramalho, Universidade Federal de Lavras, Brazil

Reviewed by:
Cheng Fang, Biogen Idec, United States
Daniel Henriques Soares Leal, Federal University of Itajubá, Brazil

***Correspondence:**
Paula Homem-de-Mello
paula.mello@ufabc.edu.br
Kathia Maria Honorio
kmhonorio@usp.br

Specialty section:
This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

Received: *27 September 2019*
Accepted: *11 March 2020*
Published: *03 April 2020*

Citation:
do Carmo AL, Bettanin F, Oliveira Almeida M, Pantaleão SQ, Rodrigues T, Homem-de-Mello P and Honorio KM (2020) Competition Between Phenothiazines and BH3 Peptide for the Binding Site of the Antiapoptotic BCL-2 Protein. Front. Chem. 8:235. doi: 10.3389/fchem.2020.00235

INTRODUCTION

Apoptosis is a highly regulated form of programmed cell death occurring physiologically in living organisms. However, alterations and defects in this process are also involved in the pathogenesis of several diseases, such as cancer, AIDS, Parkinson and Alzheimer diseases, amyotrophic lateral sclerosis and others (Thompson, 1995). Apoptotic cells exhibit morphological alterations, including

plasma membrane blebbing, chromatin condensation, internucleosomal DNA, and formation of apoptotic bodies. Such features result from the action of complex machinery, involving the regulation and execution by BCL-2 family proteins and also by cysteine proteases (initiator or executioner caspases) (Kalkavan and Green, 2018). Particularly regarding cancer, the unlimited proliferative capacity of tumor cells is due to several genetic and molecular alterations, including mechanisms for evading apoptosis (Brown and Attardi, 2005; Hanahan and Weinberg, 2011). One of these mechanisms is the altered expression and function of pro- and antiapoptotic members of B-cell lymphoma-2 (BCL-2) family proteins, directly involved with tumorigenesis and tumor progression/malignance (Coustan-Smith et al., 1996; Gobé et al., 2002). Thus, there are a plenty of molecular studies and clinical trials in course to target BCL-2 proteins to cancer therapy (Adams et al., 2019).

The BCL-2 family is currently divided in proapoptotic members, including BAX and BAK, antiapoptotic members, such as BCL-2, BCL-xL, and MCL-1, and BH3-only proteins (BIM, BID, PUMA, NOXA, and others), which are potent activators of apoptosis (Letai et al., 2002; Youle and Strasser, 2008). Structural characteristics are defined by sequence homology analysis, which allowed identifying four domains (BH1-BH4) involved in protein-protein interactions among members of the BCL-2 family. A hydrophobic slit is formed by BH1, BH2, and BH3 domains, which participates in the uptake of the BH3 domain of pro-apoptotic proteins via heterodimerization; the BH4 domain is present in antiapoptotic activity. The BH3-only proteins just possess the BH3 motif. There are complex interactions among the BCL-2 family members, which comprise a regulatory mechanism of control of cell fate in response to different stimuli. The recruitment of proapoptotic by antiapoptotic proteins occurs through the interaction between the highly conserved helical BH3 domain of the proapoptic protein and a binding groove in the antiapoptotic protein. Considering the homology and structural similarities with the BH3 domain of proapoptotic proteins, BH3-only proteins interact with the binding groove, releasing the proapoptotic proteins and neutralizing the antiapoptotic proteins. In this scenario, BH3-only members have a crucial role in the initiation of apoptotic cell death, since they can bind to the specific domains in anti- or proapoptotic BCL-2 proteins (Lomonosova and Chinnadurai, 2008). It was proposed that BH3-only members can activate directly proapoptotic BAX and BAK. Also, the interaction of BH3-only proteins with antiapoptotic BCL-2 members can disrupts their inhibitory interaction with the proapoptotic members triggering apoptosis (Du et al., 2011; Shamas-Din et al., 2011). As a result, the comprehension of the interaction of BH3-only proteins with other BCL-2 members acting as apoptosis activators resulted in the development of BH3-only mimetic molecules as a strategy to cancer therapy (Merino et al., 2018; Ewald et al., 2019).

Currently, chemicals are under development to inhibit the interactions of pro-apoptotic proteins with the hydrophobic slit of the antiapoptotic protein BCL-2, enabling the imitation of the action of pro-apoptotic proteins with the BH3 domain (Degterev et al., 2001; Delbridge and Strasser, 2015; Zacarías-Lara et al., 2016). Thus, proposals for small molecule interactions with BCL-2 proteins have enabled the development of cancer therapies, including BH3 domain mimetic molecules that bind to the BH3 binding domain in antiapoptotic BCL-2 members such as BCL-2 and BCL-xL (**Figure 1** PDB ID: 2O22). As example of these compounds, one may cite ABT-737, navitoclax (ABT-263), obatoclax mesylate (GX15-070), venetoclax (ABT-199), and gossypol and its derivatives (the structures of these compounds are presented in **Figures S1A–E**) (Oltersdorf et al., 2005; Bajwa et al., 2012; Souers et al., 2013; Pan et al., 2014; Kalkavan and Green, 2018). Thus, several studies are being conducted to identify novel small molecules or peptides able to act as BH3-only mimetics.

In this regards, a class of substances that has potential against BCL-2 refers to phenothiazines (**Figure 2**), as interactions between these compounds and BCL-2 protein may be favored due to the presence of a polycyclic ring system and different substituents modulating the BCL-2 biological activity. The ability of phenothiazines to interact with hydrophobic slits

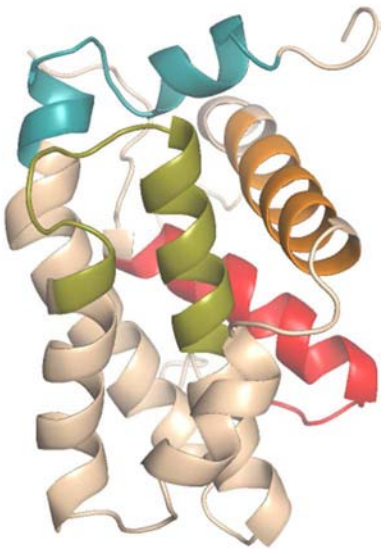


FIGURE 1 | Antiapoptotic BCL-2 protein domains (PDB_ID: 2O22): BH1 (green), BH2 (blue), BH3 (orange), BH4 (red).

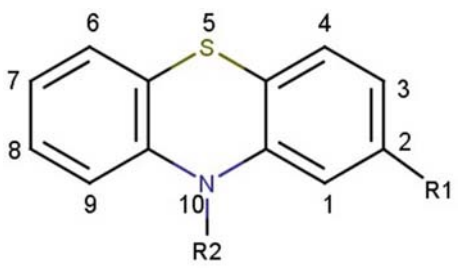


FIGURE 2 | Chemical structure of thiazine nucleus.

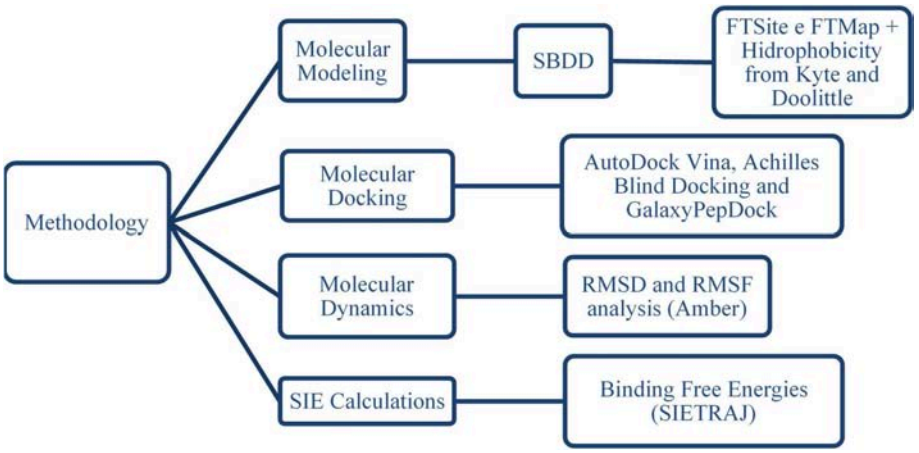


FIGURE 3 | Sequence of methods applied to study the interactions of phenothiazines and the BH3 peptide with BCL-2 protein.

was previously shown by the interactions of thioridazine with putative binding sites of human thioredoxin 1 (Philot et al., 2016). The interaction between phenothiazines and biological membranes occurs because there is amphipathic character of the molecules. The thiazine nucleus is relatively hydrophobic and its side chain can be hydrophilic and even positively charged depending on the pH of the environment (Homem-de-Mello et al., 2005, 2007; Rodrigues et al., 2006; Perussi, 2007; Rodrigues, 2007; Bettanin et al., 2015; de Faria et al., 2015; Nuñez et al., 2015). Additionally, it has been shown that antipsychotic phenothiazine derivatives possess potent cytotoxicity against several types of tumor cells by triggering of apoptosis, with involvement of mitochondrial permeabilization (de Faria et al., 2015; de Mello et al., 2016; Wu et al., 2016; Chu et al., 2019).

Thus, considering that drug design methods have been employed to understand the interactions between small molecules and biological targets, a study via computational techniques and experimental data of BCL-2 and antipsychotic phenothiazine derivatives was developed applying drug design methods to describe a relationship between chemical structure and biological activity of the selected compounds to propose new drug candidates. Bioinformatics tools were used to characterize possible binding sites and regions for anchoring compounds to the target protein (BCL-2). Molecular docking was also employed to identify the interactions of phenothiazines with the BCL-2 antiapoptotic protein, as well as comparing these results with the interactions of BCL-2 with the BH3 peptide.

MATERIALS AND METHODS

The workflow used to study the interactions between phenothiazines (and the BH3 peptide) and BCL-2 protein is presented in **Figure 3**.

To study the relationship between chemical structure and biological activity, as well as evaluating the interactions between bioactive ligands and biological targets, molecular modeling tools can be employed to plan new drug candidates (Andricopulo et al., 2009; Sant'Anna, 2009). The strategy employed in this work is known as Structure-Based Drug Design (SBDD), in which three-dimensional biological receptor structures (obtained from experimental techniques such as X-ray diffraction or nuclear magnetic resonance) are used to propose ligand modifications to improve target affinity and specificity (Andricopulo et al., 2009). In this study, the BCL-2 structures obtained from X-ray and nuclear magnetic resonance (NMR), available in the PDB database (PDB_ID: 1YSW, 2O2F, 2O21, 2O22, 2W3L, 4AQ3, 4IEH, 4LVT, 4LXD, 4MAN, 5AGW, 5AGX, 5JSN) were compared to verify significant differences by aligning these structures (**Figure S2**). The overlap of the structures (**Figure S3**) was performed in the MUSTANG v3.2.2—Multiple Structural Alignment Algorithm program (Konagurthu et al., 2006).

Given this set of alo-protein structures, multiple alignment was obtained using the Cα atom spatial information with the following steps: (I) calculation of root-mean-square deviation (RMSD) taking into account the distances between the Cα atoms for all structures to detect similar substructures between two structures and obtain a quality value for each possible residue-residue match between the two structures; (II) compute the scores of the corresponding residue-residue pairs; (III) structural alignments in pairs; (IV) recalculation of the scores of the corresponding residue-residue pairs in the context of multiple structures; (V) progressive alignment by using the Mustang algorithm (Konagurthu et al., 2006).

Some sequence failures were observed in all BCL-2 structures, so we have selected the structure obtained via NMR (PDB ID: 2O22) (Bruncko et al., 2007; Rose et al., 2018) because it preserves the loop region. Moreover, besides it is an uncut structure, the backbone alignment is similar to the X-ray structures (**Figures S2, S4**).

After choosing the more suitable 3D structure, a study was performed to detect the possible binding sites of human BCL-2, followed by the characterization of these regions, using FTSite (Brenke et al., 2009; Ngan et al., 2012; Kozakov et al., 2015) and FTMap (Brenke et al., 2009; Kozakov et al., 2011, 2015;

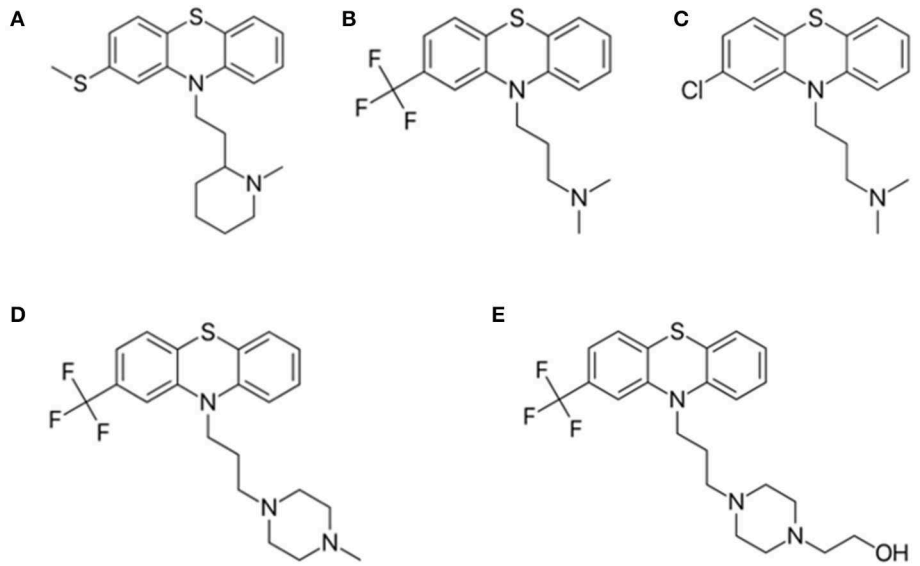


FIGURE 4 | Structure of phenothiazine derivatives analyzed in this study. **(A)** thioridazine; **(B)** triflupromazine; **(C)** chlorpromazine; **(D)** trifluoperazine, and **(E)** fluphenazine.

Bohnuud et al., 2012). FTMap is a server that identifies regions in the macromolecule that have important contributions to the ligand-binding free energy (hot spots). For this, the FTMap algorithm uses 16 probe molecules (**Table S1**) with different shapes, sizes and polarities, which run across the entire surface of the protein looking for the best "positions" for these probes. FTMap is capable of sampling billions of positions for the probe molecules, as well as clustering and ranking them according to an average energy. Consensus sites (CS) are generated, which can be defined as regions at the macromolecule that bind clusters containing different probe molecules, suggesting possible binding hot spots. It is important to highlight that FTMap serves as basis for other algorithms, for example, FTSite that is used to identify ligand binding sites. The main idea of FTSite is ranking the consensus clusters based on the number of non-bonded interactions between the protein and all probe molecules contained in the consensus cluster. So, the amino acid residues interacting with the probe molecules in the top ranked consensus cluster are considered as a possible binding site.

In addition, to better understand the main protein-phenothiazine interactions, the protein hydrophobicity surface was obtained using the UCSF Chimera 1.12 (Pettersen et al., 2004).

The phenothiazine derivatives studied here include thioridazine, triflupromazine, chlorpromazine, trifluoperazine, and fluphenazine (**Figure 4**), which yield relevant *in vitro* cytotoxicity in hepatoma HTC cells (de Faria et al., 2015).

From molecular docking simulations, information on the interaction mode and physicochemical characteristics that affect the affinity of the ligand for the macromolecule is obtained (Wang et al., 2004; Sanchez-Linares et al., 2012). Molecular docking study was performed targeting BCL-2 protein and phenothiazine compounds using the AutoDock Vina 1.5.7. For this, we employed the BCL-2 crystallographic structure (PDB 2O22) with the maximum generation of 10 conformations of each compound. The following parameters were employed in the docking simulations: grid center_x = 4.255, center_y = 1.45, center_z = −5.0, size_x = 25, size_y = 3 and size_z = 34, and exhaustiveness = 20. To validate the docking procedure, redocking analyses were performed in order to recover the original position of the ligand found in the 3D structure of the biological target (Moraes and de Azevedo, 2010).

Visual inspection of the best ligand poses at the target binding site was performed using the PyMOL 2.0, also analyzing the RMSD values calculated by the UCSF Chimera 1.12 and the representation of interactions provided by the Poseview server. It is noteworthy that the RMSD value refers to the average deviation of atoms of an initial structure from the proposed structures and generally the fit is considered successful if the value is below 2.0 Å (Verdonk et al., 2003).

In addition to the AutoDock Vina program, the Achilles Blind Docking server was used to verify molecular interactions in various regions of BCL-2, corroborating the molecular interactions established by phenothiazines in hot spots, where ligands can potentially interact (Brenke et al., 2009; Sanchez-Linares et al., 2012; Kozakov et al., 2015).

For molecular docking and analysis of the interactions between BCL-2 and the BH3 domain, the GalaxyPepDock server was used to analyze protein-protein interactions and better understand cell functions and organization (Lee et al., 2015). In this approach, one of the proteins (or receptor) refers to the origin of the fixed grid coordinate system, and the second protein (or ligand) is defined in a movable grid; interaction energy is defined as a scoring function (Kozakov et al., 2017). To verify the

accuracy of the GalaxyPepDock server, redocking analyses with calculation of RMSD values was performed.

The best poses generated by each docking program were selected based on the interactions and binding energies that were generated by the scoring functions, in order to complement the analysis of interactions obtained from the BINANA 1.2.0. This one is able to characterize hydrogen bonding, hydrophobic contact, close contacts, electrostatic interactions, π interactions and salt bridge between receptor—ligand.

After the molecular docking analyses for the five ligands and the BH3 peptide interacting with BCL-2, the next step to be carried out was the preparation of the systems for molecular dynamics (MD) simulations from the calculation of restrained electrostatic potential charges (Wang et al., 2000) of each ligand (from the conformations obtained from molecular docking). For this, we used the Hartree-Fock methodology (Echenique and Alonso, 2007), as implemented in Gaussian09 (Frisch et al., 2009), with 6-31G* basis set (Ditchfield et al., 1971).

Afterwards, the next step related to the preparation of the systems was the solvent box analysis for the target. In this step, we aim to establish the most suitable solvent box for the BCL-2 protein, where the visual analysis was performed in Chimera 1.62 (Pettersen et al., 2004) and the chosen parameters were: periodic octahedral box with a distance of 12 Å between the target and the walls of the box.

From the obtained solvent parameters, the next steps involved in the preparation of the six systems for the MD simulations were: (I) preparation of the topology of the five phenothiazines in the Antechamber module implemented in Ambertools 12 (Salomon-Ferrer et al., 2013) using the RESP charges (charges for BH3 peptide were obtained from the force field); (II) insertion of the FF99SB force field for the coordinates of the six complexes with the Tleap program; (III) total charge calculation of the six systems (the total charge obtained for the six complexes was −9, so 9 sodium ions were inserted to neutralize the system); (IV) inclusion of TIP3P-type water molecules to fill the simulation box; (V) preparation of MD scripts: isothermal-isobaric or NPT ensemble, Langevin thermostat (ntt = 3) and Monte Carlo barostat (Case et al., 2012).

Simulations were then performed following the following steps: (i) four minimizations to eliminate very close contacts between atoms; in the first minimization, the system was kept fixed (without degrees of freedom); in the second and third simulations, only ligands and peptide were kept fixed and in

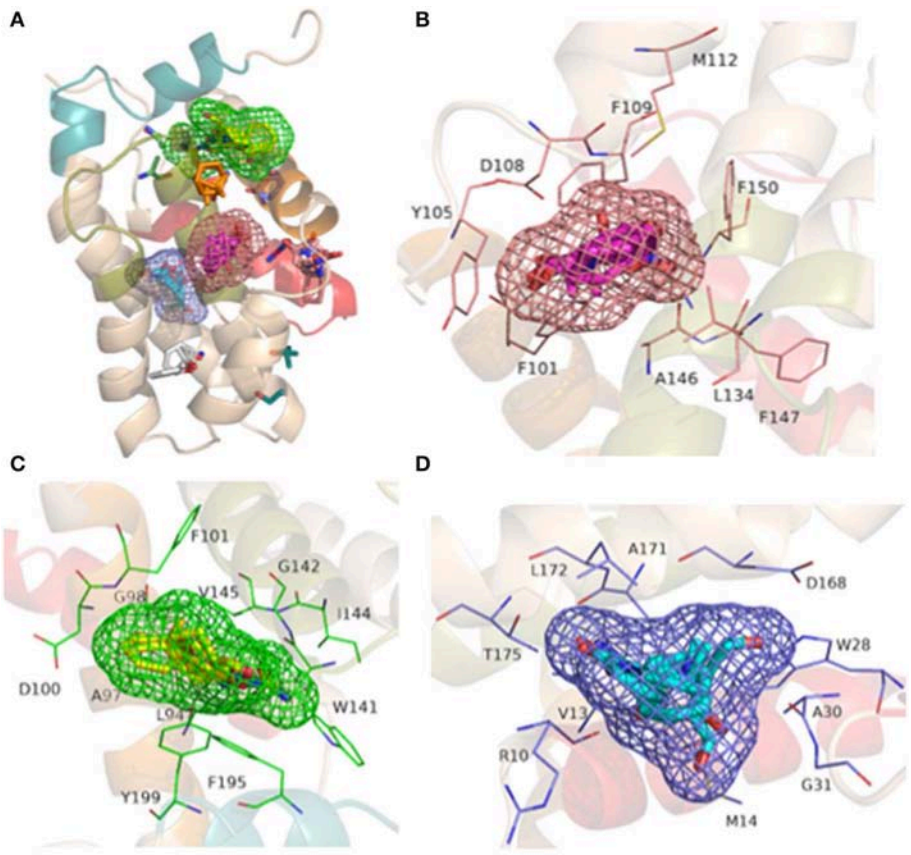


FIGURE 5 | Representation of the BH1-BH4 domains with the presence of the binding sites detected from FTSite and the coupling of probe molecules (in clusters) by the FTMap server produced in the PyMOL 2.0 program: **(A)** the three binding sites with their respective probes; **(B)** Site 1: residues (light pink) with FTSite probes (light pink mesh) and 001 cluster probes (pink); **(C)** Site 2: residues (green) with FTSite probes (green mesh) and 002 cluster probes (yellow) and **(D)** Site 3: residues (blue) with FTSite probes (blue mesh) and 000 cluster probes (light blue).

the last minimization, the whole system was free; (ii) heating (thermal bath) from 0 to 300 K, and a time period of 0.5 ns; the purpose of this step was to control the temperature and adjust the kinetic energy of the system; (iii) 10 ns for equilibration of the system, and it is finalized with the thermal equilibrium of the system; (iv) production step to obtain time subtrajectories. In this final step, the system moves freely and, in addition to simulating thermodynamic properties, a lower energy conformation is obtained for each system under study. This last step was performed over a time period of 100 ns; (v) subtrajectory RMSD value analysis (trajectory stability analysis); (vi) analysis of root-mean-square fluctuation (RMSF) values in order to verify the fluctuations that occur between BCL-2 residues in the presence of the ligands.

Finally, SIE (Solvated Interaction Energy) methodology was employed to estimate the binding free energy related to the ligand-receptor complex by applying the boundary element method (BEM) to solve the Poisson-Boltzmann equation. This method also uses implicit solvation in the study of protein-ligand complexes (Naïm et al., 2007; Silva et al., 2016). For this, SIE was used in this study to estimate the binding free energy between BCL-2 and the five phenothiazine derivatives, or the BH3 peptide, from the most stable subtrajectories generated by the MD simulations. Thus, for the SIE method to be implemented, the following steps were performed: (I) solvent removal using the cpptraj program, generating a file without periodic solvation coordinates (SIE uses implicit solvation); (II) elaboration of the file containing the initial and final coordinates of the dynamics trajectory; (III) choice of frame range over start and end frames (each sub-trajectory of molecular dynamics has 500 frames),

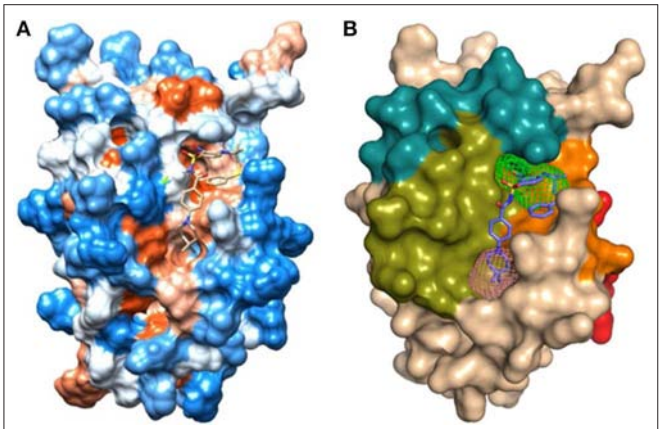


FIGURE 6 | (A) Hydrophobicity surface from Kyte and Doolittle (1982) (**Table S2**): hydrophobic orange-red; neutral white; hydrophilic blue, obtained from UCSF Chimera 1.12 with the crystallographic ligand of protein BCL-2 (PDB_ID: 2O22); **(B)** Representation of the domains: BH1 (green), BH2 (blue), BH3 (orange), BH4 (red), and binding site 1 (pink) and site 2 (green) in the human BCL-2 enzyme - crystallographic structure (PDB_ID: 2O22).

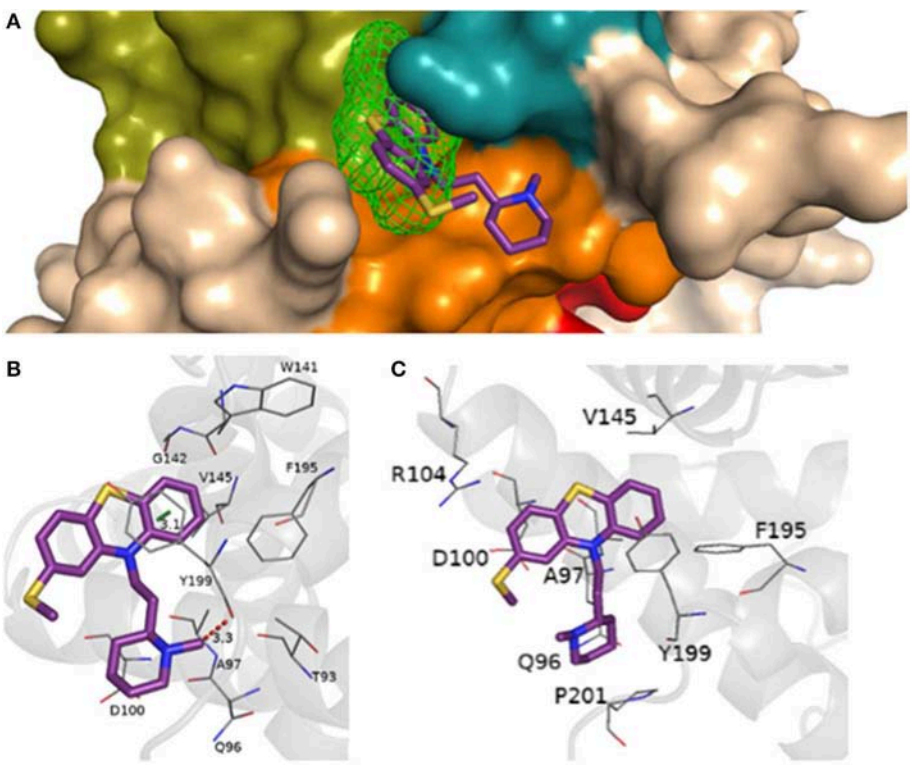


FIGURE 7 | (A) Representation of the BH1-BH4 domains with the presence of the site 2 detected from FTSite and the molecular docking of thioridazine, piperidine subclass, EC_{50} = (45.5 ± 1.0) μmol.L^{-1}, performed at AutoDock Vina 1.5.7. Representation of hydrogen bonding (red) from the molecular docking performed in AutoDock Vina 1.5.7, **(B)** in the Achilles Blind Docking Server, **(C)** and the additional interactions and/or confirmed by BINANA 1.2.0.

from frame 1 to 250; (IV) specify the number of atoms of the target, the five ligands and the peptide; (V) output file definition (sie.log), which contains the results obtained; (VI) calculations by the SIETRAJ program and analysis of the estimated free energy values for the six complexes (BCL-2 + phenothiazines and BCL-2 + BH3 peptide).

RESULTS AND DISCUSSION

Structural Analysis of BCL-2 Structure

Three binding site candidates were located on the BCL-2 protein (PDB_ID: 2W3L) using the FTSite and FTMap servers. The detected sites have the following residues: **site 1**—Phe101, Tyr105, Asp108, Phe109, Met112, Leu134, Ala146, Phe147, Glu149, Phe150; **site 2**—Leu94, Ala97, Gly98, Asp100, Phe101, Trp141, Gly142, Ile144, Val145, Phe195, Tyr199; e **site 3**—Arg10, Val13, Met14, Trp28, Ala30, Gly31, Asp168, Ala171, Leu172 e Thr175. Different probe molecules were used to determine which had the highest affinity for each binding site. The results indicate that the three sites have affinity for polar molecules, hydrogen bond donors and acceptors, hydrophobic and aromatic groups. The representation of the detected binding sites containing the clusters of probe molecules according to the affinity of the molecular interaction on the protein is illustrated in **Figure 5A**. Some probes may also have small contacts with the protein or be in small buried sites, but large CSs occur at the binding site hot spots, also depicted in **Figure 5A**, where the FTMap probes are not in place connection sites determined by FTSite. Residues of the sites 1, 2, and 3 are also shown in **Figures 5B–D**, respectively.

The hydrophobicity scale of the BCL-2 protein was used to complement the results obtained from FTMap and FTSite. **Figure 6** indicates that the crystallographic ligand interacts with the sites 1 and 2 due to its structural size and hydrophobic characteristics. In order to understand the position and the

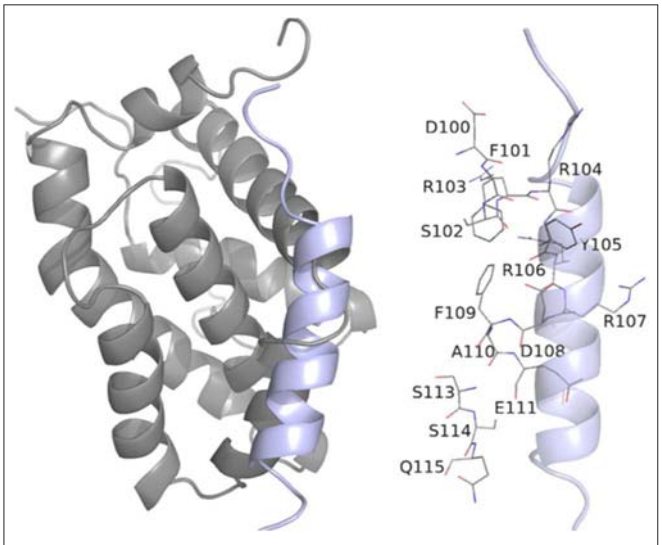


FIGURE 8 | Representation of BCL-2 protein residues with BH3 peptide obtained from molecular docking performed in the GalaxyPepDock server with additional interactions, confirmed by BINANA 1.2.0.

TABLE 1 | Description on the interactions between the main residues of BCL-2 and phenothiazines.

AutoDock Vina 1.5.7		Achilles blind docking server	
Residue	**Interaction**	**Residue**	**Interaction**
Chlorpromazine			
Ala97	Hydrogen Bond	Ala97	Hydrophobic contact
Asp100	Salt Bridge	Asp100	Salt bridge
Phe101	Hydrophobic contact	Phe101	Hydrophobic contact
Arg104	Hydrophobic contact	Arg104	Hydrophobic contact
Trp141	Hydrophobic contact	Trp141	Hydrophobic contact
Val145	Hydrophobic contact	Val145	Hydrophobic contact
Phe195	π-stacking	Phe195	Hydrophobic contact
		Tyr199	Hydrophobic contact
Trifluprazine			
Gln96	Hydrophobic contact	Leu94	Hydrophobic contact
Ala97	Hydrophobic contact	Ala97	Hydrophobic contact
Asp100	Hydrophobic contact	Phe101	Hydrophobic contact
Phe101	Hydrophobic contact	Trp141	Hydrophobic contact
Arg104	Cation-π	Val145	Hydrophobic contact
Val145	Hydrophobic contact	Phe195	Cátion-π
Phe195	Hydrophobic contact	Tyr199	π-stacking
Tyr199	Hydrogen Bond		
Pro201	Hydrophobic contact		
Fluphenazine			
Ala97	Hydrophobic contact	Gln96	Hydrophobic contact
Asp100	Hydrophobic contact	Ala97	Hydrophobic contact
Phe101	Hydrophobic contact	Asp100	Hydrophobic contact
Arg104	Cátion-π	Phe101	Hydrophobic contact
Tyr105	Hydrophobic contact	Arg104	Hydrophobic contact
Gly142	Hydrogen Bond	Val145	Hydrophobic contact
Phe195	Hydrophobic contact	Pro201	Hydrophobic contact
Tyr199	π-stacking	Ser202	Hydrogen Bond
Trifluoperazine			
Leu94	Hydrogen bond	Thr93	Hydrophobic contact
Ala97	Hydrogen bond	Leu94	Hydrophobic contact
Asp100	Salt bridge	Gln96	Hydrophobic contact
Phe101	Hydrophobic contact	Ala97	Hydrophobic contact
Arg104	Hydrophobic contact	Phe101	Hydrophobic contact
Trp141	Hydrophobic contact	Arg104	Hydrophobic contact
Val145	Hydrophobic contact	Trp141	Hydrophobic contact
Phe195	π-stacking	Val145	Hydrophobic contact
Tyr199	Hydrogen Bond	Phe195	Hydrophobic contact
		Tyr199	Hydrophobic contact
Thioridazine			
Thr93	Hydrophobic contact	Gln96	Hydrophobic contact
Gln96	Hydrophobic contact	Ala97	Hydrophobic contact
Ala97	Hydrophobic contact	Asp100	Hydrophobic contact
Asp100	Hydrophobic contact	Arg104	Hydrophobic contact
Trp141	Hydrophobic contact	Val145	Hydrophobic contact
Gly142	Hydrophobic contact	Phe195	Hydrophobic contact
Val145	Hydrophobic contact	Tyr199	Hydrophobic contact
Phe195	Hydrophobic contact	Pro201	Hydrophobic contact
Tyr199	π-stacking and hydrogen Bond		

interactions of phenothiazines with BCL-2, molecular redocking was required.

Next, the molecular docking procedure was performed with the maximum number of poses to be generated at the defined binding site. Thus, the ligand structure found in the tridimensional structure of BCL-2 was superimposed with the ligand pose obtained via molecular docking (**Figure S5**). This overlap generated a RMSD value of 1.327 Å, indicating that the parameters used for the redocking analysis reliably reproduce the experimental conformation of the ligand and can be used to dock the phenothiazine derivatives. An analysis of the images generated by the Poseview program shows that the interactions of the ligand (via redocking) with BCL-2 protein are very similar to those observed in the experimental structure (**Figure S6**).

Molecular Docking

Once the parameters were validated from the redocking analyses, molecular docking studies between BCL-2 and phenothiazine derivatives were performed. It is noteworthy that, according to the experimental data (de Faria et al., 2015), it is possible to establish the following ascending order for cytotoxicity: chlorpromazine < trifluprom azine < fluphenazine < trifluoperazine < thioridazine.

The interaction of the most cytotoxic phenothiazine derivative thioridazine with BCL-2 is depicted in **Figure 7** (docking data for the other phenothiazine derivatives are presented in **Figures S7–S10**). In addition, considering the possible competition between a BH3-only peptide and the thioridazine for BCL-2 binding site, the interaction of BCL-2 and a BH3 domain was evaluated. The docking results considering the interaction between BCL-2 and the BH3 domain obtained from the GalaxyPepDock server are presented in **Figure 8**. Validation of the docking procedure between BCL-2 and BH3 peptide resulted in a satisfactory RMSD (root-mean-square deviation) value of 1.962 Å and preservation of the alpha helix structure (**Figure S11**).

From the docking results, it was possible to verify that phenothiazine derivatives showed similar interactions at BCL-2 protein, since we can see that the thiazine nucleus is located at the site 2 recognized by the FTSite server, which has favorable affinities by aromatic and hydrophobic groups.

Both strategies employed to predict the interactions between phenothiazines and BCL-2 are in reasonable agreement with regard to energies and interactions, as can be seen in **Tables 1**, **2**. The ligand-target complexes obtained from the AutoDock Vina program proved to be adequate and confirmed by the Achilles

TABLE 3 | Description of the interactions between BCL-2 residues and the BH3 peptide.

GalaxyPepDock Server	
Residue	**Interaction**
Asp100	Hydrophobic contact
Phe101	Hydrophobic contact
Ser102	Hydrophobic contact
Arg103	Hydrophobic contact
Arg104	Hydrophobic contact
Tyr105	Hydrophobic contact
Arg106	Hydrophobic contact
Arg107	Hydrogen Bond and Salt Bridge
Asp108	Hydrophobic contact
Phe109	Hydrophobic contact
Ala110	Hydrophobic contact
Glu111	Hydrophobic contact
Met112	Hydrophobic contact
Ser114	Hydrophobic contact
Gln115	Hydrophobic contact

TABLE 2 | EC_{50} values (± 1.0 μmol.L^{-1}), interaction energies (kcal.mol^{-1}) and number (#) of interactions obtained from AutoDock Vina and Achilles Blind Docking for phenothiazines and BCL-2.

		Chlorpromazine	Trifluprom azine	Fluphenazine	Trifluoperazine	Thioridazine
	EC^{*}_{50}	125.3	105.9	63.2	56.2	45.5
AutoDock Vina 1.5.7	Interaction energy (kcal.mol^{-1})	−6.4	−6.6	−6.6	−7.1	−6.0
	#hydrophobic contacts	4	7	5	4	8
	#hydrogen bonds	1	1	1	3	1
	# π-stacking	1	0	1	1	1
	#other interactions**	1	1	1	1	0
	Total #interactions	7	9	8	9	10
Achilles Blind Docking	Energy (kcal.mol^{-1})	−6.3	−7.1	−7.4	−7.5	−6.8
	#hydrophobic contacts	7	5	7	10	8
	#hydrogen bonds	0	0	1	0	0
	# π-stacking	0	1	0	0	0
	# other interactions**	1	1	0	0	0
	Total #interactions	8	7	8	10	8

*(de Faria et al., 2015).

**salt bridge or cation- π.

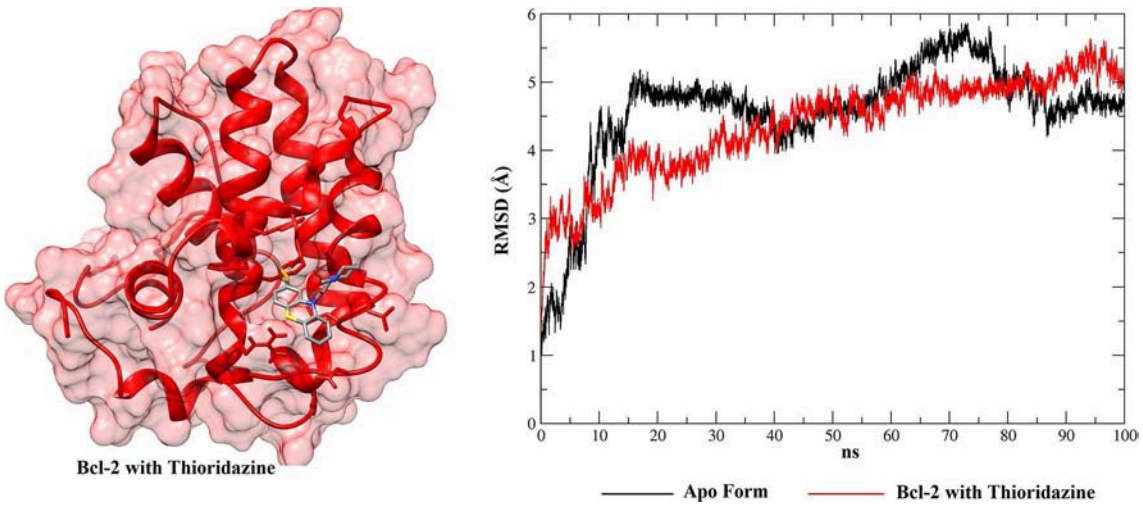


FIGURE 9 | RMSD values of BCL-2 in the apo form (black), BCL-2 with thioridazine (red) and 3D conformation obtained from the clustering of all structures obtained from the MD simulation.

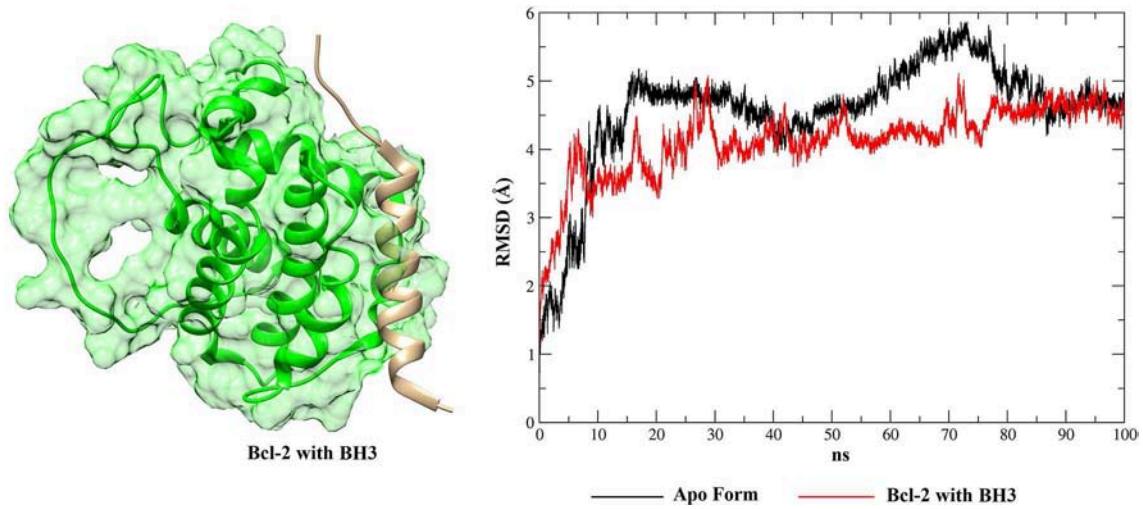


FIGURE 10 | RMSD values of BCL-2 in the apo form (black), BCL-2 with the BH3 peptide (red) and 3D conformation obtained from the clustering of all structures obtained from the MD simulation.

Blind Docking Server, and then, they were used for the molecular dynamics simulations. The main interactions between BCL-2 and the BH3-only peptide are displayed in **Table 3**, and we can see that the main interactions observed were hydrophobic contacts. This simulation was possible only with the GalaxyPepDock Server, which is specific to identify protein-peptide interactions.

Through the analysis of the bioactive conformations obtained from both programs, it was observed that the thiazine nucleus present in the studied ligands showed affinity by BCL-2, possibly due to the presence of aromatic and hydrophobic groups, which was also suggested by the FTSite server. In order to understand dynamic and energetic factors involved in the interaction between the phenothiazines and BCL-2, molecular dynamics simulations and calculations of binding free energy (ΔG) were performed.

Molecular Dynamics

From the molecular docking between phenothiazines and BCL-2 (and between the peptide BH3 and BCL-2), the possible conformations of the phenothiazine derivatives and, additionally the BH3 peptide to evaluate competition, were chosen to perform MD simulations. Then, after the end of the simulations, the RMSD plots were generated from the Cpptraj platform (Amber 12). This analysis was performed to verify the temporal evolution of the complexes BCL-2 + phenothiazines and BCL-2 + BH3, as well as the stability and the mobility of the formed systems. MD simulations were also performed with the apo BCL-2 target (no ligand at its binding site) and, in addition to the RMSD plots, the structures observed along 100 ns of MD simulations were clustered for all complexes and the protein in the form apo using Chimera 1.13.1. **Figures 9**, **10** show the RMSD plots for

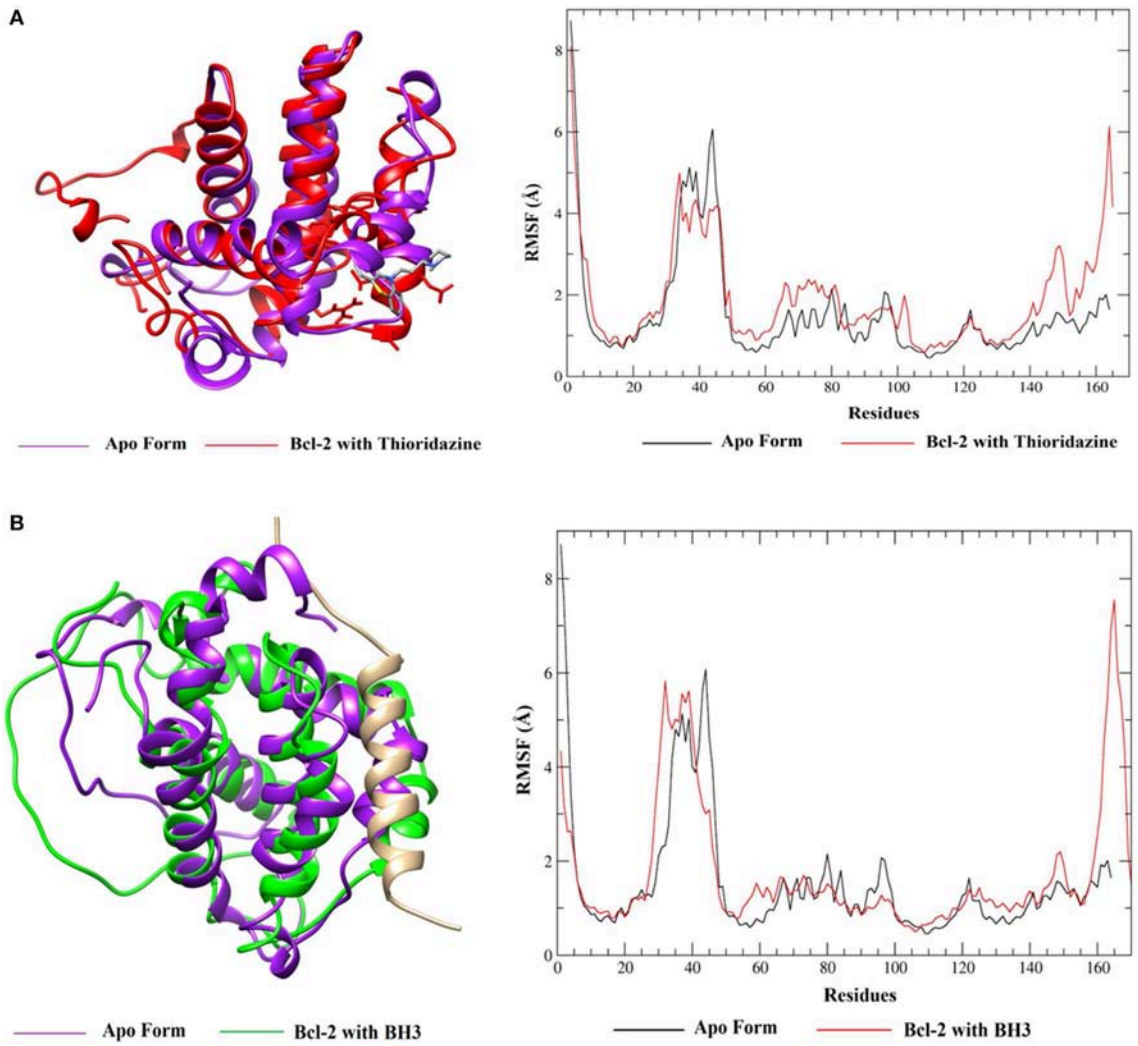


FIGURE 11 | RMSF values of BCL-2 in the presence of **(A)** thioridazine and **(B)** BH3-only peptide.

the complexes formed by BCL-2 + thioridazine and BCL-2 + BH3. The results obtained from the MD simulations for the other phenothiazines are presented in **Figures S12–S15**.

From the RMSD plots for the BCL-2 trajectories, the five phenothiazine compounds and the BH3 peptide, an analysis of trajectory stability and mobility can be performed. Thus, **Figure 9** shows that thioridazine increases the level of stability of the complex, because BCL-2 with this molecule coupled at its binding site exhibits lower RMSD variations. The second complex with the smallest variations is shown in **Figure S12** (**Supplementary Material**—BCL-2 + trifluoperazine). These two phenothiazine compounds have the lowest EC_{50} values for cytotoxicity against HTC cells (45.5 ± 1.0 and 56.2 ± 1.0 $\mu mol.L^{-1}$, respectively) and these results showed that BCL-2 has greater stability in the presence of these two ligands. Compared to other phenothiazine compounds (**Figures S13–S15**), RMSD results show greater variations, i.e., suggesting that fluphenazine, trifluopromazine and chlorpromazine decreased the target stability.

With respect to the peptide BH3, **Figure 10** shows that the receptor exhibits smaller RMSD variations than for the phenothiazines, suggesting that the peptide BH3 assists in increasing the target stability. However, **Figure 10** also shows that BH3 is the ligand that exhibits greater variations in RMSD and this can be explained by the size of this peptide relative to the phenothiazine compounds.

When these RMSD values are compared with the apo form, thioridazine and the BH3 peptide cause structural changes in the BCL-2 target; however, when this receptor has the thioridazine molecule in its binding site, the complex is more stable. It is also noted that in all cases the RMSD variations for the receptors are >4 Å and this can be explained by loops on the BCL-2 chains, which increase their mobility. These results are corroborated by the RMSF plots, which are displayed in **Figure 11** for thioridazine

TABLE 4 | Half maximal effective concentration (EC50) number of interactions obtained with the different softwares (Total #interactions), binding free energy (ΔG) and binding free energy divided by the number of interactions (ΔG/#interactions) calculated for the six complexes (BCL-2 + phenothiazines and BCL-2 + BH3).

Ligand	EC_{50} (± 1.0 $\mu mol.L^{-1}$)*	Total #interactions (Vina/Achilles)	ΔG ($kcal.mol^{-1}$)	ΔG/ #interactions ($kcal.mol^{-1}$)	
				Vina	Achilles
Thioridazine	45.5	10/8	−7.00	−0.70	−0.88
Trifluoperazine	56.2	9/10	−6.47	−0.72	−0.65
Fluphenazine	63.2	8/8	−5.51	−0.69	−0.69
Triflupromazine	105.69	9/7	−4.93	−0.55	−0.70
Chlorpromazine	125.3	7/8	−4.05	−0.58	−0.51
BH3	-	16**	−12.81	−0.80	

*de Faria et al. (2015).
** Obtained from Galaxy PepDock Server.

and the BH3 peptide, respectively. RMSF plots for the other phenothiazines can be viewed in **Figures S16–S19**.

RMSF plots were generated in order to verify the flexibility and the mobility of the backbone atoms at the complexes. The plots show the different amplitudes of fluctuations in the apo and holo forms (**Figure 11A** and **Figure S16**) and indicate that the complexes with thioridazine and trifluoperazine have the smallest fluctuations in the main residues, relative to the other phenothiazine derivatives. It is also noted that there is a large fluctuation of residues in the loop regions. For the BH3 peptide (**Figure 11B**), the fluctuations of the amino acid residues are smaller, confirming the RMSD values. Some loop regions of both BCL-2 and the peptide structure show higher fluctuations, suggesting that the peptide is more mobile compared to other compounds because it has a larger number of atoms in its structure and has side chains and a loop.

Therefore, from the results obtained from MD simulations (RMSD and RSMF values) it was possible to verify that thioridazine and trifluoperazine tend to increase the stability of the BCL-2 protein. From **Figure 11**, we can also verify the alignment of the conformations generated after the clustering, the differences in the fluctuations of the apo-BCL-2 residues and the complexes (BCL-2 + ligands). Thus, from these results, the next step of this work involved the calculation of the binding free energy between BCL-2 and the phenothiazine derivatives or the BH3-only peptide using the SIE method.

Free Energy Calculation via SIE Method

After choosing the sub-trajectories that presented the smallest variations (from RMSD values derived from MD simulations), it was possible to estimate the binding free energy (ΔG) of BCL-2 and the molecules under study, using the SIE method. These simulations were performed in order to analyze the stability of BCL-2 in relation to the six ligands. Thus, the ΔG values calculated by the SIE method are presented in **Table 4**.

The EC_{50} values of phenothiazines were directly proportional to the binding ΔG values. The interaction of BCL-2 with thioridazine has a ΔG^{SIE} value of -7.00 $kcal.mol^{-1}$, while chlorpromazine (the highest EC_{50} value) has the lowest binding free energy value when interacts with BCL-2 (-4.05 $kcal.mol^{-1}$). As expected from its physiological role, the ΔG value for the BH3 peptide was the highest (-12.81 $kcal.mol^{-1}$) that corroborates the RMSD/RSMF analysis and reinforces that BCL-2 presents a strong interaction with the peptide, increasing its stability.

The results obtained from the SIE method also showed that thioridazine and trifluoperazine could inhibit the target BCL-2, just like BH3 peptide, because complexes are favorable and stable since its ΔG values are the most negative. Since most of the interactions between ligands and protein are hydrophobic, one can estimate the free energy of interaction per site (ΔG/#interactions in **Table 4**) by combining SIE free energies and the number of interactions obtained from docking. This analysis is proposed here since the molecules are smaller than the peptide, and so are the number of possible interactions. Then, if one ligand can interact in the same intensity per site of interaction as BH3 peptide, this ligand can compete for the binding site. Of course, this is a simplistic analysis because each interaction may be stronger or weaker, but it is insightful to verify that thioridazine and trifluoperazine have interaction energy per binding site comparable to BH3 peptide.

CONCLUSIONS

From this study on the BCL-2 protein (involved in the apoptosis process) and some phenothiazine derivatives that have pharmacological properties, we can concluded that phenothiazines may compete with pro-apoptotic proteins. These results were obtained from molecular docking, RMSD and RMSF values and binding free energy.

Docking simulations were important to understand the main interactions between the target (BCL-2) and the phenothiazine compounds. RMSD results for the complex formed between BCL-2 and the two most active phenothiazine compounds (thioridazine and trifluoperazine) suggest that BCL-2 has a higher stability in the presence of these two ligands. Compared to other phenothiazine compounds, RMSD results show greater variations, i.e., the results suggest that fluphenazine, triflupromazine and chlorpromazine decrease the target stability. RMSF plots for the trajectories of BCL-2 and the five phenothiazine compounds showed that thioridazine and trifluoperazine have the smallest fluctuations considering the major residues compared to other phenothiazine compounds.

The binding free energy between BCL-2, the phenothiazine compounds and the BH3 peptide was calculated using the SIE method and the results obtained indicated that the phenothiazine compounds with lower EC_{50} values presented greater affinity (measured by means of ΔG). The net binding energy for BH3 peptide is larger than the net binding energy obtained for the phenothiazines, since BH3 is a larger compound, with many different points for interaction with BCL-2. Moreover, BH3 is the natural ligand of BCL-2, selected evolutionarily to bind. However, our data indicate that the interactions are quite specific for the compounds with greater EC_{50}; this interaction can generate competition in specific situations, including chemotherapy in

tumor cells, which would induce cell death, and can act as co-adjuvants by this mechanism. Thus, the results obtained in this study can help to better understand the mechanisms involved in the interaction of BCL-2 and phenothiazine compounds and, consequently, may help the design of new substances with improved activity against BCL-2. It should be noted that the inhibition of the antiapoptotic protein BCL-2 by phenothiazines may help explain its apoptosis-inducing effect on tumor cells.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

AC and SP carried out calculations (characterization of the protein and docking). MO performed molecular dynamics and SIE calculations, TR was the experimental collaborator (proposed the problem and obtained EC50 data). FB participated on discussions and writing the manuscript. PH-d-M and KH proposed and supervised all the computational simulations, organized discussions with experimental collaborator and wrote/revised the manuscript.

ACKNOWLEDGMENTS

The authors thank FAPESP (2006/00995-9, 2016/07367-5, 2017/23416-9, 2016/24524-7, and 2017/10118-0), CNPq (#306177/2016-1, 306585/2019-7, and 312020/2019-8) and CAPES (Finance Code 001 and fellowship EACH/CPG 027/2019). The authors are grateful to the Multiuser Central Facilities (UFABC) for the computational facility.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2020.00235/full#supplementary-material

REFERENCES

Adams, C. M., Clark-Garvey, S., Porcu, P., and Eischen, C. M. (2019). Targeting the BCL-2 family in B cell lymphoma. *Front Oncol.* 8:636. doi: 10.3389/fonc.2018.00636

Andricopulo, A. D., Salum, L. B., and Abraham, D. J. (2009). Structure-based drug design strategies in medicinal chemistry. *Curr. Top. Med. Chem.* 9, 771–790. doi: 10.2174/156802609789207127

Bajwa, N., Liao, C., and Nikolovska-Coleska, Z. (2012). Inhibitors of the anti-apoptotic BCL-2 proteins: a patent review. *Expert Opin. Ther. Pat.* 22, 37–55. doi: 10.1517/13543776.2012.644274

Bettanin, F., de Carvalho Fontinelles, T. A., Maciel, C. D., Dias, L. G., Coutinho-Neto, M. D., and Homem-de-Mello, P. (2015). Aggregation of photosensitizers: the role of dispersion and solvation on dimer formation energetics. *Theor. Chem. Acc.* 134,152. doi: 10.1007/s00214-015-1732-5

Bohnuud, T., Beglov, D., Ngan, C. H., Zerbe, B., Hall, D. R., Brenke, R., et al. (2012). Computational mapping reveals dramatic effect of Hoogsteen breathing on duplex DNA reactivity with formaldehyde. *Nucleic Acids Res.* 40, 7644–7652. doi: 10.1093/nar/gks519

Brenke, R., Kozakov, D., Chuang, G-Y., Beglov, D., Hall, D., Landon, M. R., et al. (2009). Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* 25, 621–627. doi: 10.1093/bioinformatics/btp036

Brown, M. J., and Attardi, L. D. (2005). The role of apoptosis in cancer development and treatment response. *Nat. Rev. Cancer* 5, 231–237. doi: 10.1038/nrc1560

Bruncko, M., Oost, T. K., Belli, B. A., Ding, H., Joseph, M. K., Kunzer, A., et al. (2007). Solution structure of the anti-apoptotic protein BCL-2 in complex with an acyl-sulfonamide-based *ligand. J. Med. Chem.* 50, 641–662. doi: 10.2210/pdb2o2f/pdb

Case, D. A., Darden, T., Cheatham, T. E., Simmerling, C., Wang, J., Duke, R. E., et al. (2012). *AMBER 12*. São Francisco, CA: University of California.

Chu, C-W, Ko, H-J., Chou, C-H., Cheng, T-S., Cheng, H-W., Liang, Y-H., et al. (2019). Thioridazine enhances P62-mediated autophagy and apoptosis through Wnt/β-catenin signaling pathway in glioma cells. *Int. J. Mol. Sci.* 20:473. doi: 10.3390/ijms20030473

Coustan-Smith, E., Kitanaka, A., Pui, C. H., McNinch, L., Evans, W. E., Raimondi, S. C., et al. (1996). Clinical relevance of BCL-2 overexpression in childhood acute lymphoblastic leukemia. *Blood* 87, 1140–1146. doi: 10.1182/blood.V87.3.1140.bloodjournal8731140

de Faria, P. A., Bettanin, F., Cunha, R. L. O. R., Paredes-Gamero, E. J., Homem-de-Mello, P., Nantes, I. L., et al. (2015). Cytotoxicity of phenothiazine derivatives associated with mitochondrial dysfunction: a structure-activity investigation. *Toxicology* 330, 44–54. doi: 10.1016/j.tox.2015.02.004

de Mello, J. C., Moraes, V. W. R., Watashi, C. M., da Silva, D. C., Cavalcanti, L. P., Franco, M. K. K. D., et al. (2016). Enhancement of chlorpromazine antitumor activity by pluronics F127/L81 nanostructured system against human multidrug resistant leukemia. *Pharmacol. Res.* 111, 102–112. doi: 10.1016/j.phrs.2016.05.032

Degterev, A., Lugovskoy, A., Cardone, M., Mulley, B., Wagner, G., Mitchison, T., et al. (2001). Identification of small-molecule inhibitors of interaction between the BH3 domain and BCL-XL. *Nat. Cell Biol.* 3, 173–182. doi: 10.1038/35055085

Delbridge, A. R. D., and Strasser, A. (2015). The BCL-2 protein family, BH3-mimetics and cancer therapy. *Cell Death Differ.* 22, 1071–1080. doi: 10.1038/cdd.2015.50

Ditchfield, R., Hehre, W. J., and Pople, J. A. (1971). Self-consistent molecular-orbital methods. IX. an extended gaussian-type basis for molecular-orbital studies of organic molecules. *J. Chem. Phys.* 54:724. doi: 10.1063/1.1674902

Du, H., Wolf, J., Schafer, B., Moldoveanu, T., Chipuk, J. E., and Kuwana, T. (2011). BH3 domains other than bim and bid can directly activate Bax/Bak. *J. Biol. Chem.* 286:491–501. doi: 10.1074/jbc.M110.167148

Echenique, P., and Alonso, J. L. (2007). A mathematical and computational review of hartree-fock SCF methods in quantum chemistry. *Mol. Phys.* 105, 3057–3098. doi: 10.1080/00268970701757875

Ewald, L., Dittmann, J., Vogler, M., and Fulda, S. (2019). Side-by-side comparison of BH3-mimetics identifies MCL-1 as a key therapeutic target in AML. *Cell Death Dis.* 10:917. doi: 10.1038/s41419-019-2156-2

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., et al. (2009). *Gaussian 09*. Wallingford, CT: Gaussian, Inc.

Gobé, G., Rubin, M., Williams, G., Sawczuk, I., and Buttyan, R. (2002). Apoptosis and expression of BCL-2, BCL-XL, and bax in renal cell carcinomas. *Cancer Invest.* 20, 324–332. doi: 10.1081/CNV-120001177

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Homem-de-Mello, P., Mennucci, B., Tomasi, J., and da Silva, A. B. F. (2005). The effects of solvation in the theoretical spectra of cationic dyes. *Theor. Chem. Acc.* 113, 274–280. doi: 10.1007/s00214-005-0668-6

Homem-de-Mello, P., Mennucci, B., Tomasi, J., and da Silva, A. B. F. (2007). Cationic dye dimers: a theoretical study. *Theor. Chem. Acc.* 118, 305–314. doi: 10.1007/s00214-007-0261-2

Kalkavan, H., and Green, D. R. (2018). MOMP, cell suicide as a BCL-2 family business. *Cell Death Diff.* 25, 46–55. doi: 10.1038/cdd.2017.179

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., and Lesk, A. M. (2006). MUSTANG: a multiple structural alignment algorithm. *Proteins* 64, 559–574. doi: 10.1002/prot.20921

Kozakov, D., David, R. H., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., et al. (2017). The ClusPro web server for protein-protein docking. *Nat. Protoc.* 12, 255–278. doi: 10.1038/nprot.2016.169

Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., et al. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* 10, 733–755. doi: 10.1038/nprot.2015.043

Kozakov, D., Hall, D. R., Chuang, G. Y., Cencic, R, Brenke, R., Grove, L. E., et al. (2011). Structural conservation of druggable hot spots in protein- protein interfaces. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13528–13533. doi: 10.1073/pnas.1101835108

Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157, 105–132. doi: 10.1016/0022-2836(82)90515-0

Lee, H., Heo, L., Lee, M. S., and Seok, C. (2015). GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization. *Nucleic Acids Res.* 43, W431–W435. doi: 10.1093/nar/gkv495

Letai, A., Bassik, M. C., Walensky, L. D., Sorcinelli, M. D., Weiler, S., and Korsmeyer, S. J. (2002). Distinct BH3 domains either sensitize or activate mitochondrial apoptosis, serving as prototype cancer therapeutics. *Cancer Cell.* 2, 183–192. doi: 10.1016/S1535-6108(02)00127-7

Lomonosova, E., and Chinnadurai, G. (2008). BH3-only proteins in apoptosis and beyond: an overview. *Oncogene.* 27(Suppl. 1):S2–19. doi: 10.1038/onc.2009.39

Merino, D., Kelly, G. L., Lessene, G., Wei, A. H., Roberts, A. W., and Strasser, A. (2018). BH3-mimetic drugs: blazing the trail for new cancer medicines. *Cancer Cell.* 34, 879–891. doi: 10.1016/j.ccell.2018.11.004

Moraes, F. P., and de Azevedo, W. F., Jr. (2010). *Computação Biologicamente Inspirada Aplicada Ao Estudo Da Interação Proteína-Ligante*. Porto Alegre: PUCRS.

Naïm, M., Bhat, S., Rankin, K. N., Dennis, S., Chowdhury, S., F., and Siddiqi, I., et al. (2007). Solvated Interaction Energy (SIE) for scoring protein-ligand binding affinities. 1. exploring the parameter space. *J. Chem. Inform. Model.* 47, 122–133. doi: 10.1021/ci600406v

Ngan, C-H., Hall, D. R., Zerbe, B., Grove, L. E., Kozakov, D., Vajda, S. (2012). FTSite: high accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics* 28, 286–287. doi: 10.1093/bioinformatics/btr651

Nuñez, S. C., Yoshimura, T. M., Ribeiro, M. S., Junqueira, H. C., Maciel, C., Coutinho-Neto, M. D., and Baptista, M. S. (2015). Urea enhances the photodynamic efficiency of methylene blue. *J. Photochem. Photobiol. B* 50, 31–37. doi: 10.1016/j.jphotobiol.2015.03.018

Oltersdorf, T., Elmore, S. W., Shoemaker, A. R., Armstrong, R. C., Augeri, D. J., Belli, B. A., et al. (2005). An inhibitor of BCL-2 family proteins induces regression of solid tumours. *Nature* 435, 677–681. doi: 10.1038/nature03579

Pan, R., Hogdal, L. J., Benito, J. M., Bucci, D., Han, L., Borthakur, G., et al. (2014). Selective BCL-2 inhibition by ABT-199 causes on-target cell death in acute myeloid leukemia. *Cancer Discov.* 4, 362–375. doi: 10.1158/2159-8290.CD-13-0609

Perussi, J. R. (2007). Inativação fotodinâmica de microrganismos. *Química Nova* 30, 988–994. doi: 10.1590/S0100-40422007000400039

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Philot, E. A., da Mata Lopes, D., de Souza, A. T., Braz, A. S. K., Nantes, I. L., Rodrigues, T., et al. (2016). Binding of phenothiazines into allosteric hydrophobic pocket of human thioredoxin 1. *Eur. Biophys. J.* 45, 279–286. doi: 10.1007/s00249-016-1113-6

Rodrigues, T. (2007). *Chemical and Photochemical Behavior of Phenothiazines in Heterogeneous Media Related to a Diversity of Biological Effects*. Kerala: Catalysis and Photochemistry in Heterogeneous Media. Research Signpost.

Rodrigues, T., dos Santos, C. G., Riposati, A., Barbosa, L. R. S., Di Mascio, P., Itri, R., et al. (2006). Photochemically generated stable cation radical of phenothiazine aggregates in mildly acid buffered solutions. *J. Phys. Chem. B* 110, 12257–12265. doi: 10.1021/jp0605404

Rose, A. S., Bradley, A. R., Valasatava, Y., Duarte, J. M., Prlic, A., and Rose, P. W. (2018). NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics* 34, 3755–3758. doi: 10.1093/bioinformatics/bty419

Salomon-Ferrer, R., Case, D. A., and Walker, R. C. (2013). An overview of the amber biomolecular simulation package. *Comp. Mol. Sci.* 3, 198–210. doi: 10.1002/wcms.1121

Sanchez-Linares, I., Perez-Sanchez, H., Cecilia, J. M., and Garcia, J. M. (2012). High-throughput parallel blind virtual screening using BINDSURF. *BMC Bioinformatics* 13(Suppl. 14):S13. doi: 10.1186/1471-2105-13-S14-S13

Sant'Anna, C. M. R. (2009). Molecular modeling methods in the study and design of bioactive compounds: an introduction. *Rev Virtual de Química* 1, 49–57. doi: 10.5935/1984-6835.20090007

Shamas-Din, A., Brahmbhatt, H., Leber, B., and Andrews, D. W. (2011). BH3-only proteins: orchestrators of apoptosis. *Biochim Biophys Acta.* 1813:508–20. doi: 10.1016/j.bbamcr.2010.11.024

Silva, J. R. A., Bishai, W. R., Govender, T., Lamichhane, G., Maguire, G. E. M., Kruger, H. G., et al. (2016). Targeting the cell wall of *mycobacterium tuberculosis*: a molecular modeling investigation of the interaction of imipenem and meropenem with L,D-transpeptidase 2. *J. Biomol. Struct. Dyn.* 34, 304–317. doi: 10.1080/07391102.2015.1029000

Souers, A. J., Leverson, J. D., Boghaert, E. R., Ackler, S. L., Catron, N. D., Chen, J., et al. (2013). ABT-199, a Potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. *Nat. Med.* 19, 202–208. doi: 10.1038/nm.3048

Thompson, C. B. (1995). Apoptosis in the pathogenesis and treatment of disease. *Science* 267, 1456–1462. doi: 10.1126/science.7878464

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins.* 52, 609–623. doi: 10.1002/prot.10465

Wang, J., Cieplak, P., and Kollman, P. A. (2000). How Well does a Restrained Electrostatic Potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* 21, 1049–74. doi: 10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F

Wang, R., Lu, Y., Fang, X., and Wang, S. (2004). An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J. Chem. Inf. Comput. Sci.* 44, 2114–2125. doi: 10.1021/ci049733j

Wu, C. H., Bai, L. Y., Tsai, M. H., Chu, P-C., Chiu, C. F., Chen, M. Y., et al. (2016). Pharmacological exploitation of the phenothiazine antipsychotics to develop novel antitumor agents-a drug repurposing strategy. *Sci. Rep.* 6:27540. doi: 10.1038/srep27540

Youle, R. J., and Strasser, A. (2008). The BCL-2 protein family: opposing activities that mediate cell death. *Nat. Rev. Mol. Cell Biol.* 9, 47–59. doi: 10.1038/nrm2308

Zacarías-Lara, O. J., Correa-Basurto, J., and Bello, M. (2016). Exploring the conformational and binding properties of unphosphorylated/phosphorylated monomeric and trimeric BCL-2 through docking and molecular dynamics simulations. *Biopolymers* 105, 393–413. doi: 10.1002/bip.22839

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 do Carmo, Bettanin, Oliveira Almeida, Pantaleão, Rodrigues, Homem-de-Mello and Honorio. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 09 April 2020
doi: 10.3389/fchem.2020.00167

Integrating Ligand and Target-Driven Based Virtual Screening Approaches With *in vitro* Human Cell Line Models and Time-Resolved Fluorescence Resonance Energy Transfer Assay to Identify Novel Hit Compounds Against BCL-2

Gurbet Tutumlu[1], Berna Dogan[1], Timucin Avsar[2,3,4]*, Muge Didem Orhan[3,4], Seyma Calis[4,5] and Serdar Durdagi[1,3]**

[1] Computational Biology and Molecular Simulations Laboratory, Department of Biophysics, School of Medicine, Bahcesehir University, Istanbul, Turkey, [2] Department of Medical Biology, Bahcesehir University, School of Medicine, Istanbul, Turkey, [3] Neuroscience Program, Health Sciences Institute, Bahcesehir University, Istanbul, Turkey, [4] Neuroscience Laboratory, Health Sciences Institute, Bahcesehir University, Istanbul, Turkey, [5] Molecular Biology, Genetics and Biotechnology Graduate Program, Istanbul Technical University, Istanbul, Turkey

OPEN ACCESS

Edited by:
Simone Brogi,
University of Pisa, Italy

Reviewed by:
Rodolpho C. Braga,
InsilicAll, Brazil
Dharmendra Kumar Yadav,
Gachon University, South Korea

***Correspondence:**
Berna Dogan
berna.dogan@med.bau.edu.tr
Timucin Avsar
timucin.avsar@med.bau.edu.tr
Serdar Durdagi
serdar.durdagi@med.bau.edu.tr

Specialty section:
This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

Received: 08 November 2019
Accepted: 25 February 2020
Published: 09 April 2020

Citation:
Tutumlu G, Dogan B, Avsar T, Orhan MD, Calis S and Durdagi S (2020) Integrating Ligand and Target-Driven Based Virtual Screening Approaches With in vitro Human Cell Line Models and Time-Resolved Fluorescence Resonance Energy Transfer Assay to Identify Novel Hit Compounds Against BCL-2. Front. Chem. 8:167. doi: 10.3389/fchem.2020.00167

Antiapoptotic members of *B-cell leukemia/lymphoma*-2 (BCL-2) family proteins are one of the overexpressed proteins in cancer cells that are oncogenic targets. As such, targeting of BCL-2 family proteins raises hopes for new therapeutic discoveries. Thus, we used multistep screening and filtering approaches that combine structure and ligand-based drug design to identify new, effective BCL-2 inhibitors from a small molecule database (Specs SC), which includes more than 210,000 compounds. This database is first filtered based on binary *"cancer-QSAR"* model constructed with 886 training and 167 test set compounds and common 26 toxicity quantitative structure-activity relationships (QSAR) models. Predicted non-toxic compounds are considered for target-driven studies. Here, we applied two different approaches to filter and select hit compounds for further *in vitro* biological assays and human cell line experiments. In the first approach, a molecular docking and filtering approach is used to rank compounds based on their docking scores and only a few top-ranked molecules are selected for further long (100-ns) molecular dynamics (MD) simulations and *in vitro* tests. While docking algorithms are promising in predicting binding poses, they can be less prone to precisely predict ranking of compounds leading to decrease in the success rate of *in silico* studies. Hence, in the second approach, top-docking poses of each compound filtered through QSAR studies are subjected to initially short (1 ns) MD simulations and their binding energies are calculated via molecular mechanics generalized Born surface area (MM/GBSA) method. Then, the compounds are ranked based on their average MM/GBSA energy values to select hit molecules for further long MD simulations and *in vitro* studies. Additionally, we have applied text-mining approaches to identify molecules that contain "*indol*" phrase as many of the approved drugs contain

indole and indol derivatives. Around 2700 compounds are filtered based on *"cancer-QSAR"* model and are then docked into BCL-2. Short MD simulations are performed for the top-docking poses for each compound in complex with BCL-2. The complexes are again ranked based on their MM/GBSA values to select hit molecules for further long MD simulations and *in vitro* studies. In total, seven molecules are subjected to biological activity tests in various human cancer cell lines as well as Time-Resolved Fluorescence Resonance Energy Transfer (TR-FRET) assay. Inhibitory concentrations are evaluated, and biological activities and apoptotic potentials are assessed by cell culture studies. Four molecules are found to be limiting the proliferation capacity of cancer cells while increasing the apoptotic cell fractions.

Keywords: BCL-2, molecular docking, MD simulations, virtual screening, text mining, *in vitro* human cell line models, TR-FRET assay, binary QSAR models

INTRODUCTION

Finding a cure for cancer is still a challenging task, despite the understanding of molecular mechanisms and causal relationships participating in the pathology of cancer since the mid-1980s (Fesik, 2005). As stated by Hanahan and Weinberg, multistage development of tumors consists of six biological features widely known as hallmarks of cancer: (i) maintaining proliferative signaling, (ii) avoiding growth suppressors, (iii) triggering invasion and metastasis, (iv) empowering replicative perpetuity, (v) inducing angiogenesis, and (vi) resisting cell death (Hanahan and Weinberg, 2000, 2011). The ability of cancer cells to escape from programmed cell death, namely, apoptosis, remains a critical feature of these six indicators (Mohamad Rosdi et al., 2018). Apoptosis is a molecular pathway that results with self-destruction of the cell, either following termination of physiological function or after a crucial damage to genetic material (Igney and Krammer, 2002; Reed, 2002; Verma et al., 2015). The well-defined basic apoptosis pathways, extrinsic and the intrinsic pathways, are variously stimulated, and they use determined signaling elements (Kollek et al., 2016). The extrinsic pathway is activated by outer stimulation of death receptors. Death receptors are members of the tumor necrosis factor (TNF) receptor family, which has an intracellular death domain that is able to accumulate and trigger caspase-8 followed by operation of effector caspases including caspase-3, -6, or -7 (Youle and Strasser, 2008; Eimon and Ashkenazi, 2010; Wu et al., 2018). The intrinsic pathway, also called mitochondrial pathway, is initiated by a variety of cytotoxic damages or growth signals, some of which are genetic instability, inadequate developmental stimulation, and invasion by viral pathogens. *B-cell leukemia/lymphoma*-2 (BCL-2) family proteins tightly regulate this process and subsequently leads to the activation of caspase-9 (Cory et al., 2003; Youle and Strasser, 2008; Eimon and Ashkenazi, 2010).

All members of BCL-2 protein family have retained sequence patterns regarded as the BCL-2 homology (BH) domains and could be divided into three main classes. The first class of proteins are made up of the proapoptotic activator BH domain 3 (BH3) only proteins such as BIM, BID, and PUMA. Immediately upon their activation, they serve as molecular guardians that connect outer spurs to the mitochondrial pathway. The following group contains the proapoptotic effectors, which are multidomain proteins, such as BAX and BAK, and each of them has three BH domains. These proteins distort the integrity of mitochondrial outer membrane, which leads to free movement of cytochrome C to cytoplasm, initiates downstream caspase activity, and ultimately to trigger the termination of cells. The last class of BCL-2 family are the antiapoptotic protein, BCL-XL, BCL-2, MCL-1, etc. All of these members consist of four BH domains and keeps cells safe by segregating their proapoptotic peers. The most important point in promoting apoptosis is to increase the amount of BH3-only proteins or switch off one of its antiapoptotic BCL-2 counterparts (Fesik, 2005; Dewson and Kluck, 2010; Chung, 2018; Mohamad Rosdi et al., 2018). The idea of BH3 mimetics as promising anticancer drugs is inspired by the conclusion that a great deal of cancers rely on BCL-2 family proteins and that the interaction between these proteins occurs through specific BH domains (Oltersdorf et al., 2005; Soderquist and Eastman, 2016). A genuine BH3 mimetic is expected to imitate the BH3 domain of a proapoptotic BCL-2 protein, thus deactivating the antiapoptotic family members by filling up their BH3-binding pockets.

Apoptotic cell death is an innate hurdle to growth of tumor cells; hence, one of the fundamental hallmarks of cancer cells is the avoidance of apoptosis, which comprises a crucial process in resistance to chemotherapeutics. This phenomenon led to peculiar approaches in anticancer therapies focusing on apoptosis such as suppression of survival factors that are detected to be overexpressed in numerous malignancies. In the group of survival factors, BCL-2 proteins are one of the families that step forward for drug discovery studies (Lessene et al., 2008; Hanahan and Weinberg, 2011; Billard, 2013). For example, a small molecule, named ABT-737, was issued as a potential inhibitor of BCL-2 and BCL-XL, which occupies their BH3-binding domain and further triggers apoptosis in diversified cancer types (Tse et al., 2008; Soderquist and Eastman, 2016). Ensuing pharmaceutical trials guided to clinical studies with ABT-263 (navitoclax), which had boosted bioavailability and indicated efficacy in leukemia

and a few other neoplasias. However, they also manifested toxicities such as neutropenia and thrombocytopenia, leading to dose limitations (Tse et al., 2008; Gandhi et al., 2011). The thrombocytopenia was connected to the blockage of BCL-XL, as BCL-XL is essential for survival of platelets (Zhang et al., 2007). More recently, ABT-199 (venetoclax) was designed as a selective BCL-2 inhibitor and it evades the issue of thrombocytopenia (Souers et al., 2013). However, it also carries some side effects such as diarrhea, nausea, low white blood cell counts, high K^+ ion concentrations in the blood, headache, etc. Thus, novel BCL-2 inhibitors with better pharmacodynamic as well as pharmacokinetic profiles are needed.

It is well-established, especially in the last years, that taking a new drug into the market is both a time consuming and costly process. As a result, computer-aided drug design techniques have become prominent in drug development process (Lionta et al., 2014; Yoshino et al., 2015, 2017; Chiba et al., 2017; Halim et al., 2017; Durdagi et al., 2018a, 2019; Fu et al., 2018; Is et al., 2018; Mirza et al., 2018; Erol et al., 2019; Zaka et al., 2019). Different strategies depending on the availability of target molecules have been developed; structure-based drug design in which the target structure is known and ligand-based drug design that could be applied in cases that target structure is not known. It is also possible to combine both approaches to increase the possibility of "hit molecule" discovery as has been done in our previous studies (Durdagi et al., 2018a,b; Zaka et al., 2018; Kanan et al., 2019; Mollica et al., 2019). The most widely used technique in target-driven-based drug design is the molecular docking, and there are various docking programs as well as many different scoring functions to rank binding poses. Large molecule libraries can be screened using high throughput virtual screening, and lead compounds can be identified for further studies, quickly. By the use of more sophisticated docking algorithms and scoring functions, binding modes of compounds to target can also be determined. However, as expected, each of the docking algorithms and scoring functions have their own strengths and weaknesses. Numerous studies have been conducted to evaluate the comparative assessment of the docking and scoring functions (Bissantz et al., 2000; Bursulaya et al., 2003; Chen et al., 2006; Warren et al., 2006; Cross et al., 2009; Li et al., 2014). The latest evaluation study was conducted by Li et al. for 20 scoring functions on a diverse set of protein–ligand complexes (Li et al., 2014). Their comparison of scoring functions was based on four aspects: "scoring power" (binding affinity prediction), "ranking power" (relative ranking prediction), "docking power" (binding pose prediction), and "screening power" (discrimination of true binders from random molecules). Their results showed that scoring functions were generally more promising in docking and screening power tests than scoring and ranking power tests. In addition, scoring functions that were among top-ranked in docking power test were also more successful in screening power test but poor in other two power tests. This study, which shows that every scoring function has its own weaknesses, has represented that the ordering of compounds only by their docking scores may not accomplish the correct ranking of compounds; hence, if the molecules will only be selected according to their top docking scores for further studies such as *in vitro* tests, this may lead to false positive results (Rastelli et al., 2009; Rastelli and Pinzi, 2019). Therefore, in this study, we use another approach in ranking compounds that is based on molecular dynamics (MD) simulations and molecular mechanics generalized Born surface area (MM/GBSA) calculations after initial pose prediction by molecular docking.

In the present study, in order to identify novel BCL-2 inhibitors, ligand- and target-driven-based techniques were integrated with text mining approach, and novel hit molecules were identified with the virtual screening of small molecules library (Specs SC) that includes more than 212,000 compounds. In the identification of hits, two different approaches were considered: (i) Compounds were ranked by their docking scores, and MD simulations for 100 ns were carried out for the selected compounds and average MM/GBSA energies were calculated; (ii) Short (1-ns) MD simulations were applied for top-docking poses of all selected 342 compounds from binary quantitative structure-activity relationships (QSAR) models, and average MM/GBSA scores from short MD simulations were calculated. The average MM/GBSA scores were considered in the selection of compounds for longer MD simulations (100 ns) followed by MM/GBSA calculations. Additionally, it is known that many currently used Food and Drug Administration (FDA)-approved chemotherapeutics include indole fragment. To increase the probability of discovering hit molecules with potential anticancer properties, we screened Specs-SC database to identify molecules that contain *"indol"* groups by using text mining. Around 2700 compounds were screened against BCL-2, and novel hits that includes *"indol"* fingerprints were identified.

MATERIALS AND METHODS

Binary QSAR Models

MetaCore/MetaDrug (MC/MD) platform from Clarivate Analytics provides a comprehensive tool to analyze the pharmacodynamic and pharmacokinetic profiles for screening molecules. Using MC/MD, it is possible to calculate "therapeutic activity values (TAV)" of molecules for 25 common diseases including cancer by binary QSAR disease models. Additionally, toxicities of compounds could also be predicted in 26 different toxicity QSAR models using MC/MD. The *Tanimoto Prioritization (TP)* feature was applied to detect similarity between compounds and training and test set molecules analyzed in QSAR models based on fragments within the structure. QSAR models in the platform were constructed using various compounds based on experimental evidence of their activity/function on a particular protein of interest and then tested with validation sets. Estimated QSAR values (normalized between 0 and 1) >0.5 indicate potential therapeutic activity. The details about QSAR models could be found in the following reference (Kanan et al., 2019). In the current study, we used *"cancer-QSAR"* model, which has the following parameters: Training set $N = 886$, Test set $N = 167$, Sensitivity $= 0.89$, Specificity $= 0.83$, Accuracy $= 0.86$, MCC $= 0.72$.

Ligand Preparation and Protein Preparation

The compounds screened in this study (212, 520 molecules) were downloaded from Specs SC database (https://www.specs.net/index.php). 2D structures were used in binary QSAR models both for therapeutic activity prediction and toxicity prediction. After the target-driven screening and toxicity tests, 250 compounds were filtered and these molecules were prepared with the OPLS2005 forcefield using LigPrep module (Schrödinger Release 2015-2, 2015) of Maestro program (Banks et al., 2005). The possible ionization states at neutral pH 7.4 was determined by Epik module (Shelley et al., 2007). All possible tautomers as well as stereoisomers (if any) were generated. At the end, 342 structures were obtained and used in further docking and MD simulations. Two structures of BCl-2 solved by X-ray diffraction [Protein Data Bank (PDB) IDs, 4LXD (Souers et al., 2013), and 6GL8 (Casara et al., 2018)] along with two structures solved by NMR spectroscopy were retrieved from the PDB [1YSW (Oltersdorf et al., 2005) and 2O2F (Bruncko et al., 2007)]. Here, it should be mentioned that BCL-2 has a region predicted to adopt an unstructured and flexible loop, which caused the protein to be insoluble (Petros et al., 2001; Bruncko et al., 2007). Hence, in NMR studies, as first suggested by Petros et al., residues 35–91 were replaced with residues 35–50 from BCL-XL, and the C-terminal end (residues 208–219) was truncated (Petros et al., 2001). The resulting chimeric protein was very soluble, while still retaining its biological activity. Moreover, a 3D structure of BCL-2 with an intact loop region would be obtained. For crystal structures, although chimeric protein was used, the unstructured loop region could not be resolved due to low electron density and the fact that the loop region was not connected. As this could cause problems during MD simulations, we took the loop conformation from the NMR structure (PDB, 1YSW) for our modeling studies. The numbering of residues was based on the crystal structure with PDB code 4LXD. The missing atoms of proteins were added, and the ions, small molecules used to aid in crystallization, and water molecules not near the cocrystallized ligand (>5 Å) were removed using the Protein Preparation module of Maestro (Sastry et al., 2013). PROPKA (Bas et al., 2008) was employed to adjust protonation states of amino acids at pH of 7.4, and finally, in order to relax the proteins, the target protein was minimized employing the OPLS2005 forcefield parameters (Banks et al., 2005). The binding pocket of BCL-2 was classified based on cocrystallized ligands, and the residues in these regions, together with water molecules, were considered in the construction of grid lattice boxes in molecular docking.

Molecular Docking Simulations

The docking algorithms used in this study include standard precision (SP) module of Glide (Friesner et al., 2004; Halgren et al., 2004) and Induced Fit Docking (IFD) module of Maestro (Sherman et al., 2006a,b) with flexible ligand sampling. The IFD method consists of three consequent phases, including (i) docking of the compounds while the receptor is rigid; (ii) refining the complex residues within 5 Å of the ligand using Prime module (Jacobson et al., 2004); and finally, (iii) redocking of the compounds at the refined binding pocket.

Molecular Dynamics (MD) Simulations and Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) Calculations

We performed MD simulations for apo form of BCL-2 and complexes of BCL-2 with hit compounds using Desmond program (Bowers et al., 2006). Protein–ligand complexes were placed in the cubic boxes with explicit TIP3P water models that have 10.0 Å thickness from surfaces of protein. All systems are neutralized by adding counter ions (Na^+ or Cl^- depending on the charge of the systems), and salt solution of 0.15 M NaCl was also used to adjust the concentration of the systems. The long-range electrostatic interactions were calculated by the particle mesh Ewald method (Essmann et al., 1995). A cut-off radius of 9.0 Å was used for both van der Waals and Coulombic interactions. The temperature was set as 310 K initially, and Nose–Hoover thermostat was used for adjustment (Nosé, 1984; Hoover, 1985). Martyna–Tobias–Klein protocol was employed to control the pressure, which was set at 1.01325 bar (Martyna et al., 1994). The time-step was assigned as 2.0 fs. The default values were used for minimization and equilibration steps, and finally 1-ns (for short MD simulations) and 100-ns (for long MD simulations) production run was performed for each simulation. Other details of the simulation protocols were described in our previous studies (Durdagi et al., 2016; Salmas et al., 2017; Rodrigues et al., 2018). The Prime module of Schrodinger (Jacobson et al., 2004) was used in binding free energy calculations of complexes by MM/GBSA approach (Bashford and Case, 2000). 100 trajectory frames from all MD simulation times were considered for short MD simulations. For longer MD simulations, on the other hand, 100 trajectory frames from the last half of the simulations were used for MM/GBSA calculations. OPLS3 forcefield (Banks et al., 2005) and VSGB 2.0 solvation model (Shan et al., 2011) were utilized during MM/GBSA calculations. All applied procedures for virtual screening in this study have been summarized in **Figure 1**.

Time Resolved Fluorescence Resonance Energy Transfer (TR-FRET)

The BCL-2 TR-FRET Assay (CISBio, Cat. No: 79601) was used to measure the inhibition of BCL-2 to bind its ligand in the presence of BCL-2 inhibitory molecules in a homogeneous 384 well-reaction format. The assay protocol for TR-FRET analysis was performed based on the suggestions of the manufacturer. Briefly, a sample containing terbium-labeled donor, dye-labeled acceptor, BCL-2 protein, peptide ligand, and each inhibitor were incubated for 2 h. All samples and controls were studied in triplicate. Tested concentrations for the molecules were 1 nM, 10 nM, 100 nM, 1 μM, and 10 μM. The fluorescence intensity was measured using a fluorescence reader (Varioskan LUXTM, Thermo Fischer). Two sequential measurements were conducted. First, terbium-donor emission was measured at 620 nm followed by dye-acceptor emission at 665 nm; each fluorescent read was excited at 344 nm. Data analysis was performed using the TR- FRET ratio (665 nm emission/620 nm emission) value. The percent inhibitory activity of tested molecules was calculated by:

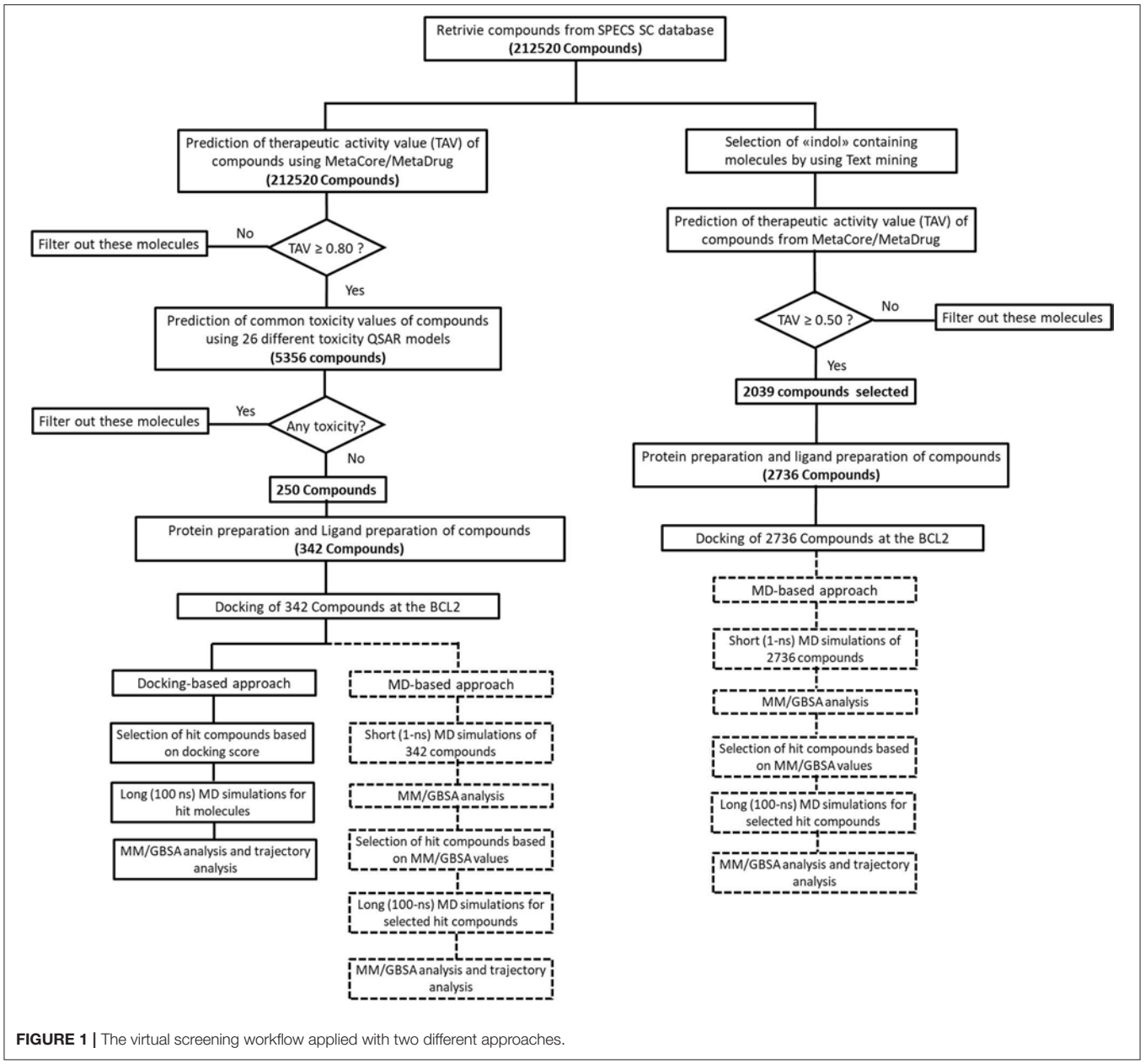


FIGURE 1 | The virtual screening workflow applied with two different approaches.

$$\%Activity = \frac{FRETs - FRETneg}{FRETp - FRETneg} x100\%$$

where *FRETs*, *FRETneg*, and *FRETp* are sample FRET, negative control FRET, and positive control FRET, respectively.

Cell Culture Experiments and 3-(4 5-dimethylthiazol-2-yl)-2 5-diphenyltetrazolium bromide (MTT) Analysis

Various cancer cell lines, such as HCT-116 colon cancer, U87-MG glial tumor, and MCF7 breast cancer cell lines, were used for cell culture experiments. Cells were seeded with high glucose *Dulbecco's* Modified Eagle Medium (DMEM) medium (Biosera) supplemented with 10% fetal bovine serum (FBS) (Gibco) and 1X penicillin/streptomycin (Multicell). Twenty-four hours prior to molecule treatment, 10,000 cells were seeded into each well of 24-well-cell culture plates. Molecules were used by preparing 4 mM stock solution in dimethyl sulfoxide (DMSO) (Amresco). For molecule treatment, molecule stock solutions were diluted in DMEM with 10% FBS and added onto cells in each corresponding well. Final concentration of vehicle DMSO was 2% at maximum. Therefore, the vehicle group in experiments only included a maximum of 2% DMSO concentration. We determined the number of cells to be seeded to make sure that none of the cells reaches more than 60% confluency during the treatment period, as higher plate confluency levels would slow down cell proliferation independently from the molecule treatment.

Values of half-maximal inhibitory concentration were (IC_{50}) determined by MTT cell proliferation assays. Different concentrations of molecules ranging between 10^{-9} and 10^{-4} M were tested on used cell lines with single treatment. Five hundred seventy nanometer absorbance values were recorded, and IC_{50} values were calculated by dose–response inhibition curves and non-linear regression analysis on GraphPad Prism 8 software. For cell proliferation assays, we performed 5-day experiments and repeated experiments at least three times with all cell lines. Survival rates did not change significantly after third day of treatment. Therefore, 3 days results were presented. MTT analysis was performed on 24-well plates with initially 1×10^4 cells/well, grown overnight, and then treated with the selected molecules with different concentrations for at least 3 days. Following the initial incubation day, molecules were added, and, after incubation with MTT at 37°C for 4 h, formazan was solubilized with DMSO (Sigma-Aldrich, St. Louis, USA) and absorbance was measured at 570 nm.

RESULTS

In this work, a small molecule library (Specs-SC) that has 212,520 available drug-like compounds as well as small molecules extracted from available literature were screened initially in MC/MD platform. The molecules were filtered based on their TAV against "cancer" disease model predicted by cancer-QSAR model of MetaCore. The cancer-QSAR model was constructed with 1,053 known compounds from literature, and obtained statistical results were found as follows: Sensitivity: 0.89; Specificity: 0.83; Accuracy: 0.86; MCC: 0.72. Thus, as it can be seen, statistical results validate the constructed cancer-QSAR model. Moreover, we have selected 30 compounds that have high ($IC_{50} \leq 10\,\mu M$) inhibitory activity against BCL-2 based on biological and cell line assays to further validate used QSAR model as suggested in various previous studies (Kumar Yadav et al., 2013, 2014a,b). These molecules are also subjected to cancer-QSAR model to predict their TAV against "cancer." Results showed that 23 out of 30 known inhibitors (more than 75% of the known inhibitors) have potential therapeutic activity (i.e., TAV $\geq$ 0.5) against cancer. Although the predicted QSAR values higher than 0.5 could potentially indicate therapeutic activity, in this study, we considered a higher threshold/cut-off value ($\geq$0.8). By this way, we only considered the molecules that would be predicted as highly therapeutically active in "Cancer-QSAR" model of MC/MD. 5356 molecules had the predicted cancer therapeutic activity values equal or higher than 0.8. Compounds may show high binding affinity, but if they carry undesired side effects, they cannot be considered for future clinical studies. Therefore, their toxicity and pharmacokinetic profiles must be investigated. In our study, we used 26 different toxicity QSAR models and these diverse toxicity models cover most of the commonly observed toxicities such as cardiotoxicity, nephrotoxicity, neurotoxicity, cytotoxicity, kidney necrosis, liver necrosis, etc. Out of 5,356 identified molecules, only 250 molecules showed no toxicities in all these 26 different toxicity QSAR models. These 250 compounds were then prepared with ligand preparation module of Maestro, and, at the end, 342 structures in total, with the possible tautomeric and protonation states, were obtained. All these molecules were then used in molecular docking studies for target protein BCL-2 along with two reference molecules venetoclax and S55746 (bcl201). For S55746, the crystal structure was available; hence, after the preparation of the complex as explained in Materials and Methods section, it was subjected to MD simulations. There was no available crystal structure of BCL-2 that was cocrystallized with venetoclax when this study was conducted, though its analogs were available. Hence, venetoclax was also prepared with LigPrep and molecular docking was used to obtain complex BCL-2/venetoclax.

The common weaknesses of docking algorithms were established by the comparative evaluation study of Li et al. (2014) as mentioned above. Molecular docking method sometimes could lead to elimination of true binders and/or false positive compounds since only top-ranked molecules would be considered as selected candidates for *in vitro* tests. Protein structure being considered as mainly rigid during docking is one of the major weaknesses. As such, here we have applied two different approaches: (i) an induced fit docking in which residues in binding pocket were considered as flexible; (ii) short MD simulations in which protein–ligand complex was relaxed to dispose clashes between protein and ligand. **Figure 1** shows the workflow applied in this study. It can be seen that we have also identified compounds that contain "*indol*" phrases using text mining to be considered for ligand- and structure-based studies of BCL-2 inhibitors.

Docking-Based Approach for Selection of Hit Molecules

The top-docking scores of five molecules and their 2D structures as well as corresponding data for reference molecules venetoclax and S55746 could be found in **Table S1**. To test the validity and reliability of docking approach as performed in previous studies (Chen et al., 2006), we have also redocked the cocrystallized ligand found in the crystal structure of BCL-2 (PDB ID, 4LXD). We have seen that the docking pose obtained with our protocol was able to reproduce the crystal pose with a root mean square deviation (RMSD) of 0.65 Å (i.e., after alignment between docked pose and cocrystallized pose, RMSD was 0.65 Å). As it can be seen from **Table S1**, venetoclax has the highest docking score (−15.46 kcal/mol) at BCL-2 cavity. However, venetoclax is a very large molecule [molecular weight (MW), 868 g/mol] that contains 61 non-hydrogen atoms, which could lead to high docking score, and in fact its ligand efficiency score (i.e., docking score per number of non-hydrogen atoms) was lower than the suggested compounds (**Table S1**). The molecules with the top-docking scores were smaller than venetoclax and S55746. However, they have high ligand efficiency scores, which could indicate that they could be lead compounds for further studies. For that reason, we performed MD simulations (100 ns) for these compounds in complex with BCL-2 protein starting from IFD docking poses. Although we carried out MD simulations for all five complexes as well as for two reference molecules in complex with BCL-2, we

will only discuss the results for three compounds: **43** (Specs ID: AO-081/41887762); compound **58** (Specs ID: AJ-292/12931005); and compound **243** (Specs ID: AN-698/40780701), as they were chosen for *in vitro* studies.

MD-Based Approach for the Selection of Hit Molecules

The top-scoring docking poses for all 342 compounds were subjected to 1-ns short MD simulations, and the average binding free energies were calculated for MD trajectory frames using MM/GBSA approach. An in-house script was used for the preparation of simulation boxes as well as for the analysis of MD simulations. Compounds were then ranked based on average MM/GBSA scores, and, as shown in **Figure S1**, the normal distribution of MM/GBSA scores of studied 342 compounds and Z-scores of the distribution curves were plotted. Then, we selected compounds that have average MM/GBSA values above $Z \leq -2$, i.e., 12 molecules were chosen. The complexes of these compounds with BCL-2 were subjected to longer (100 ns) MD simulations. The structures and average MM/GBSA values of all these selected compounds could be found in **Table S2**. Although we have performed longer MD simulations for all 12 compounds in complex with the target protein, we selected only three of the compounds for *in vitro* tests, **258** (Specs ID: AK-968/12163470), **292** (Specs ID: AK-968/11842328), and **243** (Specs ID: AN-698/40780701). It should be noted that compound **243** was also found as a hit compound and selected based on docking approach.

Text Mining Approach for Selection of Hit Compounds

Since many currently used FDA-approved chemotherapeutics include indole derivatives, Specs-SC database that includes only *"indol"* groups (i.e., indoles, indolons, bisindoles, etc.) were also screened at the binding pocket of the BCL-2. Thus, *"indol"* keyword was searched as text within the 212,000 compounds and around 2700 compounds were identified. These *"indol"* phrase containing molecules were subjected to binary QSAR tests using MC/MD platform and specifically the "cancer-QSAR" model was chosen as before. Since indole derivatives are known to have high therapeutic activity potential, here we initially used a lower TAV threshold (0.5) to begin screening with a large number of molecules that include the *"indol"* phrase. Molecules that showed higher TAV values than 0.5 were docked at the binding pocket of BCL-2 using Glide/SP. Top-docking poses of these compounds were then used in MD simulations. 2700 individual MD simulations boxes were prepared with an in-house script, and 1-ns MD simulations were conducted and the average MM/GBSA scores were calculated. The normal distribution of MM/GBSA values as well as *Z*-scores of the distribution showed that there were 83 compounds with Z-scores lower than −2 (see **Figure 2**). As performing 100-ns MD simulation for all 83 complexes would require considerable computer time and power, we instead chose to perform 10-ns MD simulations for these compounds in complex with BCL-2 and again used MM/GBSA approach to calculate their average binding free energies. After 10-ns MD simulations, top-10 MM/GBSA-scored *"indol"* phrase containing molecules were forwarded for 100-ns MD simulations and their average MM/GBSA scores were calculated. **Table S3** shows the 2D structures and average MM/GBSA scores for these 10 compounds. We have selected two of them for *in vitro* studies: **ind-199** (AG-205/12549135) **and ind-435** (AN-329/13484046).

Analysis of Selected Compounds and Their Interactions With BCL-2

Although molecular docking studies could give an initial insight into protein–ligand interactions, it is always crucial to understand the maintenance of these interactions and perform dynamical studies as MD simulations for complexes. Hence, we performed MD simulations and analyzed the interactions observed during the simulations between protein and ligands. While we conducted 100-ns MD simulations for 29 compounds in total (including the reference molecules) in complex with BCL-2, we selected seven of them for *in vitro* studies based on their docking scores, MM/GBSA values, and their interactions with the target protein. Here, we will focus our analysis and discussion on these seven compounds that could be lead compounds as BCL-2 inhibitors. Before analyzing the ligand–protein interactions, the trajectories obtained from the simulations were firstly analyzed to examine the protein and ligand structure stability. RMSD and the root mean square fluctuations (RMSF) were used to measure the displacements of atoms for each frame with respect to the initial frame/structure and to categorize the local changes along protein structures, respectively (**Figures S2, S3**). Here, we have only plotted the RMSD graphs of studied proteins based on alpha carbons (C_α). As it can be seen from the figure, for compounds other than *"indol"* phrase containing ones, the RMSD plots do not change significantly after 50-ns and they reach a plateau. However, mainly in indol-containing molecules, RMSD values did not stabilize and conformational changes were observed during MD simulations (**Figure S4**). It can be seen that it was the unstructured loop region 31–89, not alpha-helix regions, that had higher displacement (**Figures S3, S4**). For **ind-199**, some unexpected helix formation is observed for this region, though the helix could not be conserved. The RMSF plot for protein targets in complex with selected compounds also confirmed that it was the loop region for which highest displacements were observed (**Figure S3**). Additionally, we checked the RMSD of the ligand molecules by considering two different fitting modes: "fit on protein/profit" and "fit on ligand/ligfit." While the first mode indicates the structural stability of ligand with respect to protein, i.e., its translational motion, the second mode shows the internal fluctuations of the ligand atoms in its binding pocket, i.e., its rotational motion. As can been seen from **Figure 3**, the profit RMSD plot shows that after initial 50 ns, most of the compounds did not move away from the binding pocket. However, compound **58** had a very high RMSD value (around 7.0 Å), which showed that it has high mobility in the binding pocket. In fact, **Figure S4** shows that the compound **58** completely changed its initial binding pose in the hydrophobic groove just after initial 20 ns, but then it did not have high mobility and stayed in the pocket as can be seen by lower

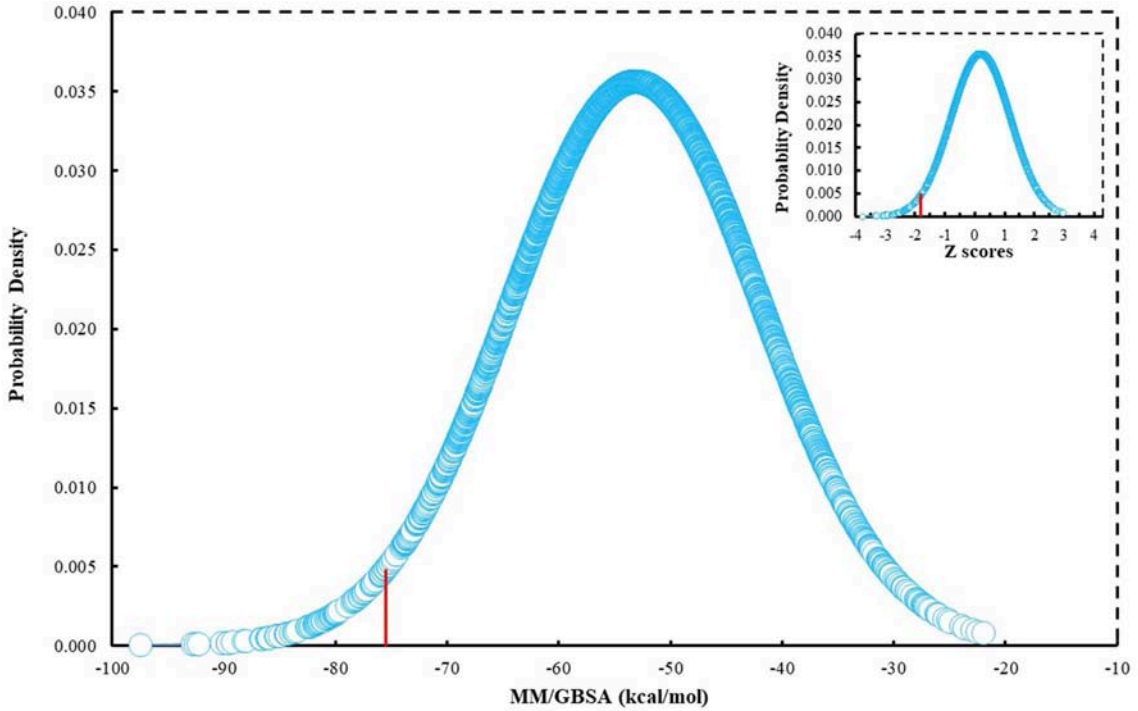


FIGURE 2 | MM/GBSA scores of around 2700 molecules that include "indol" phrases.

RMSD values and also small conformational changes. Indol-containing molecule **ind-435** also displayed high mobility in the binding pocket as can be observed in **Figure 3** and **Figure S4**. The RMSD for this compound did not really reach a plateau value. The size of this compound was actually considerably bigger than other molecules and has flexible regions. Hence, it extended from its initial binding pocket to the one next to it (e.g., hydrophobic grooves P1 to P4). Compound **243** has also higher profit RMSD values especially after 80 ns, and careful analysis of MD trajectories showed that it was mostly the phenyl ring that was attached thiazolidine group moving in pocket P1. The rotational movements (ligfit RMSDs) of the selected compounds could be seen in **Figure 4**. We observed that venetoclax did not also obtain a stable RMSD plot, though the values themselves were not higher than 3.0 Å for ligfit mode. As a large molecule, these rotational movements were not surprising. It was also not unexpected for **ind-435** as a large molecule with flexible alkyl chain to have high rotational RMSD values as seen in **Figure 4**. Compound **58** was found with higher RMSD values for ligfit mode; however, after 50 ns, the changes in RMSD values were smaller. All of the compounds at the end reached a kind of plateau value for rotational RMSD of ligands.

BCL-2 protein interacts with BH3-only proteins via hydrophobic groove on its surface, which contains four pockets: P1, P2, P3, and P4 pockets. To prevent the interaction of proapoptotic proteins such as BAX and BAK with BCL-2, which is an antiapoptotic protein, these pockets need to be filled by either small molecules or BH3-mimetics. It is important for these molecules to interact with key residues that mediate interaction between BCL-2 and BH3-only proteins. Based on the literature data, some of these crucial residues are as follows (numbering based on PDB code 4LXD): Asp100, Phe101, Arg104, Tyr105, Asp108, Phe109, Tyr199, Asn140, Gly142, Arg143, and Ala146. Additionally, we have analyzed the MD simulations of reference compounds venetoclax and S55746 in complex with BCL-2. The protein surfaces as well as 2D and 3D ligand interactions between protein and molecules were represented for venetoclax and S55746 in **Figure 5** and **Figure S5**, respectively. Consistent with the previously published data, S55746 bind and fill the pockets P1 to P2, while venetoclax could fill all four pockets on the surface from P1 to P4 (Birkinshaw et al., 2019). When this project was initiated, there was no cocrystallized venetoclax-bound form of the BCL-2. Here, we also checked the binding pose of venetoclax as the crystal structure of BCL-2 bound to venetoclax recently became available (Birkinshaw et al., 2019). A slight difference in binding pose of venetoclax was in the more flexible region of oxane fragment; however, with the rest of the compound, similar amino acid moieties interact by the identical parts of venetoclax in both poses. Based on the trajectory analysis of venetoclax, it was seen that venetoclax preserved its interactions with Asp100 and Phe101 more than 50% of the simulation time (**Figure S5**). Gly142 and Arg104 were also seen as interacting residues. Interaction with the backbone carbonyl oxygen atom of Ala143 was the most conserved interaction during MD simulations for S55746 (99%, **Figure 5**). Also, Arg143 and Phe101 formed π-cation and π-π stacking interactions with S55746, respectively. Based on these results, we analyzed the complexes of BCL-2 protein with selected hit

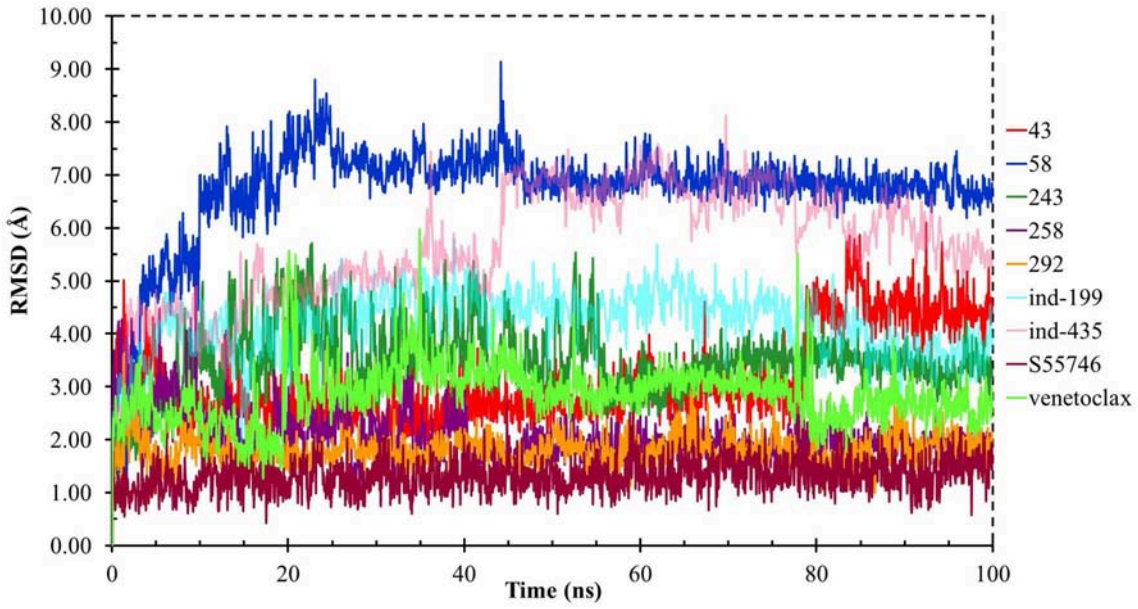


FIGURE 3 | ProFit RMSD graphs for the hit and reference compounds.

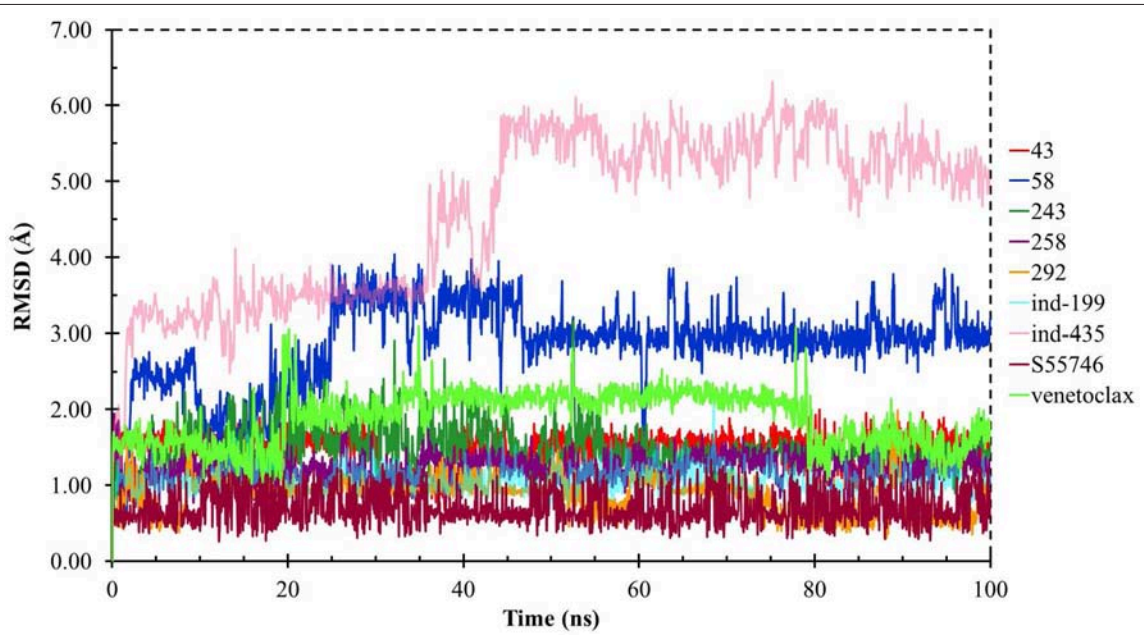


FIGURE 4 | LigFit RMSD graphs for the hit and reference compounds.

compounds as well as examined their ability to fill the pockets P1 to P4. Compound **43,** which can mainly bind BCL-2 via P2 and P3 pockets, consistently formed hydrogen bonds with Asp137 (96%) and π-π stacking interactions with residues Phe101 and Tyr105 of P2 pocket, although these later interactions were not preserved (16 and 24% of MD time, **Figure S6**). Compound **58** can mainly fill the pockets P1 to P3 of BCL-2 similar to S55746. It formed stable hydrogen bonding interactions with two key residues Asp108 and Asn140 (64 and 67%, respectively, **Figure 6**). Compound **243,** on the other hand, interacted with the residues of pockets P1 to P4. A π-π stacking interaction with Phe101 of P1 pocket was observed for 49% of MD simulation time, while its interactions with other key residues such as Tyr105, Asn140, and Arg143 were less conserved (**Figure S7**). Compound **258** was mainly found in pockets P2 to P4, and its most conserved interaction was observed to be with Tyr105.

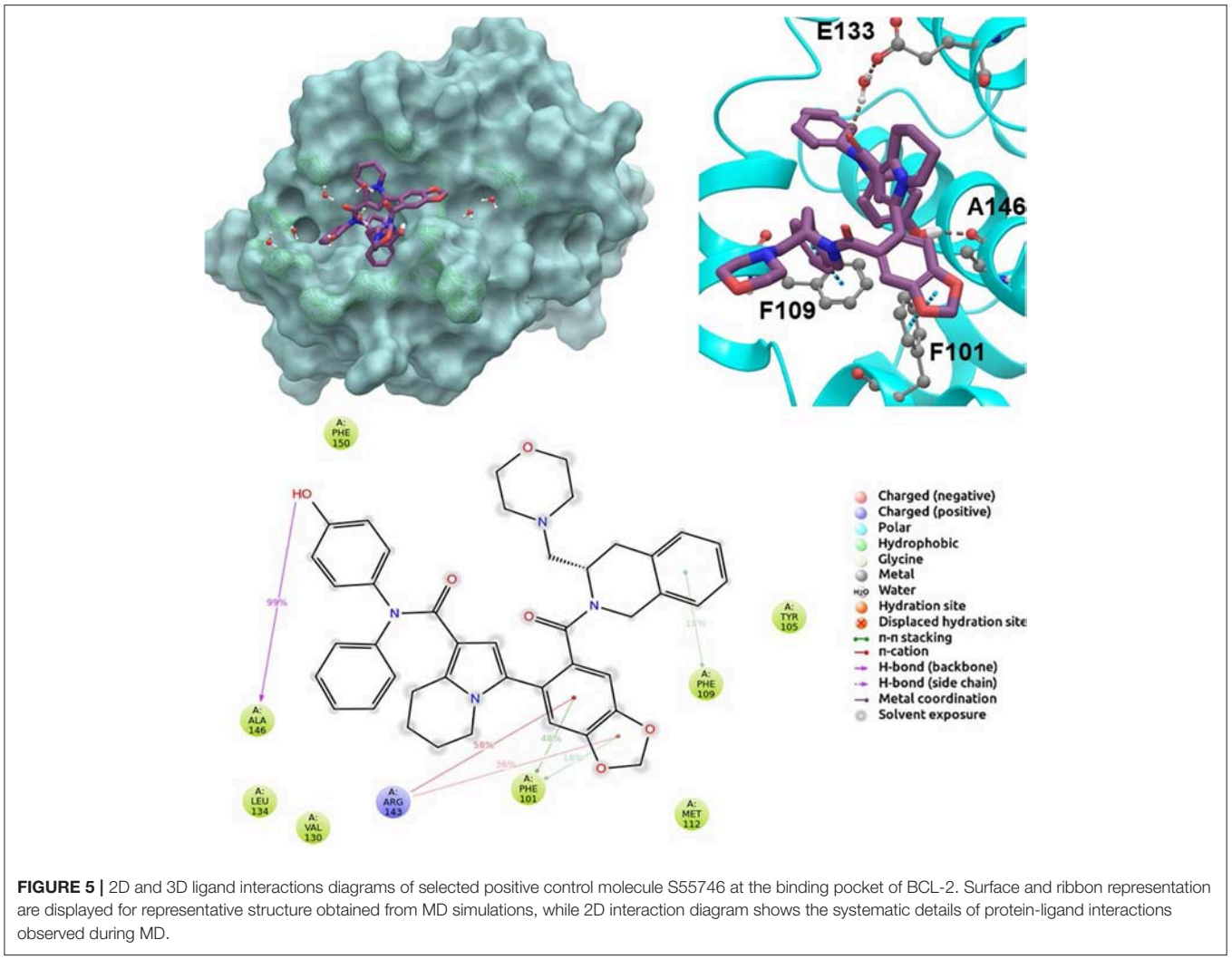


FIGURE 5 | 2D and 3D ligand interactions diagrams of selected positive control molecule S55746 at the binding pocket of BCL-2. Surface and ribbon representation are displayed for representative structure obtained from MD simulations, while 2D interaction diagram shows the systematic details of protein-ligand interactions observed during MD.

Not only did it form hydrogen bonds with Tyr105 but it also formed π-π stacking interactions, albeit they were not persistent interactions (17%, **Figure S8**). Compound **292** was an analog of compound **258;** as such, similar binding poses were expected, but their binding modes were quite different. Compound **292** formed persistent hydrogen bonding interactions with Asp108 (73%, **Figure S9**). H-bond and π-cation interactions with Arg143 were also observed for compound **292**. The chosen *"indol"* containing molecules were larger molecules compared to other five selected molecules; as such, they are able to fill the pockets P1 to P4. Compound **ind-199** formed a stable hydrogen bond with Asn140 (70%), π-cation, and salt bridge interactions with Arg104 (**Figure S10**). Although it also formed π-π stacking interactions with Phe109 and Phe150, these interactions were not persistent (13 and 22%, respectively). Compound **ind-435** not only filled all four pockets but also moved closer to carbonyl terminal part of BCL-2. A persistent hydrogen bond with Asp100 (55% of MD time) was observed, and additionally it interacts and with residues Phe100 and Arg143 (**Figure S11**).

We have also calculated the binding free energies for selected compounds as well as reference molecules in complex with BCL-2 using MM/GBSA approach after 100 ns MD simulations. In **Figure 7**, the MM/GBSA energies calculated for the trajectory frames observed during MD were plotted. It can be seen that venetoclax had lower MM/GBSA values compared to selected compounds. However, some of the selected compounds such as compound **243**, **ind-435,** and **ind-199** have considerable MM/GBSA values to other reference molecule S55746.

TR-FRET Analysis Confirms the Inhibitory Activity of Identified Hit Molecules on BCL2

TR-FRET analysis revealed that four of seven tested molecules compete with BCL2 ligand in binding. In presence of inhibitory molecules, BCL2 binding to its ligand was suppressed in a concentration- dependent manner. Compounds **58**, **ind-199**, **243**, and **292** showed the maximum inhibitory effect in ranging between 60 and 100% in 10 μM concentrations

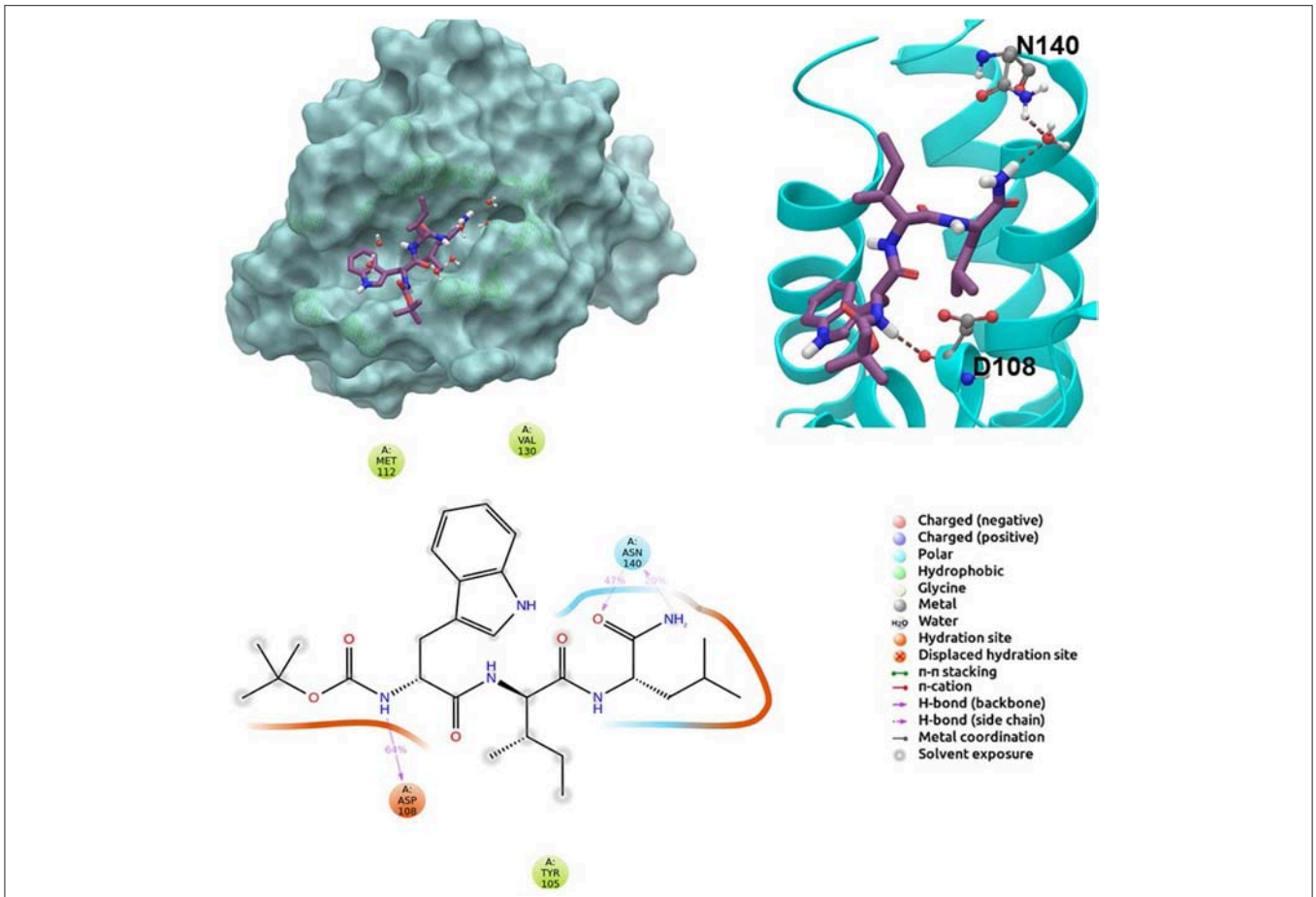


FIGURE 6 | 2D and 3D ligand interactions diagrams of selected hit molecule **58** at the binding pocket of BCL-2. Surface and ribbon representation are displayed for representative structure obtained from MD simulations, while 2D interaction diagram shows the systematic details of protein-ligand interactions observed during MD.

(**Figure 8**). However, three of selected molecules showed minimal inhibitory activity on BCL2 ranging from 10 to 40% with concentration-independent manner. We suppose that activity is correlated with solubility of the molecules since inactive molecules were found among partially soluble molecules.

Cell Proliferation Was Restricted by Using BCL2 Inhibitory Molecules

Together with TR-FRET analysis, selected hit compounds were also evaluated in various cancer cell lines, such as HCT-116 colon cancer, U87-MG glial tumor, MCF7 breast cancer cell lines, and IC_{50} values of selected molecules were calculated (**Table 1**). Of seven selected hits, five of them showed micromolar level of inhibitory concentrations, which is an acceptable range in cell culture experiments. Compound **258** did not show inhibitory activity on any of the tested cell-line assays.

To test whether BCL-2 inhibitory molecules had any effect on biological activity, we conducted cell proliferation assays and also evaluated apoptosis by observing cell structure and counting apoptotic cells. All molecules were tested on three different cancer cell lines, and all molecules showed similar effects on different cell types. The dose–response curves for all cell lines were shown in **Figures S12–S14**. Of seven molecules, four showed biological activity on MTT experiments. Molecules **58**, **ind-199**, **43,** and **243** with 100 μM concentration significantly limited the cell proliferation capacity of cancer cells. Four biologically active molecules showed their efficacy starting from the first hour of treatment by decreasing the number of proliferating cells. At the first day of treatment, only 60–70% of cells survived, while the number of proliferating cells decreased to <40% at the end of third day (**Figure 9**). MTT assay results for cell lines U87-MG and MCF7 were displayed in **Figures S15, S16**, respectively. Compared to untreated and DMSO only treated (vehicle) group, these four hit molecules showed significant activity. Lack of activity of other three compounds might be due to their low/moderate solubilities. Furthermore, the activity of the molecules was more dominant at cancer cells. We also tested these molecules on non-cancerous HUVEC cells, and none of the molecules showed significant reduction in cell viability (**Figure S17**). When we observed the cells under a microscope for the inactive molecules, we saw precipitates of molecules. Further, despite all efforts for

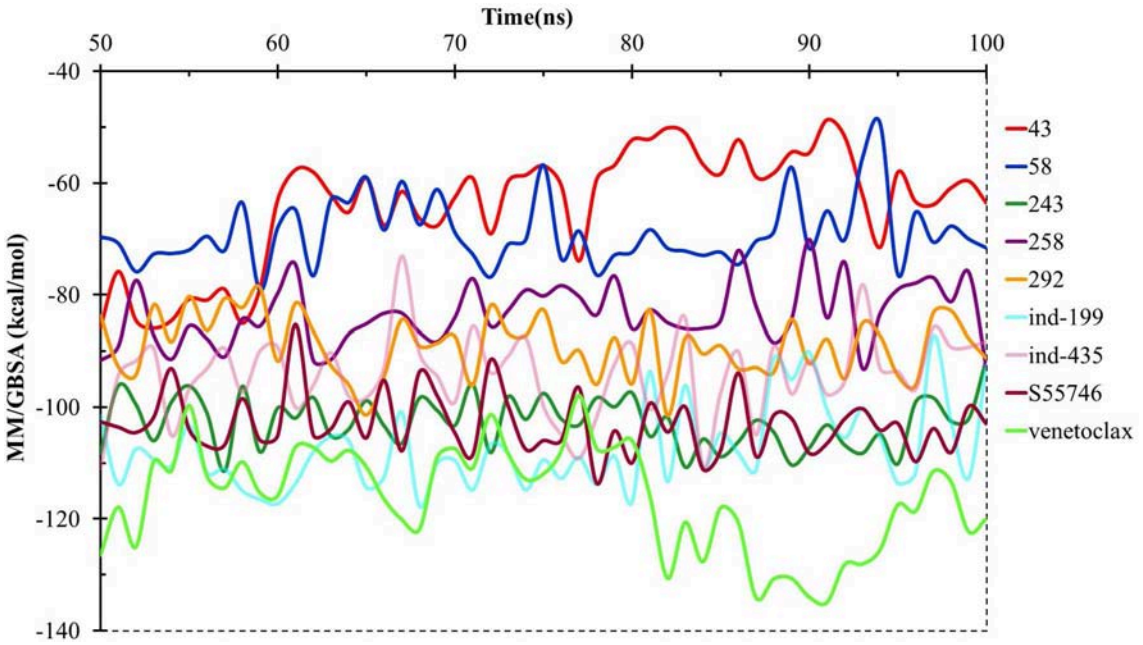


FIGURE 7 | MM/GBSA free energy analysis for the studied molecules at the binding pocket of BCL-2 throughout the last half of the MD simulations.

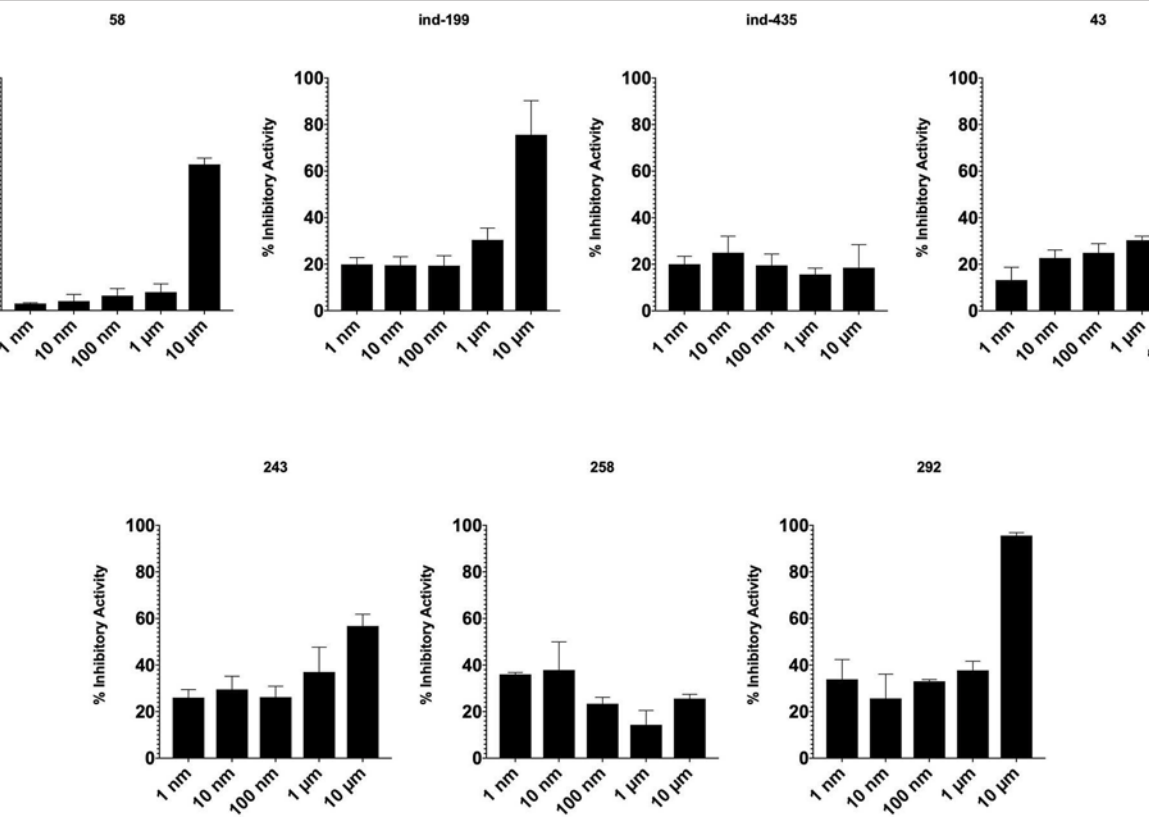
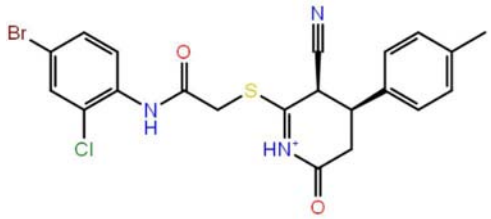
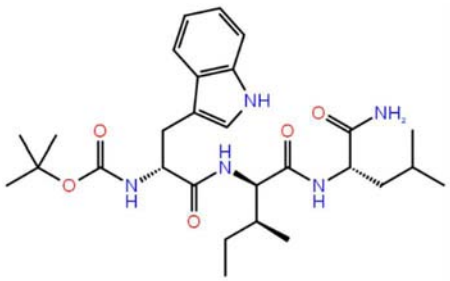
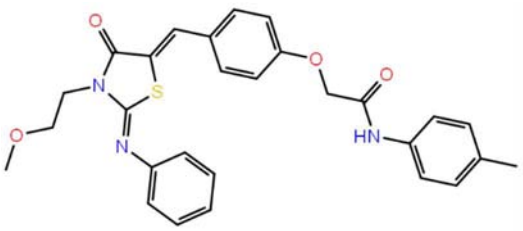
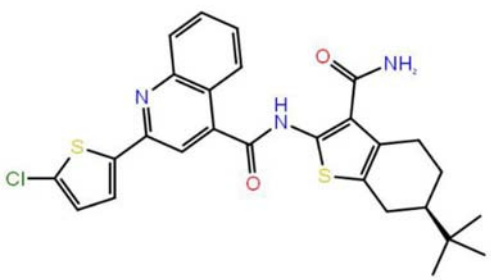
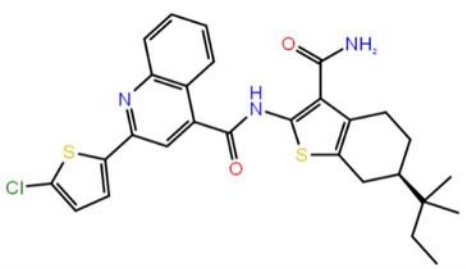


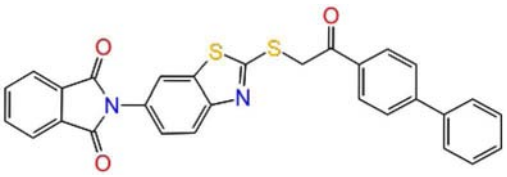
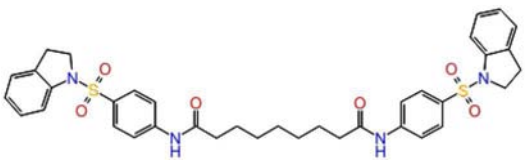
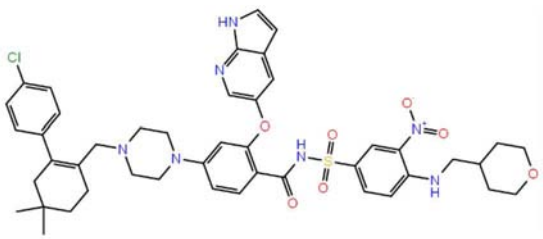
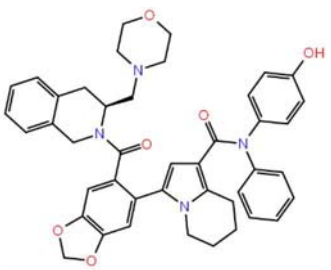
FIGURE 8 | Inhibitory activity of tested molecules. Activity of the molecules was assessed by the formula given in Materials and Methods section.

TABLE 1 | The specs ID, 2D structure, average MM/GBSA values, and maximal % inhibitory activity at 10μM concentrations, as well as IC_{50} values of selected hit compounds and reference molecules.

Compounds (specs ID)	2D structure	MM/GBSA (kcal/mol)	% inhibitory activity at 10 μM	IC_{50} (μM) for HTC116 cells	IC_{50} (μM) for U87-MG cells	IC_{50} (μM) for MCF7 cells
43 (AO-081/41887762)		−64.15 ± 10.37	41.78 ± 1.76	24 ± 2.53	26 ± 1.73	30 ± 8.27
58 (AJ-292/12931005)		−68.84 ± 6.03	62.84 ± 2.73	17 ± 1.59	18 ± 2.29	21 ± 3.34
243 (AN-698/40780701**)**		−102.65 ± 4.56	56.78 ± 5.02	25 ± 1.77	29 ± 2.55	31 ± 3.96
258 (AK-968/12163470)		−83.50 ± 5.46	25.60 ± 1.84	NA†	NA†	NA†
292 (AK-968/11842328)		−88.93 ± 5.46	95.63 ± 1.27	130 ± 25.59	122 ± 33.21	150 ± 33.50

(Continued)

TABLE 1 | Continued

Compounds (specs ID)	2D structure	MM/GBSA (kcal/mol)	% inhibitory activity at 10 μM	IC_{50} (μM) for HTC116 cells	IC_{50} (μM) for U87-MG cells	IC_{50} (μM) for MCF7 cells
ind-199 (AG-205/12549135)		−107.41 ± 7.49	75.64 ± 14.65	22 ± 3.41	28 ± 4.46	26 ± 4.89
ind-435 (AN-329/13484046)		−94.17 ± 7.28	18.51 ± 9.91	NA†	NA†	NA†
venetoclax		−116.31 ± 9.20	NA†	3.5*	22.9*	24.7*
S55746 (CHEMBL3958369)		−101.43 ± 4.20	NA†	NA†	NA†	NA†

†*Not applied.* **Taken from The Genomics of Drug Sensitivity in Cancer (GDSC) database (www.cancerRxgene.org) for HTC-116 cell line.*

increasing the solubility levels of these inactive compounds in DMSO, it failed to obtain a pure solubilized form of the molecules.

We also observed massive cell detachment and dying cells in cell culture plates (**Figure 10**), and formation of apoptotic cell bodies having circular structure rather than normal. Especially for the compound **58**, apoptotic cell bodies were abundant, and it immediately affected the cells upon the first hour of treatment. However, for the compounds **43** and **243**, despite apoptotic cell bodies having formed, there were still some unaffected cell residues that survived and proliferated. Cell detachment, cell death, and apoptotic bodies indicate apoptotic cell death. Our data altogether suggest that at 100 μM, concentrations of compounds **58**, **ind-199**, **43,** and **243** induce apoptotic cell death.

Inhibition of antiapoptotic protein BCL-2 that is overexpressed in cancer cells is one of the most studied approaches in cancer research. Currently, venetoclax is the only approved drug by the FDA for the treatment of chronic lymphocytic leukemia (CLL), and it is a selective BCL-2 protein inhibitor. Although it has a very high affinity for BCL-2 as shown in various studies performed on different cancer cell lines (Yang et al., 2012) (www.cancerrxgene.org), resistance to this drug has already been observed (Birkinshaw et al., 2019). Hence, it is necessary to suggest new compounds and scaffolds as BCL-2 inhibitors that could be more efficient against mutations on the target structure and have no side effects. As such, in this study, we performed combined ligand- and structure-based approaches as well as text mining to propose new inhibitors against BCL-2 target protein. We selected seven hit compounds of which two are "*indol*"-based molecules. These compounds were considered in cancer cell line assays, and their IC_{50} values are calculated (**Table 1**). Based on the experimental findings, compound **58,** which has an IC_{50} value of 17 μM, was found to promote apoptosis, and it was the most effective of the seven compounds. Some of the selected hit molecules (compounds **258** and **292**) for *in vitro* analysis could only be partially solubilized in DMSO; therefore, solubility issue may restrict the activity of the compounds.

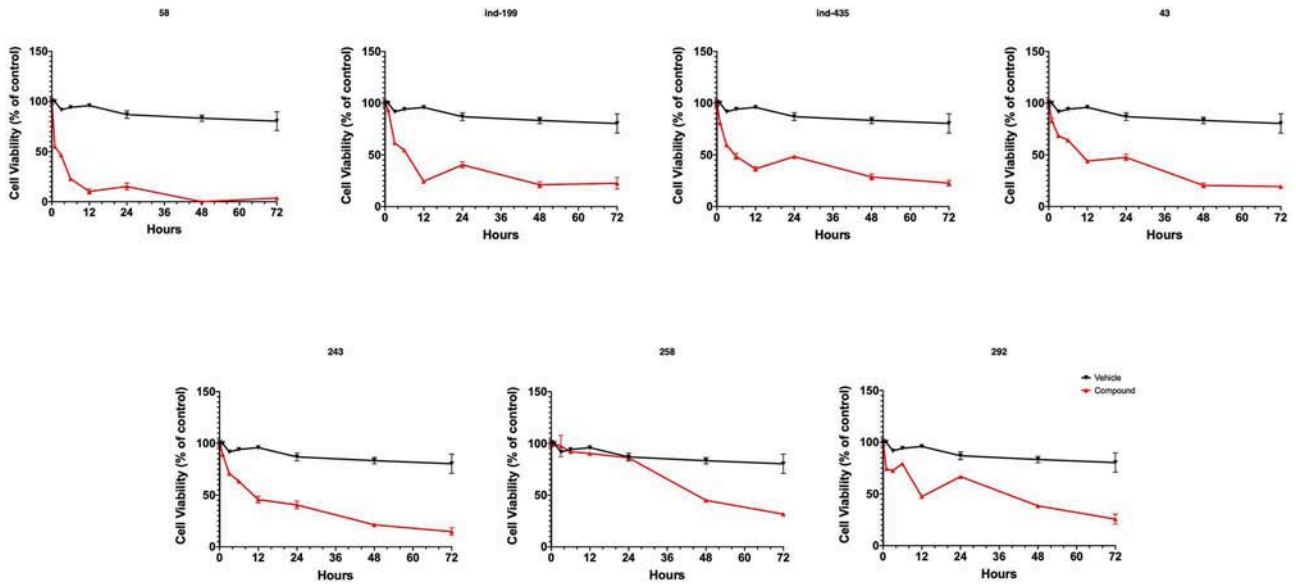


FIGURE 9 | MTT cell proliferation assay. Molecules having inhibitory activity were shown. Tested concentration for each drug was 100 μM. Statistical significance in graphs was determined by comparing each treatment group with DMSO control using ANOVA testing, and significance is considered as $p < 0.001$. Error bars show standard deviation.

When we have considered and compared the structures of selected hit molecules, specifically compounds that inhibited the proliferation of cancer cells such as compounds **58**, **ind-199**, **43,** and **243**, we have observed that all except compound **58** contains sulfur-containing groups such as sulfanyl or thiazole derivatives. All four molecules contain aromatic rings as well as amide groups in their structures, which are also groups observed in FDA-approved drug venetoclax and under development compound S55746. It must be noted that together with identified two indol-based compounds, **58** also involves an indole ring in its structure; thus, it validates the importance of indoles or indole derivatives in the scaffolds of potent BCL2 inhibitors. It must also be noted that the proposed hit molecules have lower molecular weights (MW) (between 490 and 529 g/mol) than the known BCL-2 inhibitors venetoclax (MW, 868 g/mol) and S55746 (MW, 710 g/mol) which indicate that the suggested compounds could be used as lead compounds for further optimization studies by small modifications. See .sdf file for structures of hit compounds along with their properties such as TAV, docking score, toxicity values, etc. in Data Sheet 1 in the **Supplementary Material**.

Overall, our results surprisingly show that docking-initiated screening has a better success rate compared to MD-initiated screening. However, this surprising result may be due to unexpected partial solubilities of some of the tested compounds that showed limited activity on cells.

CONCLUSIONS

In this work, a molecular library (Specs-SC) composed of 212,520 molecules was first filtered for their therapeutic effect against cancer, and then obtained molecules again filtered to remove toxic compounds using MC/MD from Clarivate Analytics. Identified 342 non-toxic and potent compounds using MC/MD were then screened based on target-driven approaches using available BCL-2 structures. In order to compare the both structural and energetic results, known BCL-2 inhibitors were used as positive control molecules and same computational protocols were applied for these compounds. Identified hit molecules from both docking and short (i.e., 1-ns) MM/GBSA calculations that have similar/better binding energies were compared to known inhibitors, then subjected to longer (i.e., 100-ns) MD simulations. In the virtual screening, two different strategies were considered and compared in the selection of hit compounds: (i) Compounds were ranked by their docking scores and long (100-ns) MD simulations were performed for the selected compounds and average MM/GBSA energies were calculated; (ii) Short (1-ns) MD simulations were performed for top-docking poses of all 342 compounds and average MM/GBSA scores were considered for the selection of molecules in long (100-ns) MD simulations of small molecules database. At the end, seven molecules were suggested as new scaffolds for inhibition of BCL-2. Compounds **58** (AJ-292/12931005), **ind-199** (AG-205/12549135), **43** (AO-081/41887762), and **243** (AN-698/40780701) with 100 μM concentration significantly limited the cell proliferation capacity of cancer cells. Four biologically active molecules showed their efficacy starting from the first hour of treatment by decreasing the number of proliferating cells. At the first day of treatment, only 60–70% of cells survived, while the number of proliferating cells decreased to <40% at the end of third day. TR-FRET analysis revealed that hit compounds **58**, **ind-199**, **243,** and **292** showed the maximum inhibitory effect ranging between 60 and 100% in 10 μM concentration. Thus, most of the active compounds found in the cell line tests

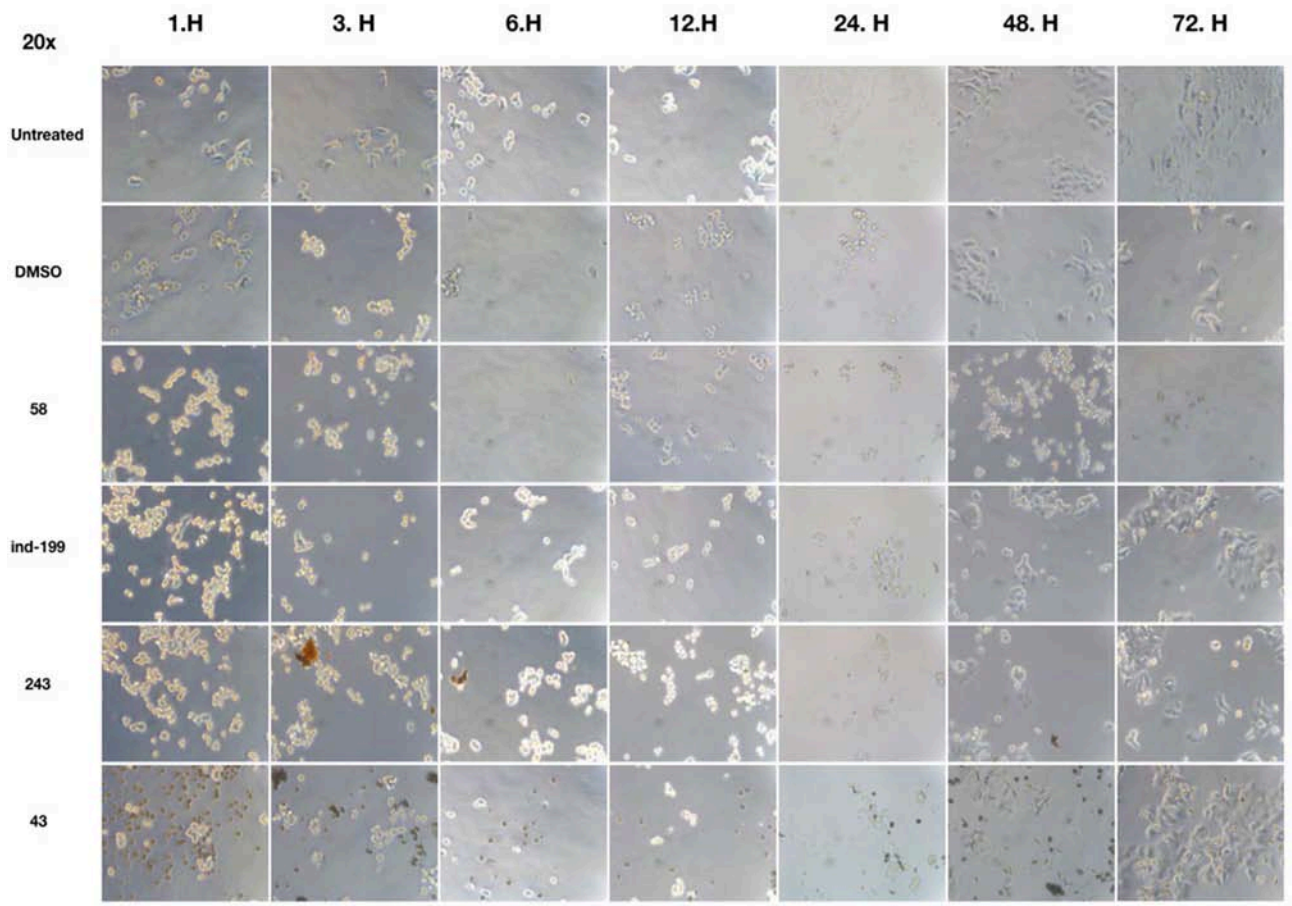


FIGURE 10 | Microscopic evaluation of HCT-116 cells for compounds **58** (AJ-292/12931005), **ind-199** (AG-205/12549135), **43** (AO-081/41887762), and **243** (AN-698/40780701). Cells were photographed and observed under microscope for 3 days. Vehicle group showed neat proliferation of cells as untreated group did, whereas molecule-treated groups reduced proliferation and showed apoptotic cell structures. Compound **58** (AJ-292/12931005) showed clear apoptotic activity from the first day, whereas other molecules showed limited activity. At 48 and 72 h of treatment in other groups, resistant cells are visible together with the unresolved molecule precipitates. However, there are no resistant cells and all cells seemed to be affected by the molecule **58** on treated group for all 3 days.

were also found potent in enzymatic assays. Results showed that compounds identified via integrated text-mining and docking initiated MM/GBSA-scores based approach has higher success rate. Thus, the results of this study may open new avenues for the designing of new BCL-2 inhibitor scaffolds.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

BD and SD constructed the methodology and wrote the manuscript. SD also guided the studies, revised the paper, and managed laboratory resources and research funding. GT and BD performed molecular modeling studies. MO, SC, and TA conducted the experimental studies and wrote the manuscript.

FUNDING

BD has been supported by The Scientific and Technological Research Council of Turkey (TUBITAK) under the program of TUBITAK-2218. This work has been supported by Bahcesehir University, Project Code: BAP.2019-02.10. The numerical calculations reported in this paper were partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

ACKNOWLEDGMENTS

We would like to thank Ismail Erol, Umit Yilmaz, and Canberk Ozbaykus for a helpful discussion.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2020.00167/full#supplementary-material

REFERENCES

Banks, J. L., Beard, H. S., Cao, Y., Cho, A. E., Damm, W., Farid, R., et al. (2005). Integrated modeling program, applied chemical theory (IMPACT). *J. Comput. Chem.* 26, 1752–1780. doi: 10.1002/jcc.20292

Bas, D. C., Rogers, D. M., and Jensen, J. H. (2008). Very fast prediction and rationalization of pKa values for protein–ligand complexes. *Proteins* 73, 765–783. doi: 10.1002/prot.22102

Bashford, D., and Case, D. A. (2000). Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* 51, 129–152. doi: 10.1146/annurev.physchem.51.1.129

Billard, C. (2013). BH3 mimetics: status of the field and new developments. *Mol. Cancer Ther.* 12, 1691–1700. doi: 10.1158/1535-7163.MCT-13-0058

Birkinshaw, R. W., Gong, J. N., Luo, C. S., Lio, D., White, C. A., Anderson, M. A., et al. (2019). Structures of BCL-2 in complex with venetoclax reveal the molecular basis of resistance mutations. *Nat. Commun.* 10:2385. doi: 10.1038/s41467-019-10363-1

Bissantz, C., Folkers, G., and Rognan, D. (2000). Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* 43, 4759–4767. doi: 10.1021/jm001044l

Bowers, K. J., Chow, D. E., Xu, H., Dror, R. O., Eastwood, M. P., Gregersen, B. A., et al. (2006). "Scalable algorithms for molecular dynamics simulations on commodity clusters," *Paper Presented at the SC'06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing* (Tampa, FL). doi: 10.1145/1188455.1188544

Bruncko, M., Oost, T. K., Belli, B. A., Ding, H., Joseph, M. K., Kunzer, A., et al. (2007). Studies leading to potent, dual inhibitors of Bcl-2 and Bcl-xL. *J. Med. Chem.* 50, 641–662. doi: 10.1021/jm061152t

Bursulaya, B. D., Totrov, M., Abagyan, R., and Brooks, C. L. (2003). Comparative study of several algorithms for flexible ligand docking. *J. Comput. Aided Mol. Des.* 17, 755–763. doi: 10.1023/B:JCAM.0000017496.76572.6f

Casara, P., Davidson, J., Claperon, A., Le Toumelin-Braizat, G., Vogler, M., Bruno, A., et al. (2018). S55746 is a novel orally active BCL-2 selective and potent inhibitor that impairs hematological tumor growth. *Oncotarget* 9, 20075–20088. doi: 10.18632/oncotarget.24744

Chen, H., Lyne, P. D., Giordanetto, F., Lovell, T., and Li, J. (2006). On evaluating molecular-docking methods for pose prediction and enrichment factors. *J. Chem. Inf. Model* 46, 401–415. doi: 10.1021/ci0503255

Chiba, S., Ishida, T., Ikeda, K., Mochizuki, M., Teramoto, R., Taguchi, Y. H., et al. (2017). An iterative compound screening contest method for identifying target protein inhibitors using the tyrosine-protein kinase yes. *Sci. Rep.* 7:12038. doi: 10.1038/s41598-017-10275-4

Chung, C. (2018). Restoring the switch for cancer cell death: targeting the apoptosis signaling pathway. *Am. J. Health Syst. Pharm.* 75, 945–952. doi: 10.2146/ajhp170607

Cory, S., Huang, D. C., and Adams, J. M. (2003). The Bcl-2 family: roles in cell survival and oncogenesis. *Oncogene* 22, 8590–8607. doi: 10.1038/sj.onc.1207102

Cross, J. B., Thompson, D. C., Rai, B. K., Baber, J. C., Fan, K. Y., Hu, Y., et al. (2009). Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* 49, 1455–1474. doi: 10.1021/ci900056c

Dewson, G., and Kluck, R. (2010). Bcl-2 family-regulated apoptosis in health and disease. *Cell Health Cytoskeleton* 2, 9–22. doi: 10.2147/CHC.S6228

Durdagi, S., Aksoydan, B., Erol, I., Kantarcioglu, I., Ergun, Y., Bulut, G., et al. (2018a). Integration of multi-scale molecular modeling approaches with experiments for the *in silico* guided design and discovery of novel herg-neutral antihypertensive oxazalone and imidazolone derivatives and analysis of their potential restrictive effects on cell proliferation. *Eur. J. Med. Chem.* 145, 273–290. doi: 10.1016/j.ejmech.2017.12.021

Durdagi, S., Dogan, B., Erol, I., Kayik, G., and Aksoydan, B. (2019). Current status of multiscale simulations on GPCRs. *Curr. Opin. Struct. Biol.* 55, 93–103. doi: 10.1016/j.sbi.2019.02.013

Durdagi, S., Salmas, R. E., Stein, M., Yurtsever, M., and Seeman, P. (2016). Binding interactions of dopamine and apomorphine in D2High and D2Low states of human dopamine D2 receptor using computational and experimental techniques. *ACS Chem. Neurosci.* 7, 185–195. doi: 10.1021/acschemneuro.5b00271

Durdagi, S., Tahirul Qamar, M., Salmas, R. E., Tariq, Q., Anwar, F., and Ashfaq, U. A. (2018b). Investigating the molecular mechanism of staphylococcal DNA gyrase inhibitors: a combined ligand-based and structure-based resources pipeline. *J. Mol. Graph. Modell.* 85, 122–129. doi: 10.1016/j.jmgm.2018.07.010

Eimon, P. M., and Ashkenazi, A. (2010). The zebrafish as a model organism for the study of apoptosis. *Apoptosis* 15, 331–349. doi: 10.1007/s10495-009-0432-9

Erol, I., Cosut, B., and Durdagi, S. (2019). Toward understanding the impact of dimerization interfaces in angiotensin II type 1 receptor. *J. Chem. Inf. Model.* 59, 4314–4327. doi: 10.1021/acs.jcim.9b00294

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *J. Chem. Phys.* 103, 8577–8593. doi: 10.1063/1.470117

Fesik, S. W. (2005). Promoting apoptosis as a strategy for cancer drug discovery. *Nat. Rev. Cancer* 5, 876–885. doi: 10.1038/nrc1736

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* 47, 1739–1749. doi: 10.1021/jm0306430

Fu, Y., Sun, Y.-N., Yi, K.-H., Li, M.-Q., Cao, H.-F., Li, J.-Z., et al. (2018). Combination of virtual screening protocol by *in silico* toward the discovery of novel 4-hydroxyphenylpyruvate dioxygenase inhibitors. *Front. Chem.* 6:14. doi: 10.3389/fchem.2018.00014

Gandhi, L., Camidge, D. R., De Oliveira, M. R., Bonomi, P., Gandara, D., Khaira, D., et al. (2011). Phase I study of navitoclax (ABT-263), a novel Bcl-2 family inhibitor, in patients with small-cell lung cancer and other solid tumors. *J. Clin. Oncol.* 29, 909–916. doi: 10.1200/JCO.2010.31.6208

Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* 47, 1750–1759. doi: 10.1021/jm030644s

Halim, S. A., Khan, S., Khan, A., Wadood, A., Mabood, F., Hussain, J., et al. (2017). Targeting dengue virus NS-3 helicase by ligand based pharmacophore modeling and structure based virtual screening. *Front. Chem.* 5:88. doi: 10.3389/fchem.2017.00088

Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* 100, 57–70. doi: 10.1016/S0092-8674(00)81683-9

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Hoover, W. G. (1985). Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A* 31:1695. doi: 10.1103/PhysRevA.31.1695

Igney, F. H., and Krammer, P. H. (2002). Death and anti-death: tumour resistance to apoptosis. *Nat. Rev. Cancer* 2, 277–288. doi: 10.1038/nrc776

Is, Y. S., Durdagi, S., Aksoydan, B., and Yurtsever, M. (2018). Proposing novel MAO-B hit inhibitors using multidimensional molecular modeling approaches and application of binary QSAR models for prediction of their therapeutic activity, pharmacokinetic and toxicity properties. *ACS Chem. Neurosci.* 9, 1768–1782. doi: 10.1021/acschemneuro.8b00095

Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E., et al. (2004). A hierarchical approach to all-atom protein loop prediction. *Proteins* 55, 351–367. doi: 10.1002/prot.10613

Kanan, T., Kanan, D., Erol, I., Yazdi, S., Stein, M., and Durdagi, S. (2019). Targeting the NF-κB/IκBα complex via fragment-based E-Pharmacophore virtual screening and binary QSAR models. *J. Mol. Graph. Modell.* 86, 264–277. doi: 10.1016/j.jmgm.2018.09.014

Kollek, M., Müller, A., Egle, A., and Erlacher, M. (2016). Bcl-2 proteins in development, health, and disease of the hematopoietic system. *FEBS J.* 283, 2779–2810. doi: 10.1111/febs.13683

Kumar Yadav, D., Dhawan, S., Chauhan, A., Qidwai, T., Sharma, P., Singh Bhakuni, R., et al. (2014a). QSAR and docking based semi-synthesis and *in vivo* evaluation of artemisinin derivatives for antimalarial activity. *Curr. Drug Targets* 15, 753–761. doi: 10.2174/1389450115666140630102711

Kumar Yadav, D., Kalani, K., Singh, A. K., Khan, F., Srivastava, S. K., and Pant, A. B. (2014b). Design, synthesis and *in vitro* evaluation of 18β-glycyrrhetinic acid derivatives for anticancer activity against human breast cancer cell line MCF-7. *Curr. Med. Chem.* 21, 1160–1170. doi: 10.2174/09298673113206660330

Kumar Yadav, D., Kalani, K., Khan, F., and Kumar Srivastava, S. (2013). QSAR and docking based semi-synthesis and *in vitro* evaluation of 18 β-glycyrrhetinic

acid derivatives against human lung cancer cell line A-549. *Med. Chem.* 9, 1073–1084. doi: 10.2174/1573406411309080009

Lessene, G., Czabotar, P. E., and Colman, P. M. (2008). BCL-2 family antagonists for cancer therapy. *Nat. Rev. Drug Discov.* 7, 989–1000. doi: 10.1038/nrd2658

Li, Y., Han, L., Liu, Z., and Wang, R. (2014). Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results. *J. Chem. Inf. Model.* 54, 1717–1736. doi: 10.1021/ci500081m

Lionta, E., Spyrou, G., K., Vassilatis, D., and Cournia, Z. (2014). Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr. Top. Med. Chem.* 14, 1923–1938. doi: 10.2174/1568026614666140929124445

Martyna, G. J., Tobias, D. J., and Klein, M. L. (1994). Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* 101, 4177–4189. doi: 10.1063/1.467468

Mirza, S. B., Lee, R. C. H., Chu, J. J. H., Salmas, R. E., Mavromoustakos, T., and Durdagi, S. (2018). Discovery of selective dengue virus inhibitors using combination of molecular fingerprint-based virtual screening protocols, structure-based pharmacophore model development, molecular dynamics simulations and *in vitro* studies. *J. Mol. Graph. Modell.* 79, 88–102. doi: 10.1016/j.jmgm.2017.10.010

Mohamad Rosdi, M. N., Mohd Arif, S., Abu Bakar, M. H., Razali, S. A., Mohamed Zulkifli, R., and Ya'akob, H. (2018). Molecular docking studies of bioactive compounds from Annona muricata Linn as potential inhibitors for Bcl-2, Bcl-w and Mcl-1 antiapoptotic proteins. *Apoptosis* 23, 27–40. doi: 10.1007/s10495-017-1434-7

Mollica, A., Zengin, G., Durdagi, S., Ekhteiari Salmas, R., Macedonio, G., Stefanucci, A., et al. (2019). Combinatorial peptide library screening for discovery of diverse α-glucosidase inhibitors using molecular dynamics simulations and binary QSAR models. *J. Biomol. Struct. Dyn.* 37, 726–740. doi: 10.1080/07391102.2018.1439403

Nosé, S. (1984). A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* 81, 511–519. doi: 10.1063/1.447334

Oltersdorf, T., Elmore, S. W., Shoemaker, A. R., Armstrong, R. C., Augeri, D. J., Belli, B. A., et al. (2005). An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature* 435, 677–681. doi: 10.1038/nature03579

Petros, A. M., Medek, A., Nettesheim, D. G., Kim, D. H., Yoon, H. S., Swift, K., et al. (2001). Solution structure of the antiapoptotic protein bcl-2. *Proc. Natl Acad. Sci. U.S.A.* 98, 3012–3017. doi: 10.1073/pnas.041619798

Rastelli, G., Degliesposti, G., Del Rio, A., and Sgobba, M. (2009). Binding estimation after refinement, a new automated procedure for the refinement and rescoring of docked ligands in virtual screening. *Chem. Biol. Drug Des.* 73:283–86. doi: 10.1111/j.1747-0285.2009.00780.x

Rastelli, G., and Pinzi, L. (2019). Refinement and rescoring of virtual screening results. *Front. Chem.* 7:498. doi: 10.3389/fchem.2019.00498

Reed, J. C. (2002). Apoptosis-based therapies. *Nat. Rev. Drug Discov.* 1, 111–121. doi: 10.1038/nrd726

Rodrigues, M. J., Slusarczyk, S., Pecio, Ł., Matkowski, A., Salmas, R. E., Durdagi, S., et al. (2018). *In vitro* and *in silico* approaches to appraise *Polygonum maritimum* L. as a source of innovative products with anti-ageing potential. *Indust. Crops Products* 111, 391–399. doi: 10.1016/j.indcrop.2017.10.046

Salmas, R. E., Seeman, P., Aksoydan, B., Erol, I., Kantarcioglu, I., Stein, M., et al. (2017). Analysis of the glutamate agonist LY404,039 binding to nonstatic dopamine receptor D2 dimer structures and consensus docking. *ACS Chem. Neurosci.* 8, 1404–1415. doi: 10.1021/acschemneuro.7b00070

Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., and Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* 27, 221–234. doi: 10.1007/s10822-013-9644-8

Schrödinger Release 2015-2 (2015) *LigPrep*. New York, NY: Schrödinger, LLC.

Shan, Y., Kim, E. T., Eastwood, M. P., Dror, R. O., Seeliger, M. A., and Shaw, D. E. (2011). How does a drug molecule find its target binding site? *JACS* 133, 9181–9183. doi: 10.1021/ja202726y

Shelley, J. C., Cholleti, A., Frye, L. L., Greenwood, J. R., Timlin, M. R., and Uchimaya, M. (2007). Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. *J. Comput. Aided Mol. Des* 21, 681–691. doi: 10.1007/s10822-007-9133-z

Sherman, W., Beard, H. S., and Farid, R. (2006a). Use of an induced fit receptor structure in virtual screening. *Chem. Biol. Drug Des.* 67, 83–84. doi: 10.1111/j.1747-0285.2005.00327.x

Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A., and Farid, R. (2006b). Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* 49, 534–553. doi: 10.1021/jm050540c

Soderquist, R. S., and Eastman, A. (2016). BCL2 inhibitors as anticancer drugs: a plethora of misleading BH3 mimetics. *Mol. Cancer Ther.* 15, 2011–2017. doi: 10.1158/1535-7163.MCT-16-0031

Souers, A. J., Leverson, J. D., Boghaert, E. R., Ackler, S. L., Catron, N. D., Chen, J., et al. (2013). ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. *Nat. Med* 19, 202–208. doi: 10.1038/nm.3048

Tse, C., Shoemaker, A. R., Adickes, J., Anderson, M. G., Chen, J., Jin, S., et al. (2008). ABT-263: a potent and orally bioavailable Bcl-2 family inhibitor. *Cancer Res.* 68, 3421–3428. doi: 10.1158/0008-5472.CAN-07-5836

Verma, S., Singh, A., and Mishra, A. (2015). Complex disruption effect of natural polyphenols on Bcl-2-Bax: molecular dynamics simulation and essential dynamics study. *J. Biomol. Struct. Dyn.* 33, 1094–1106. doi: 10.1080/07391102.2014.931823

Warren, G. L., Andrews, C. W., Capelli, A.-M., Clarke, B., LaLonde, J., Lambert, M. H., et al. (2006). A critical assessment of docking programs and scoring functions. *J. Med. Chem.* 49, 5912–5931. doi: 10.1021/jm050362n

Wu, H., Medeiros, L. J., and Young, K. H. (2018). Apoptosis signaling and BCL-2 pathways provide opportunities for novel targeted therapeutic strategies in hematologic malignances. *Blood Rev.* 32, 8–28. doi: 10.1016/j.blre.2017.08.004

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., et al. (2012). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* 41, D955–D961. doi: 10.1093/nar/gks1111

Yoshino, R., Yasuo, N., Hagiwara, Y., Ishida, T., Inaoka, D. K., Amano, Y., et al. (2017). *In silico*, *in vitro*, X-ray crystallography, and integrated strategies for discovering spermidine synthase inhibitors for Chagas disease. *Sci. Rep.* 7:6666. doi: 10.1038/s41598-017-06411-9

Yoshino, R., Yasuo, N., Inaoka, D. K., Hagiwara, Y., Ohno, K., Orita, M., et al. (2015). Pharmacophore modeling for anti-chagas drug design using the fragment molecular orbital method. *PLoS ONE* 10:e0125829. doi: 10.1371/journal.pone.0125829

Youle, R. J., and Strasser, A. (2008). The BCL-2 protein family: opposing activities that mediate cell death. *Nat. Rev. Mol. Cell Biol.* 9, 47–59. doi: 10.1038/nrm2308

Zaka, M., Abbasi, B. H., and Durdagi, S. (2018). Proposing novel TNFα direct inhibitor Scaffolds using fragment-docking based e-pharmacophore modeling and binary QSAR-based virtual screening protocols pipeline. *J. Mol. Graph. Modell.* 85, 111–121. doi: 10.1016/j.jmgm.2018.07.007

Zaka, M., Abbasi, B. H., and Durdagi, S. (2019). Novel tumor necrosis factor-α (TNF-α) inhibitors from small molecule library screening for their therapeutic activity profiles against rheumatoid arthritis using target-driven approaches and binary QSAR models. *J. Biomol. Struct. Dyn.* 37, 2464–2476. doi: 10.1080/07391102.2018.1491423

Zhang, H., Nimmer, P., Tahir, S., Chen, J., Fryer, R., Hahn, K., et al. (2007). Bcl-2 family proteins are essential for platelet survival. *Cell Death Differ.* 14, 943–951. doi: 10.1038/sj.cdd.4402081

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Tutumlu, Dogan, Avsar, Orhan, Calis and Durdagi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 21 April 2020
doi: 10.3389/fchem.2020.00246

High-Throughput Docking Using Quantum Mechanical Scoring

Claudio N. Cavasotto[1,2,3]* *and M. Gabriela Aucar*[1]

[1] *Computational Drug Design and Biomedical Informatics Laboratory, Translational Medicine Research Institute (IIMT), CONICET-Universidad Austral, Pilar, Argentina,* [2] *Facultad de Ciencias Biomédicas and Facultad de Ingeniería, Universidad Austral, Pilar, Argentina,* [3] *Austral Institute for Applied Artificial Intelligence, Universidad Austral, Pilar, Argentina*

Today high-throughput docking is one of the most commonly used computational tools in drug lead discovery. While there has been an impressive methodological improvement in docking accuracy, docking scoring still remains an open challenge. Most docking programs are rooted in classical molecular mechanics. However, to better characterize protein-ligand interactions, the use of a more accurate quantum mechanical (QM) description would be necessary. In this work, we introduce a QM-based docking scoring function for high-throughput docking and evaluate it on 10 protein systems belonging to diverse protein families, and with different binding site characteristics. Outstanding results were obtained, with our QM scoring function displaying much higher enrichment (screening power) than a traditional docking method. It is acknowledged that developments in quantum mechanics theory, algorithms and computer hardware throughout the upcoming years will allow semi-empirical (or low-cost) quantum mechanical methods to slowly replace force-field calculations. It is thus urgently needed to develop and validate novel quantum mechanical-based scoring functions for high-throughput docking toward more accurate methods for the identification and optimization of modulators of pharmaceutically relevant targets.

Keywords: high-throughput docking, structure-based drug design, molecular docking, quantum mechanics, semi-empirical methods

OPEN ACCESS

Edited by:
Simone Brogi,
University of Pisa, Italy

Reviewed by:
Tingjun Hou,
Zhejiang University, China
Stefano Moro,
University of Padova, Italy
Vladimir Borisovich Sulimov,
Lomonosov Moscow State University, Russia

***Correspondence:**
Claudio N. Cavasotto
ccavasotto@austral.edu.ar;
cnc@cavasotto-lab.net

Specialty section:
This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

Received: *02 February 2020*
Accepted: *16 March 2020*
Published: *21 April 2020*

Citation:
Cavasotto CN and Aucar MG (2020) High-Throughput Docking Using Quantum Mechanical Scoring. Front. Chem. 8:246. doi: 10.3389/fchem.2020.00246

INTRODUCTION

The cost to bring a new drug to the market could be as high as 2.6 billion US dollars, and can take up to 15 years (DiMasi et al., 2016). For many years, both the identification and optimization of novel drug lead compounds were accomplished within the drug discovery process by the experimental high-throughput screening of large chemical libraries. In spite of multiple efforts to improve its performance, drug discovery remains a costly and time consuming technique (Phatak et al., 2009). However, for the last 25 years, theoretical developments, better computational algorithms, faster computing resources, and improved visualization tools enabled the routine use of computational methods to model and visualize protein-ligand (PL) interactions, calculate binding free energy to different degrees of accuracy, and *in silico* screen chemical libraries using ligand-based and structure-based approaches. Today, computational chemistry is firmly established as a valuable tool in any drug lead discovery endeavor, aimed at saving time, effort, resources, and reducing costs (Cavasotto and Orry, 2007; Jorgensen, 2009, 2012; Spyrakis and Cavasotto, 2015; Pagadala et al., 2017).

During the last three decades, molecular docking has been one of the most commonly used computational methods in drug lead discovery (for review, cf., Kitchen et al., 2004; Rognan, 2011; Ciancetta and Moro, 2015; Sotriffer, 2015; Spyrakis and Cavasotto, 2015; Sulimov et al., 2019b). The aim of protein-small-molecule docking is the characterization of the optimal binding modes (poses) of a molecule within the binding site, and an estimation of its binding free energy. In high-throughput docking (HTD), where the protein is usually considered rigid or with very few degrees of freedom, and thousands to millions of molecules from a chemical library are screened, the goal is to generate a sub-library enriched with potential ligands, which will be prioritized for further experimental evaluation. In HTD, two different stages can be distinguished: the assessment of the best binding mode(s) of each molecule of the library ("docking stage"), and, on each *in silico* generated protein-small-molecule complex, the calculation of a score reflective of the likelihood that the molecule will actually bind to the target ("scoring stage") (Cavasotto and Orry, 2007; Guedes et al., 2018). In the docking stage, the docking energy (DE) is used to select, for each molecule, the lowest-energy pose(s) from a large amount of conformations generated, while the docking score (DS) is generally calculated as a fast approximation to the binding free energy ($\Delta G_{binding}$), and depends on several factors, such as the energy representation of the system, the model used to represent the aqueous environment and the consideration of explicit water molecules within the active site (Cozzini et al., 2006; Amadasi et al., 2008), and the degree of consideration of receptor flexibility (Cavasotto and Singh, 2008; Spyrakis et al., 2011; Spyrakis and Cavasotto, 2015). Thus, DE discriminates among poses of the same molecule, while the DS characterizes each molecule of the docked chemical library and is used to rank them according to the likelihood of binding. Many docking programs, however, use a single function as DE and DS.

It should be stressed that one of the main advantages of docking is that *in silico* generated poses usually serve as the starting point for *in silico* ligand optimization, using for example molecular dynamics-based calculation of binding free energies, such as Molecular Mechanics-Poisson Boltzmann Surface Area (MM-PBSA) and MM-Generalized Born Surface Area (MM-GBSA) methods (Kerrigan, 2013; Reddy et al., 2014; Genheden and Ryde, 2015; Sun et al., 2018; Wang et al., 2019).

While docking accuracy depends on the program, it is acknowledged that most of them are usually successful in identifying the correct pose (RMSD $<$ 2 Å) with respect to the native structure (Warren et al., 2006; Wang et al., 2016). Moreover, an extensive recent benchmark of the Comparative Assessment of Scoring Functions (CASF) (Su et al., 2019) highlighted that docking programs display a better performance in terms of docking accuracy than in any of these three scoring-related metrics: correlation with experimental binding data (scoring power), ranking of ligands by their binding affinity data provided their correct poses are known (ranking power), and identification of actual ligands from a sub-library of top-ranking small-molecules (screening power). This was in agreement with other works (Cavasotto and Abagyan, 2004; Slater and Kontoyianni, 2019).

Most docking developments have been mainly rooted in molecular mechanics (MM) force-fields (FF). However, to better characterize protein-ligand interactions, at least in some cases, the use of a quantum mechanical (QM) description would be necessary (Cavasotto et al., 2019). The QM formulation is theoretically exact, as in principle, it accounts for all contributions to the energy (including terms or effects usually missing in FFs, such as electronic polarization, charge transfer, halogen bonding, and covalent-bond formation). Moreover, the QM framework is general across the chemical space so that all elements and interactions can be considered on equal footing, thus avoiding MM parameterizations.

Following the pioneering work of Raha and Merz (2004, 2005) where a QM-based score was used to discriminate ligand from decoy poses, there have been recently some applications of QM methods in docking, mainly aiming for accurate ligand binding mode assessment (for a survey of recent related works cf., Mucs and Bryce, 2013; Cavasotto et al., 2018; Aucar and Cavasotto, 2020). In a significant step forward, Pecina et al. obtained impressive results on the discrimination of native from decoy docking poses on four challenging systems (Pecina et al., 2016) using a docking energy function (Lepšík et al., 2013) based on the semi-empirical QM PM6 Hamiltonian (Stewart, 2007) supplemented with the D3H4X correction for dispersion, hydrogen- and halogen-bonding interactions (Rezáč and Hobza, 2012). In a follow-up contribution (Pecina et al., 2017), an even superior performance was achieved for accurate pose assessment using a self-consistent-charge density-functional tight-binding method (SCC-DFTB) formulation coupled with D3H4 corrections for dispersion and hydrogen-bond interactions, though at a higher computational cost. This docking energy score function was further used to obtain a reliable ranking on 10 inhibitors binding to carbonic anhydrase II (CAII) (Pecina et al., 2018).

However, the development of QM-based docking scoring functions aiming at the ranking of molecules within HTD (screening power) has progressed at a significantly slower pace. Only very recently, a QM-based approach was presented displaying a very good performance on discriminating ligands and decoys on a single system (heat shock protein 90, HSP90) (Eyrilmez et al., 2019). In fact, the development of fast yet accurate docking scoring functions still constitutes an area of active research (Cavasotto, 2012; Guedes et al., 2018). Moreover, the blind challenges ran by the Drug Design Data Resource (D3R) for ligand-pose and affinity prediction in 2015 (Gathiaka et al., 2016), 2016 (Gaieb et al., 2018), and 2018 (Gaieb et al., 2019), have shown the importance of method development and benchmarking in pose prediction and binding affinity ranking of ligands.

In this work, we introduce a QM-based docking scoring function and evaluate it in terms of ligand enrichment on 10 protein systems belonging to diverse protein families in terms of different binding site characteristics, the presence of co-factors and water molecules, and the enrichment factors computed with a standard HTD method. Excellent results were obtained by displaying our QM-based scoring function a much higher enrichment (screening power) than a traditional docking

method. We stress that our goal is to present and to validate an initial straightforward approach, which could serve as a starting point for further developments and improvement. A wider and extensive benchmarking on more systems and a systematic comparison with most of the standard docking programs, and the assessment of the optimal combination of the different components of our approach (QM formalism and continuum solvent model, energy minimization strategies, use of single or multiple docking poses for scoring, and entropy contribution) are considerations of their importance. However, they exceed the purpose of our work and will be published in due course.

Assuming a continuous development in QM theory, algorithms and computer hardware, it is likely that semi-empirical methods [or low-cost Density Functional Theory (DFT) methods] will replace FF over the next 25 years (Grimme and Schreiner, 2018). Therefore, it is absolutely justified and there is an urgent need to start developing the next generation of QM-based scoring functions for HTD toward better methods for the identification of small-molecule modulators of pharmaceutically relevant targets.

TABLE 1 | Target proteins used in the evaluation of QM-based scoring functions.

Receptor name	Receptor code	PDB code	Co-factor[a]	Number of water molecules[b]	*EF*(1)[c]
Cyclin-dependent Kinase 2	CDK2	1FVV	–	–	8.0
Estrogen receptor α	ESR1	3ERT	–	–	16.5
Cyclooxygenase-1	COX1	2OYU	–	–	1.3
Neuraminidase	NRAM	1B9V	Ca^{2+}	–	0.
Heat shock protein 90 α	HSP90a	1UYG	–	4	0.
Hexokinase type IV	HXK4	3F9M	–	–	1.1
Coagulation factor VII	FA7	1W7X	–	1	20.2
Thymidine kinase	KITH	2B8T	–	–	35.1
Fatty acid binding protein adipocyte	FABP4	2NNQ	–	4	31.9
Phospholipase A2	PA2GA	1KVO	Ca^{2+}	–	2.0

[a] *Within 8 Å of the cyrstallographic ligand.*
[b] *Within 4 Å of the cyrstallographic ligand.*
[c] *Enrichment factor at 1% corresponding to docking with AutoDock Vina.*

MATERIALS AND METHODS

Protein Systems Preparation

The following targets were downloaded from the PDB (cf. **Table 1**): Cyclin-dependent Kinase 2 (CDK2, PDB 1FVV), Estrogen Receptor α (ESR1, PDB 3ERT), Cyclooxygenase-1 (COX1, PDB 2OYU), Neuraminidase (NRAM, PDB 1B9V), Heat Shock Protein 90 α (HSP90a, PDB 1UYG), Hexokinase Type IV (HXK4, PDB 3F9M), Coagulation Factor VII (FA7, PDB 1W7X), Thymidine kinase (KITH, PDB 2B8T), Fatty Acid Binding Protein Adipocyte (FABP4, PDB 2NNQ), and Phospholipase A2 (PA2GA, PDB 1KVO). All water molecules and co-factors were deleted, except in the following cases: NRAM and PA2GA, the Ca^{2+} atom within 8 Å of the bound ligand; HSP90a, water molecules 2059, 2121, 2123, and 2236; FA7, water molecule 2440; FABP4 water molecules 303, 623, 634, 665.

Each target was prepared using ICM software (MolSoft, San Diego, CA, 2019; Abagyan et al., 1994) in a similar fashion as in earlier works (Phatak et al., 2010). Succinctly, hydrogen atoms were added, followed by a local energy minimization of the complete system, and polar and water hydrogen positions were determined by optimizing the hydrogen bonding network within the torsional coordinates space. All Asp and Glu residues were assigned a −1 charge, and all Arg and Lys residues were assigned a +1 charge. Histidine tautomers were chosen according to their corresponding hydrogen bonding pattern. For docking with AutoDock Vina (Trott and Olson, 2010), the systems were pre-processed with AutoDock Tools (Morris et al., 2009).

Docking Library Preparation

For each target, the docking libraries were built by merging a set of ligands and a set of decoys, where the latter had similar physico-chemical properties to the ligands, but dissimilar 2-D topology. This has been shown to be necessary to ensure unbiased results when benchmarking docking programs (Huang et al., 2006; Gatica and Cavasotto, 2012). Ligands and decoys were extracted from the Directory of Useful Decoys (DUD, Huang et al., 2006), the NRLiSt binding data base for nuclear receptors (Lagarde et al., 2014), or the Directory of Useful Decoys- Enhanced (DUD-E, Mysinger et al., 2012), according to: CDK2, DUD (72, 2074) (number of ligands, number of decoys); ESR1, NRLiSt (133, 6555); COX1, DUD-E (210, 6955); NRAM, DUD-E (222, 6227); HSP90a (125, 4942); HXK4, DUD-E (127, 4802); FA7, DUD-E (185, 6300); KITH, DUD-E (132, 2866); FABP4, DUD-E (57, 2855); PA2GA (127, 5215). The protonation state and chirality of all molecules were conserved as in their original database.

High-Throughput Docking With AutoDock Vina

Molecular docking of the chemical libraries onto the associated targets using AutoDock Vina (Trott and Olson, 2010) was performed in a similar fashion as in our recent work (Palacio-Rodriguez et al., 2019).

Protein-Molecule Complex Generation, Structural Relaxation, and Unbound Protein and Ligand States Characterization

Protein-molecule complexes for QM-scoring were generated using the ICM docking module, keeping for each molecule its lowest DE conformation (docking RMSD values of native ligands are shown in **Table 2**). These protein-molecule complexes were also relaxed through cycles of local energy minimization in ICM according to the following procedure: (i) For each protein, the collected dihedral angles of amino-acids within 4 Å of any docked ligand of the corresponding chemical library were considered free; (ii) For each protein-molecule complex, five cycles of local energy minimization were performed restraining the heavy atoms with a harmonic potential with respect to their initial conformation; in each cycle the weight of this added potential

TABLE 2 | RMSD values of docked native ligands.

Receptor name	Receptor code	PDB ligand ID	RMSD (Å)	
			ICM	AD Vina
Cyclin-dependent Kinase 2	CDK2	107	0.74	2.68
Estrogen receptor α	ESR1	oht	1.48	4.69
Cyclooxygenase-1	COX1	ims	1.40	0.36
Neuraminidase	NRAM	ra2	0.54	0.91
Heat shock protein 90 α	HSP90a	pu2	0.52	0.26
Hexokinase type IV	HXK4	mrk	0.47	7.56
Coagulation factor VII	FA7	413	0.31	0.60
Thymidine kinase	KITH	thm	0.22	0.70
Fatty acid binding protein adipocyte	FABP4	t4b	0.37	0.79
Phospholipase A2	PA2GA	oap	0.54	1.37

was reduced in the following way: 50, 10, 5, 1, and 0 kcal/mol (no restraint). During this local energy minimization, the protein system was optimized in the torsional space (Abagyan et al., 1994), and the small-molecule in the Cartesian space.

To generate the unbound states, local energy minimization was performed on both protein and small-molecule in isolation from the crystallographic structure and the docked conformation, respectively.

System Cutout

For each target, a reduced-system was defined by first listing all the amino-acids within 8 Å of any docked molecule with ICM (only heavy atoms were considered in this threshold). Then, upon visual inspection, other amino-acids were eventually added to the list in order to avoid intra-helix or intra-β-sheet fragmentation, or loop fragments with just one amino-acid. A reduced-system was then built by deleting from the structure all amino-acids not included in the list, capping the N- and C-terminal of each fragment with hydrogens.

Entropy Calculation

Binding small-molecule conformational entropy was estimated as

$$\Delta S = -R \ln \Omega \tag{1}$$

where it is assumed that, upon binding, the molecule adopts a single conformation state (thus $S_{bound} = 0$), and Ω is the number of conformations in the free state, which was estimated in two different ways: i) by assigning each of the N free torsional bonds three rotational degrees of freedom (and thus $\Omega = 3^N$); ii) by performing a Monte-Carlo (MC) sampling with local energy minimization in the torsional space using ICM (Abagyan and Totrov, 1994; Abagyan et al., 1994), collecting all distinct conformations within the lowest 3 kcal/mol energy, and assuming all conformers are equally probable (a similar low-level sampling approach was used to explore the conformational flexibility of small-molecules, Forti et al., 2012). The MC approach was considered since rotamer count is known to over-estimate the number of low-energy conformations, and thus the entropy (Anisimov and Cavasotto, 2011).

Quantum Mechanical Calculations

All QM calculations were performed using the QM package MOPAC2016 (Stewart, 2016) and its linear-scaling module MOZYME (Stewart, 1996), using the semi-empirical PM7 Hamiltonian (Stewart, 2013). In agreement with other authors (Sulimov et al., 2017a), we selected PM7 since it accounts for dispersion interactions, and hydrogen and halogen bonding have been taken into consideration at the paramterization stage, while it also includes several corrections to the PM6 Hamiltonian. Moreover, PM7 exhibited a very good performance on energy calculations aimed at discriminating native ligand positions in crystallographic complexes (Sulimov et al., 2017b). The solvation energy contribution in aqueous environment was calculated using the Conductor Like Screening Model (COSMO, Klamt and Schüürmann, 1993) continuum solvent model, with default atomic radii and surface tension parameters. The solvent-accessible surface area was taken from the program output [cf. (Stewart, 2016) for details on how the surface is built]. Those molecules which did not complete the QM calculation were excluded when computing the enrichment.

Evaluation Metrics

The enrichment factor (*EF*) measures the enrichment of actual ligands in a docked hit-list given a specific percentage of the dataset (threshold). The *EF* is defined as the ratio between actual number ligands (hits) found at the top x% of the screened database ($Hits_{x\%}$) and the number of molecules at that threshold $N_{x\%}$, normalized by the ratio between the total number of actual ligands within the entire dataset ($Hits_{\text{total}}$) and the total number of molecules of the latter (N_{total}).

$$EF(x) = \frac{Hits_{x\%}}{N_{x\%}} / \frac{Hits_{total}}{N_{total}} \tag{2}$$

Thus, the EF represents the probability of finding an actual ligand within the x% of the screened database with respect to the probability of finding an actual ligand at random. Whenever a molecule is represented within a chemical library with different states according to its protonation or chirality, each state is assigned an individual score, and the lowest score is used in the hit-list, and thus to calculate the *EF*. Throughout this work we report *EF*(1) and *EF*(2), since they are more representative of early enrichment.

We also report receiver operating characteristics (ROC) curves for each of the studied systems, measuring the area under the curve (AUC).

THEORETICAL FRAMEWORK

The binding free energy ($\Delta G_{binding}$) corresponding to Protein-Ligand (PL) association is expressed within the end-point molecular mechanics-quantum mechanics surface area method

(MM-QMSA) (Anisimov and Cavasotto, 2011; Anisimov et al., 2011) as

$$\Delta G_{binding} = \Delta \left\langle G^{QM} \right\rangle - T\Delta S \quad (3)$$

where the difference in the first term is calculated between the bound (PL) and unbound (P, L) states, <…> represents the average over QM-minimized classical molecular dynamics (MD) trajectories, G^{QM} is the QM energy including a continuum solvation term in an aqueous environment, and the second term represents the entropy change of P and L upon binding. We prefer to note the first term as a free energy, since it also includes the change in solvation free energy.

Since $\Delta G_{binding}$ in Equation (3) is obviously too costly to be used to score and rank large chemical libraries of small-molecules in HTD, a reasonable QM docking scoring function (*QMDS*) can be defined as an approximation to Equation (3), namely

$$QMDS = \Delta G^{QM} - T\Delta S \quad (4)$$

where averages over MD trajectories have been replaced by single-point QM calculations on the docked PL structure, and the free unbound L and P structures. The L and P deformation penalty contributions due to changes in L and P conformations upon binding are expressed as

$$\Delta G^{QM}_{conf}(X) = G^{QM}_{o}(X) - G^{QM}(X)\ , \text{ with } X = \text{L, P} \quad (5)$$

where G_o(X) is the energy of the isolated X in the conformation of the docked PL complex, and G(X) is the energy of X in the free unbound state. Considering Equation (5), Equation (4) can be now be written out making the deformation contributions explicit as

$$QMDS = \Delta G^{QM}_{o} + \Delta G^{QM}_{conf}(P) + \Delta G^{QM}_{conf}(L) - T\Delta S \quad (6)$$

where the "o" subscript in the first term refers to calculations using the PL, P, and L conformations from the docked complex. It should be pointed out that Equation 6 is formally identical to another formulation (Eyrilmez et al., 2019).

Two types of QM docking scoring functions were defined according to the relaxation of the reference docked PL complexes: (i) $QMDS_1$, with no relaxation, that is, the QM calculations are performed directly on the docked PL complex, and (ii) $QMDS_2$, where docked PL complexes are relaxed through local energy minimization (see Methods). When the deformation contributions (second and third terms in Equation 6) were included, the suffix "d" is added ($QMDS_{1d}$ and $QMDS_{2d}$).

RESULTS AND DISCUSSION

Improved HTD Enrichment Using QM-Based Scoring

Ten target proteins were selected based on different characteristics such as protein family, binding site properties, presence of co-factors and water molecules (within or close to the binding site), and enrichment factor at 1% calculated after docking with AutoDock Vina (**Table 1**). Only crystallographic and/or conserved water molecules within 4 Å of the native ligand were included.

TABLE 3 | Comparison of the enrichment factors [*EF*(1)] for docking and scoring ($QMDS_1$) using a complete and reduced protein systems.

Receptor	Complete system	Reduced system
CDK2	20.1	24.2
ESR1	33.7	36.4

Throughout all this work, the *QMDS* was calculated in all its variants on PL complexes generated with ICM docking, since it is acknowledged to generate high quality protein-molecule poses (Bursulaya et al., 2003; Neves et al., 2012), as confirmed by the RMSD values of the docked native ligands in **Table 2**). Clearly, better enrichment is strongly coupled to scoring over correct docking poses. In this regard, the use of multiple docked conformations for each molecule, stemming from the same docking program or not, might clearly enhance the results of our QM-scoring scheme. However, we preferred to use a single pose from a single program, to keep our methodology straightforward, and to establish a clear baseline from which to start looking for improvement.

Since a target receptor protein is usually very large for QM calculations, to calculate the *QMDS* we used a reduced system by cutting out amino acids farther than ∼8 Å from any docked molecule (cf. the Methods section for full details on the cutout process), since a threshold of <6 Å has been reported to seriously deteriorate the results (Ehrlich et al., 2017); moreover, it should be highlighted that the smaller the threshold, the greater the impact of the continuous solvent surface replacing the cutout amino-acids. To further validate our approach, quantum mechanical docking scores *QMDS* considering the complete protein and its associated reduced system were calculated on CDK2 and ESR1 (**Table 3**). We observe that using a cutout system has no impact on the calculation. Thus, throughout this work, a reduced representation of the target protein will be used for all QM calculations.

In **Table 4**, we display the enrichment factors *EF*(1) for the 10 target systems comparing AutoDock Vina with four schemes of QM docking scoring (for HSP90a, enrichment values including and excluding the 19 macrocycle containing ligands are shown). The conformational entropy change upon ligand binding was estimated in two ways: (i) ΔS^{rot}, based on a term proportional to the number of N free rotatable bonds of the molecule ($\Omega_{conf} = 3^N$), and (ii) ΔS^{conf}, by estimating Ω_{conf} as the number of low-energy diverse conformations generated using Monte-Carlo sampling with local energy minimization (cf. Methods). We found that the use of S^{rot} deteriorates the *EF* (data not shown), so S^{conf} is used in all calculations. In $QMDS_2$ and $QMDS_{2d}$ the reference docked PL complexes were local energy minimized using MM (see Methods). Obviously, a QM minimization would have been desirable, but this would render any QM docking scoring function useless due to the computational times involved,

TABLE 4 | Enrichment factors calculated at 1% [*EF*(1)] and 2% [*EF*(2)] (in parenthesis) for AutoDock Vina and QM docking scoring.

Receptor	AD Vina	$QMDS_1$	$QMDS_{1d}$	$QMDS_2$	$QMDS_{2d}$
CDK2	8.0 (5.0)	24.2 (15.1)	26.2 (18.1)	26.3 (16.2)	26.3 (18.2)
ESR1	16.5 (11.0)	36.4 (26.1)	30.6 (22.4)	44.0 (29.2)	43.2 (26.4)
COX1	1.3 (0.7)	2.8 (2.8)	3.5 (3.9)	2.8 (1.4)	3.5 (3.5)
NRAM	0. (0.)	9.3 (8.7)	9.3 (10.2)	21.4 (15.8)	21.4 (16.3)
HSP90a	0. (0.)	15.3 (11.9)	16.7 (11.9)	28.3 (16.1)	31.1 (16.1)
HSP90a[a]	0. (0.)	14.1 (10.5)	15.2 (10.5)	23.3 (14.5)	26.8 (14.0)
HXK4	1.1 (1.1)	11.4 (7.9)	9.1 (6.2)	15.6 (8.8)	15.6 (9.9)
FA7	20.2 (20.2)	49.0 (42.5)	47.0 (42.5)	54.0 (41.0)	52.0 (41.0)
KITH	35.1 (21.1)	35.4 (23.8)	31.7 (22.9)	34.1 (24.2)	30.5 (26.9)
FABP4	31.9 (16.0)	33.9 (18.9)	36.2 (20.9)	32.8 (18.6)	32.8 (23.0)
PA2GA	2.0 (2.0)	6.5 (7.5)	10.9 (8.6)	18.2 (12.9)	16.1 (12.9)

[a]*Excluding the macrocycle containing molecules for calculating the EF.*

even for reduced systems. Moreover, in this case further caution should be exerted not to artificially deform the molecular system.

As stated before, a wide range of enrichment factors calculated from docking with AD Vina was taken into account for selecting the target proteins for this benchmark. It can be readily seen from **Table 4** that using any variant of QM docking scoring has an impressive improvement over AD Vina, especially in those cases with low AD Vina *EF*. This happens even in the simplest case of $QMDS_1$, where no relaxation is performed on the PL complexes.

It is clear that PL relaxation, even using a MM-based approach, has on average a positive effect for calculating the QM docking score. Moreover, in those cases where the *EF*(1) slightly decreases (KITH, FABP4), the *EF*(2) is conserved. Focusing in the analysis of $QMDS_2$ and $QMDS_{2d}$, inclusion of the deformation contribution (second and third term in Equation 6) slightly deteriorates the results in ESR1, FA7, KITH, and PA2GA. However, in all but ESR1, *EF*(2) improves after inclusion of the deformation term (as it also happens in the other cases where *EF*(1) increases or is constant, CDK2, COX1, NRAM, HSP90a, HXK4, and FABP4). Considering that the effect on *EF*(1) is in no way dramatic, and that *EF*(2) (which also refers to early enrichment), improves except in one case, we state that the deformation terms are necessary to obtain better enrichment factors, though this should obviously be validated in a larger-scale benchmark. We hypothesize that this slight deterioration might be related to a small noise introduced upon energy minimization, which is canceled out in the $QMDS_2$ case. In the special case of HSP90a, the consideration of 19 macrocycle containing molecules has a negative effect in the *EF* calculation. We hypothesize that the strong performance of QM-scoring is due to a better representation of intra- and inter-molecular interactions, though of course further validation and benchmarking is still needed to confirm this.

In **Figure 1**, the ROC plots of $QMDS_2$ and AD Vina for the 10 systems are shown, including the corresponding AUC values. Analysis of the curves confirm what has been noted above based on *EF*, exhibiting the QM-score excellent results.

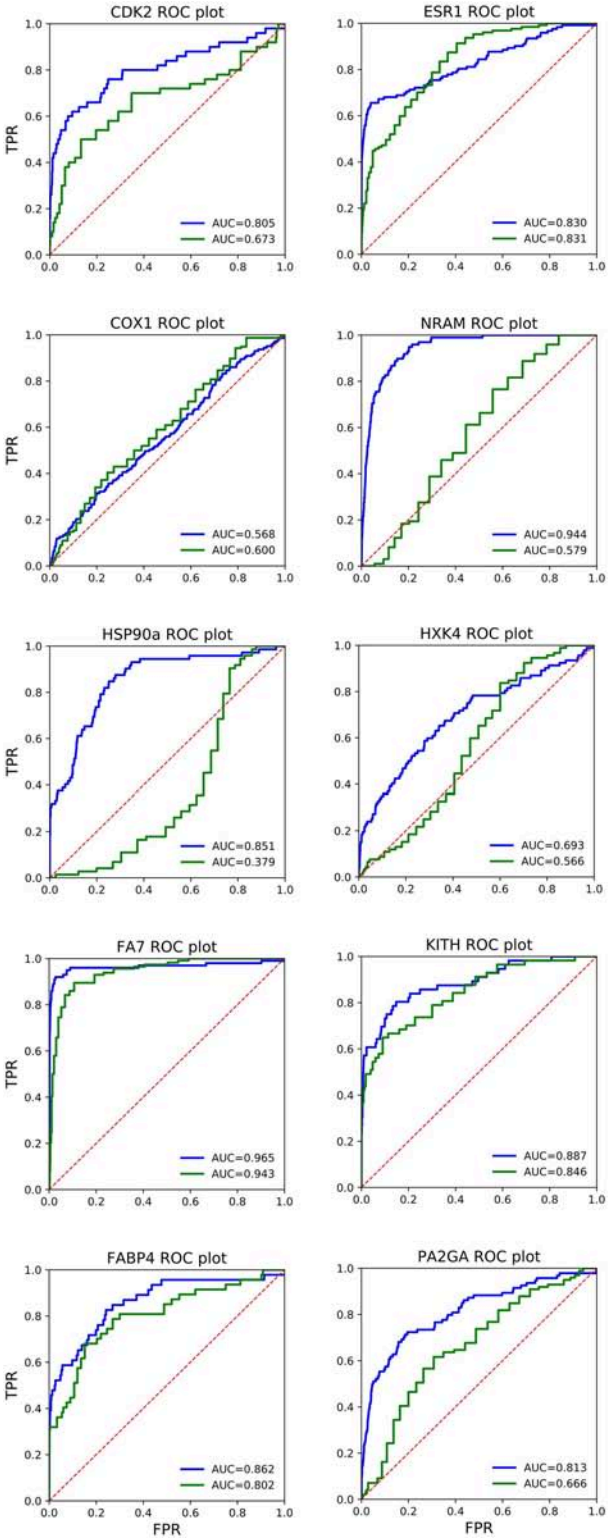


FIGURE 1 | Receiver operating characteristic (ROC) plots of AutoDock Vina (red line) and QM-scoring $QMDS_{2d}$ (blue line) for the 10 systems studied. The dotted line corresponds to random selection (AUC = 0.5). FPR, False Positive Rate; TPR, True Positive Rate; AUC, Area under the curve.

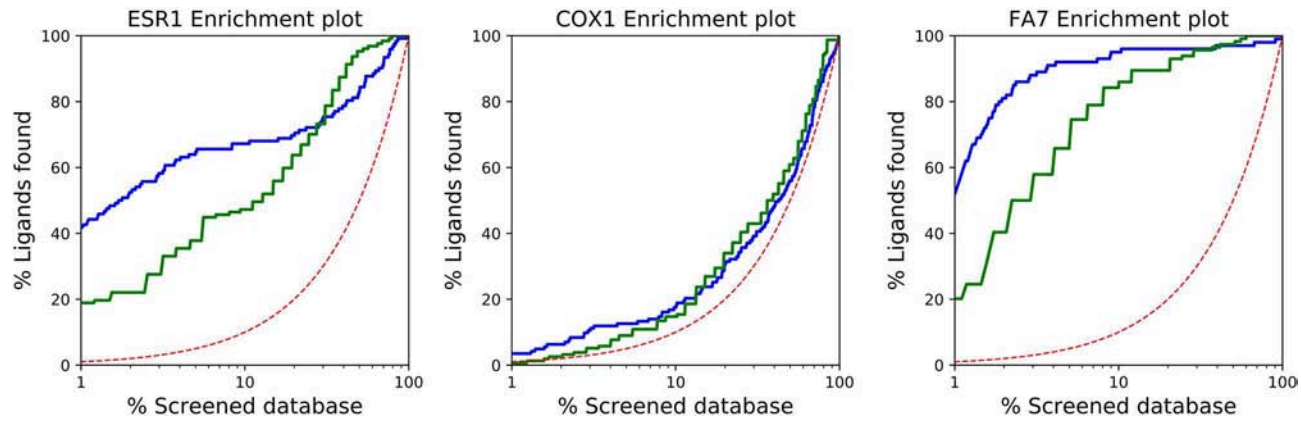


FIGURE 2 | Enrichment plots for ESR1, COX1, and FA7 using AutoDock Vina (red line) and QM-scoring $QMDS_{2d}$ (blue line). The dotted line corresponds to random selection.

Interestingly, in ESR1 both scoring methods show basically the same AUC, which is in conflict with the large difference in EF values reported in **Table 4**. To clarify this issue, in **Figure 2** we show the enrichment plot associated to ESR1. It can be seen that AD out-performs $QMDS_2$ after 30% of the screened database, a region of no importance for drug discovery; for early enrichment, the enrichment plot in **Figure 2** confirms the trend observed in **Table 4** that $QMDS_2$ is remarkable superior in the initial part of the ranking. A similar behavior is observed for FA7 (cf. **Table 4** and **Figures 1**, **2**). In the case of COX1, while the AUC of the QM-score is slightly less than AD Vina, the enrichment plot in **Figure 2** shows that for early enrichment, QM-scoring out-performs AD Vina.

While our QM-score appears to be a very promising for HTD, and QM calculations are in principle more accurate than classical ones to describe molecular interactions, there are still a number of approximations which prevent the direct use of *QMDS* as a measure of actual absolute binding free energy. We mention three, among many: (i) QM local energy minimization was not performed (for computational efficiency, as said above); (ii) Vibrational entropies were not included; (iii) PM7 has not been parameterized to reproduce binding free energies. Our QM calculations were in the order of −70 kcal/mol, in agreement with recent binding enthalpy calculations on protein-ligand complexes using a PM7+COSMO approach (Sulimov et al., 2019a), where in spite of the difference between experimental and calculated absolute binding enthalpies, very good correlation with experimental values was obtained. It should be added that it is also well-known that traditional scoring functions correlate poorly with binding energy (cf. Enyedy and Egan, 2008, among others). Moreover, among traditional scoring functions there is no uniform scale: While AutoDock and Glide (Friesner et al., 2004; Halgren et al., 2004) are roughly in the range of −10 kcal/mol and higher, others are around −60 kcal/mol. Moreover, even end-point methods such as MM/PBSA or MM/GBSA exhibited calculated binding free energies in the order of −60 kcal/mol, or even lower when changes in vibrational entropy are not included (Zhong and Carlson, 2005), and even when including those terms (Woo and Roux, 2005; Anisimov and Cavasotto, 2011; Anisimov et al., 2011). Thus, we stress that *QMDS* should be considered a score, not a measure of absolute binding energy. It is aimed for relative binding energy estimation, and thus for compound ranking.

On average, the computing time of this QM docking score on a single core is ∼6–8 minutes (depending on the size of the system, and on whether the deformation energy term is considered), around an order of magnitude slower than a MM-based DS.

CONCLUSIONS AND PERSPECTIVES

Docking programs have been so far based on molecular mechanics force-fields. However, a better description of protein-ligand interactions could be achieved, in principle, with quantum mechanical methods, which are theoretically exact, capture the underlying physics of the molecular system, and account for all contributions to the energy, including those effects usually missing in force-fields, such as electronic polarization, covalent-bond formation, and charge transfer. Moreover, a quantum mechanical formulation is generally valid across the chemical space, thus avoiding the force-field parameterizations.

We present a new QM-based high-throughput docking scoring function, which has been evaluated on 10 protein systems belonging to different protein families, displaying diverse binding site properties, and covering a wide range of enrichment factors computed with a traditional docking program. As shown in **Table 4**, even the simplest QM docking scoring function (where no relaxation is performed on the reference docked protein-small-molecule complex) shows excellent results in terms of enrichment (screening power). In fact, the improvement over AutoDock Vina on all systems is remarkable, especially in those cases with very low AD Vina enrichment. Upon complex relaxation, the improvement is even larger, regardless of whether the protein and ligand deformation terms are included or not.

We highlight that our main aim is to develop and validate a simple, straightforward approach for QM docking scoring, from which further developments can be built. Clearly, to further improve this methodology, several aspects should be analyzed: (i) a wider and extensive benchmark on many more target systems; (ii) comparison with other MM-based standard docking scoring functions; (iii) evaluation of other QM formalisms, continuum solvent models and their associated parameters (atomic radii and surface tension parameters); (iv) structural relaxation strategies; (v) use of single or multiple poses for scoring; (vi) the vibrational entropy changes upon binding. All of these considerations are important. They are currently being investigated and will be published in due course. Considering the outstanding improvements to our methods, we highlight that the *QMDS* should be used as a score and not an estimation to the absolute binding energy.

In terms of CPU time, our QM docking scoring function is approximately 10 times slower than MM-based standard scores on a single core. In spite of this, our impressive results on a set of 10 different protein targets highlight the huge potential of QM-based scoring. Moreover, considering future developments in QM theory, algorithms and computer hardware, it can be hypothesized that semi-empirical methods (or low-cost DFT methods) will replace FF over the following years (Grimme and Schreiner, 2018). We thus believe it is fully justified and of the utmost importance to develop the next generation of QM-based scoring functions for HTD toward highly accurate methods for the identification and optimization of small-molecule modulators of pharmaceutically relevant targets.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

CC conceived and designed the research, performed simulations, analysis, interpretation, and wrote the paper. MA performed quantum mechanical simulations and enrichment calculations and contributed with interpretation and analysis. All authors approved the manuscript for publication.

FUNDING

This work was supported by the National Agency for the Promotion of Science and Technology (ANPCyT) (PICT-2014-3599 and PICT-2017-3767).

ACKNOWLEDGMENTS

The Authors thank Prof. F. Javier Luque for helpful discussions. CNC thanks Molsoft LLC (San Diego, CA) for providing an academic license for the ICM program. The authors thank the National System of High Performance Computing (Sistemas Nacionales de Computación de Alto Rendimiento, SNCAD), the Centro de Computación de Alto Rendimiento (Computational Centre of High Performance Computing, CeCAR), and the Centro de Cálculo de Alto Desempeño (Universidad Nacional de Córdoba) for granting the use of their computational resources.

REFERENCES

Abagyan, R., and Totrov, M. (1994). Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* 235, 983–1002. doi: 10.1006/jmbi.1994.1052

Abagyan, R., Totrov, M., and Kuznetsov, D. (1994). ICM - a new method for protein modeling and design - applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* 15, 488–506. doi: 10.1002/jcc.540150503

Amadasi, A., Surface, J. A., Spyrakis, F., Cozzini, P., Mozzarelli, A., and Kellogg, G. E. (2008). Robust classification of "relevant" water molecules in putative protein binding sites. *J. Med. Chem.* 51, 1063–1067. doi: 10.1021/jm701023h

Anisimov, V. M., and Cavasotto, C. N. (2011). Quantum mechanical binding free energy calculation for phosphopeptide inhibitors of the Lck SH2 domain. *J. Comput. Chem.* 32, 2254–2263. doi: 10.1002/jcc.21808

Anisimov, V. M., Ziemys, A., Kizhake, S., Yuan, Z., Natarajan, A., and Cavasotto, C. N. (2011). Computational and experimental studies of the interaction between phospho-peptides and the C-terminal domain of BRCA1. *J. Comput. Aided Mol. Des.* 25, 1071–1084. doi: 10.1007/s10822-011-9484-3

Aucar, M. G., and Cavasotto, C. N. (2020). Molecular docking using quantum mechanical-based methods. *Methods Mol. Biol.* 2114, 269–284. doi: 10.1007/978-1-0716-0282-9_17

Bursulaya, B. D., Totrov, M., Abagyan, R., and Brooks, C. L. 3rd. (2003). Comparative study of several algorithms for flexible ligand docking. *J. Comput. Aided Mol. Des.* 17, 755–763. doi: 10.1023/B:JCAM.0000017496.76572.6f

Cavasotto, C. N. (2012). "Binding free energy calculations and scoring in small-molecule docking," in *Physico-Chemical and Computational Approaches to Drug Discovery,* eds F. J. Luque and X. Barril (London: Royal Society of Chemistry), 195–222.

Cavasotto, C. N., and Abagyan, R. A. (2004). Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* 337, 209–225. doi: 10.1016/j.jmb.2004.01.003

Cavasotto, C. N., Adler, N. S., and Aucar, M. G. (2018). Quantum chemical approaches in structure-based virtual screening and lead optimization. *Front. Chem.* 6:188. doi: 10.3389/fchem.2018.00188

Cavasotto, C. N., Aucar, M. G., and Adler, N. S. (2019). Computational chemistry in drug lead discovery and design. *Int. J. Quantum Chem.* 119:e25678. doi: 10.1002/qua.25678

Cavasotto, C. N., and Orry, A. J. (2007). Ligand docking and structure-based virtual screening in drug discovery. *Curr. Top. Med. Chem.* 7, 1006–1014. doi: 10.2174/156802607780906753

Cavasotto, C. N., and Singh, N. (2008). Docking and high throughput docking: successes and the challenge of protein flexibility. *Curr. Comput. Aided Drug Design* 4, 221–234. doi: 10.2174/157340908785747474

Ciancetta, A., and Moro, S. (2015). "Protein-ligand docking: virtual screening and applications to drug discovery," in *In silico Drug Discovery and Design: Theory, Methods, Challenges, and Applications,* ed C.N. Cavasotto (Boca Raton, FL: CRC Press, Taylor & Francis Group), 189–213.

Cozzini, P., Fornabaio, M., Mozzarelli, A., Spyrakis, F., Kellogg, G. E., and Abraham, D. J. (2006). Water: how to evaluate its contribution in protein-ligand interactions. *Int. J. Quantum Chem.* 106, 647–651. doi: 10.1002/qua.20812

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* 47, 20–33. doi: 10.1016/j.jhealeco.2016.01.012

Ehrlich, S., Goller, A. H., and Grimme, S. (2017). Towards full quantum-mechanics-based protein-ligand binding affinities. *ChemPhysChem* 18, 898–905. doi: 10.1002/cphc.201700082

Enyedy, I. J., and Egan, W. J. (2008). Can we use docking and scoring for hit-to-lead optimization? *J. Comput. Aided Mol. Des.* 22, 161–168. doi: 10.1007/s10822-007-9165-4

Eyrilmez, S. M., Kopruluoglu, C., Rezac, J., and Hobza, P. (2019). Impressive enrichment of semiempirical quantum mechanics-based scoring function: HSP90 protein with 4541 inhibitors and decoys. *ChemPhysChem* 20, 2759–2766. doi: 10.1002/cphc.201900628

Forti, F., Cavasotto, C. N., Orozco, M., Barril, X., and Luque, F. J. (2012). A multilevel strategy for the exploration of the conformational flexibility of small molecules. *J. Chem. Theor. Comput.* 8, 1808–1819. doi: 10.1021/ct300097s

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.* 47, 1739–1749. doi: 10.1021/jm0306430

Gaieb, Z., Liu, S., Gathiaka, S., Chiu, M., Yang, H., Shao, C., et al. (2018). D3R grand challenge 2: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. *J. Comput. Aided Mol. Design* 32, 1–20. doi: 10.1007/s10822-017-0088-4

Gaieb, Z., Parks, C. D., Chiu, M., Yang, H., Shao, C., Walters, W. P., et al. (2019). D3R grand challenge 3: blind prediction of protein-ligand poses and affinity rankings. *J. Comput. Aided Mol. Design* 33, 1–18. doi: 10.1007/s10822-018-0180-4

Gathiaka, S., Liu, S., Chiu, M., Yang, H., Stuckey, J. A., Kang, Y. N., et al. (2016). D3R grand challenge 2015: evaluation of protein-ligand pose and affinity predictions. *J. Comput. Aided Mol. Design* 30, 651–668. doi: 10.1007/s10822-016-9946-8

Gatica, E. A., and Cavasotto, C. N. (2012). Ligand and decoy sets for docking to g protein-coupled receptors. *J. Chem. Inf. Model* 52, 1–6. doi: 10.1021/ci200412p

Genheden, S., and Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* 10, 449–461. doi: 10.1517/17460441.2015.1032936

Grimme, S., and Schreiner, P. R. (2018). Computational chemistry: the fate of current methods and future challenges. *Angew. Chem. Int. Ed. Engl.* 57, 4170–4176. doi: 10.1002/anie.201709943

Guedes, I. A., Pereira, F. S. S., and Dardenne, L. E. (2018). Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. *Front. Pharmacol.* 9:1089. doi: 10.3389/fphar.2018.01089

Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *J. Med. Chem.* 47, 1750–1759. doi: 10.1021/jm030644s

Huang, N., Shoichet, B. K., and Irwin, J. J. (2006). Benchmarking sets for molecular docking. *J. Med. Chem.* 49, 6789–6801. doi: 10.1021/jm0608356

Jorgensen, W. L. (2009). Efficient drug lead discovery and optimization. *Acc. Chem. Res.* 42, 724–733. doi: 10.1021/ar800236t

Jorgensen, W. L. (2012). Challenges for academic drug discovery. *Angew. Chem. Int. Ed. Engl.* 51, 11680–11684. doi: 10.1002/anie.201204625

Kerrigan, J. E. (2013). Molecular dynamics simulations in drug design. *Methods Mol. Biol.* 993, 95–113. doi: 10.1007/978-1-62703-342-8_7

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* 3, 935–949. doi: 10.1038/nrd1549

Klamt, A., and Schüürmann, G. (1993). COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc. Perkin Trans.* 2, 799–805. doi: 10.1039/P2993000 00799

Lagarde, N., Ben Nasr, N., Jeremie, A., Guillemain, H., Laville, V., Labib, T., et al. (2014). NRLiSt BDB, the manually curated nuclear receptors ligands and structures benchmarking database. *J. Med. Chem.* 57, 3117–3125. doi: 10.1021/jm500132p

Lepšík, M., Rezáč, J., Kolár, M., Pecina, A., Hobza, P., and Fanfrlík, J. (2013). The semiempirical quantum mechanical scoring function for *in silico* drug design. *ChemPlusChem* 78, 921–931. doi: 10.1002/cplu.201300199

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* 30, 2785–2791. doi: 10.1002/jcc.21256

Mucs, D., and Bryce, R. A. (2013). The application of quantum mechanics in structure-based drug design. *Expert Opin. Drug Discov.* 8, 263–276. doi: 10.1517/17460441.2013.752812

Mysinger, M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E) - better ligands and decoys for better benchmarking. *J. Med. Chem.* 55, 6582–6594. doi: 10.1021/jm300687e

Neves, M. A., Totrov, M., and Abagyan, R. (2012). Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J. Comput. Aided Mol. Design* 26, 675–686. doi: 10.1007/s10822-012-9547-0

Pagadala, N. S., Syed, K., and Tuszynski, J. (2017). Software for molecular docking: a review. *Biophys. Rev.* 9, 91–102. doi: 10.1007/s12551-016-0247-1

Palacio-Rodriguez, K., Lans, I., Cavasotto, C. N., and Cossio, P. (2019). Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. *Sci. Rep.* 9:5142. doi: 10.1038/s41598-019-41594-3

Pecina, A., Brynda, J., Vrzal, L., Gnanasekaran, R., Horejsi, M., Eyrilmez, S., et al. (2018). Ranking power of the SQM/COSMO scoring function on carbonic anhydrase II-inhibitor complexes. *ChemPhysChem* 19, 873–879. doi: 10.1002/cphc.201701104

Pecina, A., Haldar, S., Fanfrlik, J., Meier, R., Rezac, J., Lepsik, M., et al. (2017). SQM/COSMO scoring function at the DFTB3-D3H4 level: unique identification of native protein-ligand poses. *J. Chem. Infor. Model.* 57, 127–132. doi: 10.1021/acs.jcim.6b00513

Pecina, A., Meier, R., Fanfrlik, J., Lepsik, M., Rezac, J., Hobza, P., et al. (2016). The SQM/COSMO filter: reliable native pose identification based on the quantum-mechanical description of protein-ligand interactions and implicit COSMO solvation. *Chem. Commun.* 52, 3312–3315. doi: 10.1039/C5CC09499B

Phatak, S. S., Gatica, E. A., and Cavasotto, C. N. (2010). Ligand-steered modeling and docking: a benchmarking study in class A G-protein-coupled receptors. *J. Chem. Inf. Model.* 50, 2119–2128. doi: 10.1021/ci100285f

Phatak, S. S., Stephan, C. C., and Cavasotto, C. N. (2009). High-throughput and *in silico* screenings in drug discovery. *Exp. Opin. Drug Discov.* 4, 947–959. doi: 10.1517/17460440903190961

Raha, K., and Merz, K. M. Jr. (2004). A quantum mechanics-based scoring function: study of zinc ion-mediated ligand binding. *J. Am. Chem. Soc.* 126, 1020–1021. doi: 10.1021/ja038496i

Raha, K., and Merz, K. M. Jr. (2005). Large-scale validation of a quantum mechanics based scoring function: predicting the binding affinity and the binding mode of a diverse set of protein-ligand complexes. *J. Med. Chem.* 48, 4558–4575. doi: 10.1021/jm048973n

Reddy, M. R., Reddy, C. R., Rathore, R. S., Erion, M. D., Aparoy, P., Reddy, R. N., et al. (2014). Free energy calculations to estimate ligand-binding affinities in structure-based drug design. *Curr. Pharm. Des.* 20, 3323–3337. doi: 10.2174/13816128113199990604

Rezáč, J., and Hobza, P. (2012). Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods. *J. Chem. Theor. Comput.* 8, 141–151. doi: 10.1021/ct200751e

Rognan, D. (2011). "Docking methods for virtual screening: principles and recent advances," in *Virtual Screening. Principles, Challenges and Practical Guidelines, ed C. Sotriffer* (Weinheim: Wiley-VCH Verlag), 153–176.

Slater, O., and Kontoyianni, M. (2019). The compromise of virtual screening and its impact on drug discovery. *Expert Opin. Drug Discov.* 14, 619–637. doi: 10.1080/17460441.2019.1604677

Sotriffer, C. A. (2015). "Protein-ligand docking: from basic principles to advanced applications," in *In silico Drug Discovery and Design: Theory, Methods, Challenges, and Applications*, ed C.N. Cavasotto (Boca Raton, FL: CRC Press, Taylor & Francis Group), 155–188.

Spyrakis, F., Bidon-Chanal, A., Barril, X., and Luque, F. J. (2011). Protein flexibility and ligand recognition: challenges for molecular modeling. *Curr. Top. Med. Chem.* 11, 192–210. doi: 10.2174/156802611794863571

Spyrakis, F., and Cavasotto, C. N. (2015). Open challenges in structure-based virtual screening: Receptor modeling, target flexibility consideration and active site water molecules description. *Arch. Biochem. Biophys.* 583, 105–119. doi: 10.1016/j.abb.2015.08.002

Stewart, J. J. (2013). Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* 19, 1–32. doi: 10.1007/s00894-012-1667-x

Stewart, J. J. P. (1996). Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. *Int. J. Quant. Chem.* 58, 133–146.

Stewart, J. J. P. (2007). Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* 13, 1173–1213. doi: 10.1007/s00894-007-0233-4

Stewart, J. J. P. (2016). *"MOPAC2016"*. Colorado Spring, CO: Stewart Computational Chemistry. Available online at: www.OpenMOPAC.net

Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., et al. (2019). Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Inform. Model.* 59, 895–913. doi: 10.1021/acs.jcim.8b00545

Sulimov, A., Kutov, D., Gribkova, A., Ilin, I., Tashchilova, A., and Sulimov, V. (2019a). "Search for approaches to supercomputer quantum-chemical docking," in *Supercomputing,* eds V. Voevodin and S. Sobolev (Cham: Springer International Publishing), 363–378.

Sulimov, A. V., Kutov, D. C., Katkova, E. V., Ilin, I. S., and Sulimov, V. B. (2017a). New generation of docking programs: supercomputer validation of force fields and quantum-chemical methods for docking. *J. Mol. Graph. Model.* 78, 139–147. doi: 10.1016/j.jmgm.2017.10.007

Sulimov, A. V., Kutov, D. C., Katkova, E. V., and Sulimov, V. B. (2017b). Combined docking with classical force field and quantum chemical semiempirical method PM7. *Adv. Bioinformatics* 2017:7167691. doi: 10.1155/2017/7167691

Sulimov, V. B., Kutov, D. C., and Sulimov, A. V. (2019b). Advances in docking. *Curr. Med. Chem.* 26, 7555–7580. doi: 10.2174/0929867325666180904115000

Sun, H., Duan, L., Chen, F., Liu, H., Wang, Z., Pan, P., et al. (2018). Assessing the performance of MM/PBSA and MM/GBSA methods. 7. entropy effects on the performance of end-point binding free energy calculation approaches. *Phys. Chem. Chem. Phys.* 20, 14450–14460. doi: 10.1039/C7CP07623A

Trott, O., and Olson, A. J. (2010). AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455–461. doi: 10.1002/jcc.21334

Wang, E., Sun, H., Wang, J., Wang, Z., Liu, H., Zhang, J. Z. H., et al. (2019). End-point binding free energy calculation with MM/PBSA and MM/GBSA: strategies and applications in drug design. *Chem. Rev.* 119, 9478–9508. doi: 10.1021/acs.chemrev.9b00055

Wang, Z., Sun, H., Yao, X., Li, D., Xu, L., Li, Y., et al. (2016). Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys. Chem. Chem. Phys.* 18, 12964–12975. doi: 10.1039/C6CP01555G

Warren, G. L., Andrews, C. W., Capelli, A. M., Clarke, B., LaLonde, J., Lambert, M. H., et al. (2006). A critical assessment of docking programs and scoring functions. *J. Med. Chem.* 49, 5912–5931. doi: 10.1021/jm050362n

Woo, H. J., and Roux, B. (2005). Calculation of absolute protein-ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. U.S.A.* 102, 6825–6830. doi: 10.1073/pnas.0409005102

Zhong, H., and Carlson, H. A. (2005). Computational studies and peptidomimetic design for the human p53-MDM2 complex. *Proteins* 58, 222–234. doi: 10.1002/prot.20275

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Cavasotto and Aucar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 23 April 2020
doi: 10.3389/fchem.2020.00296

Exploring the Use of Compound-Induced Transcriptomic Data Generated From Cell Lines to Predict Compound Activity Toward Molecular Targets

Benoît Baillif[1], *Joerg Wichard*[2], *Oscar Méndez-Lucio*[1,3†] *and David Rouquié*[1*†]

[1] Bayer SAS, Bayer CropScience, Sophia Antipolis, France, [2] Department of Genetic Toxicology, Bayer AG, Berlin, Germany, [3] Bloomoon, Villeurbanne, France

OPEN ACCESS

Edited by:
Kamil Kuca,
University of Hradec Králové, Czechia

Reviewed by:
Chanin Nantasenamat,
Mahidol University, Thailand
Rajeev K. Singla,
West China Hospital, Sichuan University, China

***Correspondence:**
David Rouquié
david.rouquie@bayer.com

† These authors jointly supervised this work

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: 02 December 2019
Accepted: 25 March 2020
Published: 23 April 2020

Citation:
Baillif B, Wichard J, Méndez-Lucio O and Rouquié D (2020) Exploring the Use of Compound-Induced Transcriptomic Data Generated From Cell Lines to Predict Compound Activity Toward Molecular Targets. Front. Chem. 8:296. doi: 10.3389/fchem.2020.00296

Pharmaceutical or phytopharmaceutical molecules rely on the interaction with one or more specific molecular targets to induce their anticipated biological responses. Nonetheless, these compounds are also prone to interact with many other non-intended biological targets, also known as off-targets. Unfortunately, off-target identification is difficult and expensive. Consequently, QSAR models predicting the activity on a target have gained importance in drug discovery or in the de-risking of chemicals. However, a restricted number of targets are well characterized and hold enough data to build such *in silico* models. A good alternative to individual target evaluations is to use integrative evaluations such as transcriptomics obtained from compound-induced gene expression measurements derived from cell cultures. The advantage of these particular experiments is to capture the consequences of the interaction of compounds on many possible molecular targets and biological pathways, without having any constraints concerning the chemical space. In this work, we assessed the value of a large public dataset of compound-induced transcriptomic data, to predict compound activity on a selection of 69 molecular targets. We compared such descriptors with other QSAR descriptors, namely the Morgan fingerprints (similar to extended-connectivity fingerprints). Depending on the target, active compounds could show similar signatures in one or multiple cell lines, whether these active compounds shared similar or different chemical structures. Random forest models using gene expression signatures were able to perform similarly or better than counterpart models built with Morgan fingerprints for 25% of the target prediction tasks. These performances occurred mostly using signatures produced in cell lines showing similar signatures for active compounds toward the considered target. We show that compound-induced transcriptomic data could represent a great opportunity for target prediction, allowing to overcome the chemical space limitation of QSAR models.

Keywords: target prediction, compound-induced transcriptomic data, QSAR, machine learning, cellular context

INTRODUCTION

Biologically active molecules rely on the interaction with one or more molecular targets (Hughes et al., 2000). In the context of hit discovery both in pharmaceutical or in phytopharmaceutical industries, a major objective is to be able to screen molecule candidates for their activity toward a target of interest, and assessing compound activity toward off-targets, that can cause adverse effects *in vivo* (Rouquié et al., 2015). Testing activity of every candidate on a battery of targets represent a complex task that requires major R&D costs. A potential solution to predict candidate's activity with a lower cost is to perform computational methods using more general measured or calculated descriptors (Chen et al., 2016; Vamathevan et al., 2019).

A commonly used technique is to compute descriptors from chemical structures, like the extended-connectivity fingerprints (ECFPs) and use them for prediction, relying on the quantitative structure-activity relationship (QSAR) principle, i.e., molecules sharing a similar structure may share a similar activity profile (Rogers and Hahn, 2010; Cherkasov et al., 2014). However, such molecule descriptors show limitations: they do not perform well for all target prediction tasks depending on the quantity and quality of available activity data, prediction is limited to the applicability domain (depending on the training set used), and a small change in chemical structure can lead to a large change in biological response (activity cliffs) (Cruz-Monteagudo et al., 2014).

Additional descriptors have been proposed to circumvent such QSAR drawbacks, such as measurements from large scale biological assays (Petrone et al., 2012; Laufkötter et al., 2019). Results from high throughput screening (HTS) assays, such as bioactivity experiments, can be used as fingerprints (HTSFPs) in predictive models for specific targets. Petrone et al. (2012) showed that models using HTSFPs were outperforming models using ECFPs for certain targets, and that HTSFP models' predictions were covering a large structural diversity. The main limiting factor of such models is the sparsity of available activity data. Besides bioactivity data, more integrative large-scale biological measurements, like transcriptomics or cell morphology readouts can be used for target prediction (Aliper et al., 2016; Pabon et al., 2018; Scheeder et al., 2018; Simm et al., 2018; Hofmarcher et al., 2019; Kuthuru et al., 2019; Lapins and Spjuth, 2019).

Compound-induced gene expression data are gathered from biological experiments reflecting how the compound acted on one or multiple targets in a specific biological context. Cancer cell lines, being easily cultured, are a commonly used model to generate gene expression data. Hughes et al. (2000) proved that enough data allows to use pattern-matching algorithms to study similarity between signatures coming from drug induction (Hughes et al., 2000). Lamb et al. (2006) invented the concept of Connectivity Map (CMAP), creating relationships between small molecules, genes and diseases (Lamb et al., 2006). Since then, transcriptomics data have been shown to be useful to identify new molecules with biological activity (Hieronymus et al., 2006; Wei et al., 2006). Recently, a large public CMAP L1000 dataset was released representing more than 300,000 Gene Expression Signatures (GESs) of cell line responses to so-called perturbagens (Subramanian et al., 2017). GESs were produced for more than 20,000 compounds in 80 human cancer cell lines, tested at various concentration and exposition time. The large scale of this dataset allows the use of GESs in machine learning models for target prediction or drug repurposing (Lee et al., 2016; De Wolf et al., 2018).

In the current work, we investigated whether we could predict compound activity toward a larger number of molecular targets based on their GESs extracted from the CMAP L1000 dataset. In addition, we were interested to reveal how machine learning models using GESs perform compared to models using more traditional QSAR descriptors, such as the Morgan fingerprints.

We show that random forest models built using compound-induced GES were able to effectively predict targets, especially if they were produced from a cell line showing similar GESs between active compounds on the evaluated target. For 25% of the target prediction tasks, GESs models had similar or higher performances than models using Morgan fingerprints, offering an opportunity to escape from the chemical space limitation associated with QSAR approaches.

MATERIALS AND METHODS

Gene Expression Signatures (GESs) Acquisition

The CMAP L1000 dataset was obtained from two GEO repositories: GSE92742, corresponding to the first phase of L1000 (pilot, 2012–2015) and GSE70138, which is the second phase (production, on-going). GESs generation was described by Subramanian et al. (2017).

For this study, we only used Level 5 GESs meaning that each GES is represented by an instance, that is a combination of a perturbagen (chemical or gene deletion), cell line, concentration and time point, and is composed by the plate-normalized expression z-scores of the whole genome, inferred from 978 landmark genes (measured gene that can be used for whole transcriptome inference). We focused on landmark signatures of compound perturbagens, which comprises 333,273 GESs for 21,300 unique compounds. GESs obtained in the exact same condition were averaged, to have one signature per condition.

Among all obtained GESs, the ones generated at a 10 μM and 24 h time point were selected (as shown in **Figure 1**), as this condition was the most represented in the dataset and facilitate the comparison of results. GESs from the 8 most profiled cell lines were used; cell line and number of GESs are presented in **Table 1**. Also, only GESs generated by compounds with known structure were selected. In total, the working dataset contains 39,544 GESs obtained from 9,035 compounds.

Activity Data Acquisition

Annotations about activity or inactivity was retrieved from the PubChem BioAssay database, using available CIDs documented in the L1000 signature metadata, excepted for TUBB actives, that were extracted from the Drug Repurposing Hub of the LINCS (Wang et al., 2014; Corsello et al., 2017). Activity data were compiled in a binary activity matrix (1 for active, 0 for

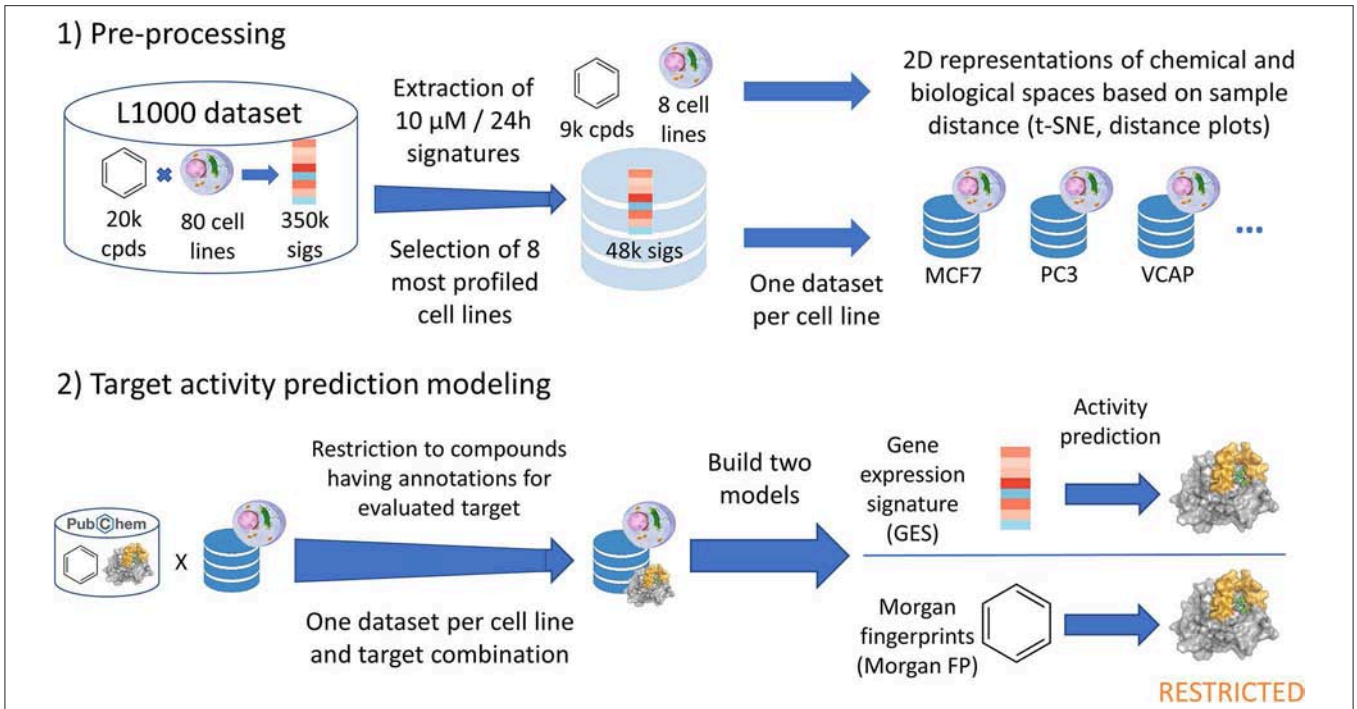


FIGURE 1 | Data analysis pipeline performed in current work. Starting from the CMAP L1000 dataset, signatures produced at 10 µM and 24 h from 8 cell lines were extracted and used in t-SNE and distance plots. One dataset was built per cell line (GES and corresponding compound structure), and each of these datasets were restricted to compounds having known annotations (active or inactive) for the evaluated target. For each target—cell line dataset, a first model was built using the gene expression signatures (GES model). Alongside, a second counterpart model was built using the Morgan fingerprints of compounds whose signatures were used in the first model (Morgan FP model).

TABLE 1 | The 8 core cell lines used in this work, with their corresponding number of GESs for compounds with known structure tested at 10 µM/24 h.

Cell line	Primary site	Subtype	Number of 10 µM−24 h signatures
A375	Skin	Malignant melanoma	3,525
A549	Lung	Non small cell lung cancer| carcinoma	5,267
HA1E	Kidney	Normal kidney	3,646
HCC515	Lung	Carcinoma	1,932
HT29	Large intestine	Colorectal adenocarcinoma	3,192
MCF7	Breast	Adenocarcinoma	7,546
PC3	Prostate	Adenocarcinoma	8,071
VCAP	Prostate	Carcinoma	6,365

inactive, empty if unknown). At least one annotation among 1,388 targets was found for 7,804 of the 9,035 compounds (512,406 annotations were found, representing 4.8% of the full activity matrix).

Representation of Chemical and Biological Spaces

For each compound, binary Morgan fingerprints were computed. The Morgan fingerprints were employed as input of a t-SNE (t-distributed stochastic neighbor embedding) algorithm (using the sklearn implementation) using Dice distance as metric, to reduce the data to a two-dimensional output that can be plotted to represent the chemical space (Van Der Maaten and Hinton, 2008; Pedregosa et al., 2011). Information of the number of targets per compound was included as color-code using a blue gradient in the plot.

The whole extracted 10 µM/24 h signature dataset was used as input for a second t-SNE using the cosine distance metric, representing the overall biological (response) space wherein each cell line was color-coded in the plot. For every cell line, a t-SNE using the cosine distance metric was performed using all GESs profiled in the cell line, generating 2D biological space.

Machine Learning Modeling

Targets for which we know at least 50 active compounds (representing between 1 and 63% of active per target) were selected for machine learning modeling in order to have a minimum number of actives in test sets to evaluate the model performances, and for computational time purposes. Complete information on the number of active and inactive compounds for these selected targets is listed in **Table 2**.

Subsequently, for each cell line GES dataset, we created a target—cell line GES dataset, restricting to compounds for which target activity was known as shown in **Figure 1** (this step caused the number of possible models to drop from 1,104 to 990). Datasets for each target prediction task were split into a training set (67% of the data) and a test set (remaining 33% of the data). Two models for target activity prediction were trained using

TABLE 2 | Molecular targets used in this work, with number of active and inactive compounds in total, and in each cell line GES dataset.

Gene name	Total inactive	Total active	Description	A375		A549		HA1E		HCC515		HT29		MCF7		PC3		VCAP	
				Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active
ABCB1	801	96	ATP binding cassette subfamily B member 1 [HGNC:40]	331	42	545	60	424	50	211	32	317	32	770	95	772	95	608	68
ABHD5	2,458	57	Abhydrolase domain containing 5 [HGNC:21396]	908	22	1,533	47	-	-	-	-	-	-	2,030	57	2,045	57	1,730	52
ALOX15	1,136	101	Arachidonate 15-lipoxygenase [HGNC:433]	508	27	764	92	808	52	496	33	498	27	1,062	97	1,066	96	804	91
AR	1,085	103	Androgen receptor [HGNC:644]	580	36	639	64	757	62	374	36	572	37	1,038	94	1,036	93	682	69
ATAD5	2,213	97	Atpase family, AAA domain containing 5 [HGNC:25,752]	1,007	43	1,466	70	1288	60	638	41	922	42	2,087	90	2,090	88	1,629	72
ATXN2	1,897	143	Ataxin 2 [HGNC:10555]	695	36	1,139	104	950	50	501	41	652	41	1556	123	1,558	129	1,243	102
BAZ2B	1,252	143	Bromodomain adjacent to zinc finger domain 2B [HGNC:963]	516	63	873	101	653	76	319	37	470	56	1,199	135	1,197	136	974	112
BRCA1	3,008	160	BRCA1, dna repair associated [HGNC:1100]	1,117	67	1945	116	1,519	63	800	33	1,014	50	2,537	148	2,549	150	2,134	131
CBX1	1,999	80	Chromobox 1 [HGNC:1551]	899	41	1,412	64	1,120	61	532	38	809	40	1,899	75	1,904	76	1,576	67
CHRM1	2,433	86	Cholinergic receptor muscarinic 1 [HGNC:1950]	906	49	1,544	62	1,055	59	460	39	791	48	2,036	84	2051	83	1,739	65
CHRM4	2,476	70	Cholinergic receptor muscarinic 4 [HGNC:1953]	908	45	1,552	54	1,057	55	460	40	793	45	2,049	68	2,064	67	1,751	54
CHRM5	2,478	62	Cholinergic receptor muscarinic 5 [HGNC:1954]	908	41	1,553	47	1,057	49	461	36	793	39	2,050	62	2,065	61	1,751	50
CYP1A2	307	526	Cytochrome P450 family 1 subfamily A member 2 [HGNC:2596]	145	265	183	399	219	428	136	277	144	261	299	505	294	505	183	410
CYP2C19	717	276	Cytochrome P450 family 2 subfamily C member 19 [HGNC:2621]	329	151	514	213	486	228	289	138	324	148	688	271	684	272	524	225
CYP2C9	708	270	Cytochrome P450 family 2 subfamily C member 9 [HGNC:2623]	310	157	510	197	476	222	285	126	305	154	679	264	674	263	517	207
CYP3A4	1,153	164	Cytochrome P450 family 3 subfamily A member 4 [HGNC:2637]	472	113	780	104	847	133	561	64	467	110	1,070	160	1069	161	802	118
DRD1	1,843	99	Dopamine receptor D1 [HGNC:3020]	807	54	1,295	71	1028	78	526	55	725	54	1,762	91	1,762	91	1,450	71
DRD2	2,262	95	Dopamine receptor D2 [HGNC:3023]	769	58	1,371	73	956	84	474	55	683	57	1,858	93	1873	93	1,541	74
DRD3	2,446	142	Dopamine receptor D3 [HGNC:3024]	877	76	1,432	110	1,129	114	551	78	823	75	2,004	139	2,017	139	1,569	111
EPAS1	2,443	70	Endothelial PAS domain protein 1 [HGNC:3374]	–	–	1,524	52	–	–	–	–	–	–	2,021	64	2,033	67	1,723	57
FEN1	2,100	53	Flap structure-specific endonuclease 1 [HGNC:3650]	961	23	1,496	29	1,213	28	–	–	866	21	1,990	46	1,999	46	1,669	32
GFER	1,589	89	Growth factor, augmenter of liver regeneration [HGNC:4236]	679	37	1,153	59	813	46	363	21	600	29	1,519	81	1,519	81	1,294	70

(Continued)

TABLE 2 | Continued

Gene name	Total inactive	Total active	Description	A375		A549		HA1E		HCC515		HT29		MCF7		PC3		VCAP	
				Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active
GLS	2,989	66	Glutaminase [HGNC:4331]	1,240	22	1,878	46	1,515	31	–	–	–	–	2,560	58	2,574	59	2,072	53
GMNN	2,079	161	Geminin, DNA replication inhibitor [HGNC:17493]	969	67	1,392	121	1,224	72	569	43	884	63	1,972	146	1,974	153	1,552	128
HPGD	1,464	92	15-Hydroxyprostaglandin dehydrogenase [HGNC:5154]	575	38	1,000	74	962	62	618	39	567	37	1363	89	1,363	86	1,060	71
HSD17B10	1,211	107	Hydroxysteroid 17-beta dehydrogenase 10 [HGNC:4800]	516	48	827	84	858	81	548	47	506	48	1,134	99	1,135	95	866	81
HSP90AA1	666	56	Heat shock protein 90 alpha family class A member 1 [HGNC:5253]	295	25	453	39	419	38	–	–	288	25	640	50	637	50	502	40
HSPB1	876	76	Heat shock protein family B (small) member 1 [HGNC:5246]	461	40	522	45	600	53	304	26	454	41	837	72	837	71	558	47
HTR1A	412	60	5-Hydroxytryptamine receptor 1A [HGNC:5286]	186	34	279	49	232	55	122	37	180	34	401	58	400	58	315	50
IL1B	1,773	206	Interleukin 1 beta [HGNC:5992]	589	54	1,005	165	768	78	382	54	541	54	1,385	190	1,391	196	1,122	163
JAK2	895	80	Janus kinase 2 [HGNC:6192]	378	39	663	58	478	43	248	29	364	40	867	71	867	74	723	55
JUN	842	97	Jun proto-oncogene, AP-1 transcription factor subunit [HGNC:6204]	442	49	491	67	570	72	279	41	435	49	801	91	799	91	523	70
KCNH2	363	190	Potassium voltage-gated channel subfamily H member 2 [HGNC:6251]	174	119	212	136	250	161	128	104	173	119	331	183	329	184	228	139
KDM4A	1,607	192	Lysine demethylase 4A [HGNC:22978]	693	76	1,130	125	834	87	379	42	603	69	1,529	173	1,536	175	1,286	140
KDM4E	1,389	124	Lysine demethylase 4E [HGNC:37098]	543	43	999	88	880	70	547	40	530	42	1,320	109	1,321	110	1,057	95
MITF	3,626	132	Melanogenesis associated transcription factor [HGNC:7105]	1,170	42	2,238	91	1,562	51	858	42	1,083	37	2,832	116	2,871	120	2,460	93
MLLT3	14,566	101	MLLT3, super elongation complex subunit [HGNC:7136]	–	–	2,244	26	–	–	–	–	–	–	3,002	33	3,461	50	3,095	46
MPHOSPH8	506	52	M-Phase phosphoprotein 8 [HGNC:29810]	278	21	365	39	403	41	253	29	278	21	490	48	485	50	382	43
MYC	2,069	121	MYC proto-oncogene, bHLH transcription factor [HGNC:7553]	–	–	1,067	113	–	–	–	–	–	–	1,230	114	1,249	117	1,151	115
NFE2L2	2,850	226	Nuclear factor, erythroid 2 like 2 [HGNC:7782]	1,142	94	1,816	148	1,355	153	620	83	1,013	95	2,425	204	2,439	204	2,023	152
NFKB1	2,875	107	Nuclear factor kappa B subunit 1 [HGNC:7794]	730	23	1,608	91	1,237	37	814	29	716	22	1,978	100	2,000	101	1,742	94
NOD1	1,056	51	Nucleotide binding oligomerization domain containing 1 [HGNC:16390]	–	–	754	43	–	–	–	–	409	21	1,010	47	1,010	49	844	40
NOD2	2,578	59	Nucleotide binding oligomerization domain containing 2 [HGNC:5331]	952	23	1,629	53	1,124	21	–	–	836	23	2,152	57	2,165	59	1,837	48
NPSR1	1,007	55	Neuropeptide S receptor 1 [HGNC:23631]	–	–	712	44	554	21	–	–	–	–	959	52	956	54	777	50

(Continued)

TABLE 2 | Continued

Gene name	Total inactive	Total active	Description	A375		A549		HA1E		HCC515		HT29		MCF7		PC3		VCAP	
				Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive	Active
NR3C1	925	54	Nuclear receptor subfamily 3 group C member 1 [HGNC:7978]	–	–	636	38	692	34	451	25	–	–	896	54	890	54	659	47
NR5A1	419	69	Nuclear receptor subfamily 5 group A member 1 [HGNC:7983]	190	22	285	50	239	29	–	–	–	–	408	60	407	64	322	50
OPRK1	1,122	51	Opioid receptor kappa 1 [HGNC:8154]	455	29	805	41	564	34	–	–	428	29	1,068	51	1,070	50	895	41
PIP4K2A	1,898	88	Phosphatidylinositol-5-phosphate 4-kinase type 2 alpha [HGNC:8997]	609	31	1,082	66	812	34	–	–	566	23	1,486	84	1,501	83	1,206	72
PLA2G7	1,907	57	Phospholipase A2 group VII [HGNC:9040]	828	27	1,367	37	995	37	–	–	737	28	1,825	52	1,826	51	1,534	37
PLK1	1,935	108	Polo like kinase 1 [HGNC:9077]	631	45	1,125	82	836	39	431	23	590	36	1,531	100	1,542	102	1,241	88
POLB	1,166	53	DNA polymerase beta [HGNC:9174]	480	22	850	37	646	30	–	–	462	21	1,113	50	1,113	50	932	39
POLH	2,202	70	DNA polymerase eta [HGNC:9181]	858	26	1,342	44	1,015	37	–	–	749	23	1,827	61	1,838	64	1,532	48
POLI	1,726	79	DNA polymerase iota [HGNC:9182]	775	29	1,210	52	945	43	448	24	689	27	1,629	71	1,637	72	1,359	55
POLK	2,895	79	DNA polymerase kappa [HGNC:9183]	1,248	30	2,048	49	1,799	43	986	27	1,154	26	2,713	69	2,722	70	2,225	54
PRMT1	2,886	80	Protein arginine methyltransferase 1 [HGNC:5187]	1,114	28	1,704	59	1,073	23	–	–	–	–	2,394	74	2,415	74	2,081	68
RAD52	14,593	132	RAD52 homolog, DNA repair protein [HGNC:9824]	–	–	2,291	25	–	–	–	–	–	–	3,043	40	3,496	54	3,121	49
SIRT5	14,103	141	Sirtuin 5 [HGNC:14933]	–	–	2,086	30	–	–	–	–	–	–	2,769	40	3,211	44	2,844	42
SLC6A3	1,006	94	Solute carrier family 6 member 3 [HGNC:11049]	461	49	773	71	584	72	252	54	453	49	976	91	973	90	823	73
SMN2	1,633	53	Survival of motor neuron 2, centromeric [HGNC:11118]	–	–	1,136	44	1,059	28	–	–	–	–	1,520	49	1,521	49	1,209	45
STK33	3,358	423	Serine/threonine kinase 33 [HGNC:14568]	1,127	101	2,077	304	1,458	163	776	131	1,034	102	2,660	329	2,663	360	2,268	328
TARDBP	1,802	60	TAR DNA binding protein [HGNC:11571]	–	–	1,044	50	748	25	–	–	–	–	1,409	58	1,418	59	1,156	52
TNFRSF10B	2,429	80	TNF receptor superfamily member 10b [HGNC:11905]	–	–	1,510	66	1,056	26	–	–	786	27	2,008	73	2,021	75	1,718	58
TP53	2,310	198	Tumor protein p53 [HGNC:11998]	974	97	1,554	137	1,380	130	737	84	891	98	2,172	179	2,174	181	1,714	137
TSHR	2,259	70	Thyroid stimulating hormone receptor [HGNC:12373]	968	25	1,579	58	1,317	43	727	35	–	–	2,133	68	2,131	67	1,739	64
TUBB	697	51	Tubulin beta class I [HGNC:20778]	–	–	503	32	373	32	–	–	–	–	692	48	693	49	563	32
USP1	2,356	64	Ubiquitin specific peptidase 1 [HGNC:12607]	877	30	1,425	46	1,260	45	697	29	833	30	1,972	55	1,985	58	1,557	44
VDR	2,696	140	Vitamin D receptor [HGNC:12679]	1,161	44	1,901	101	1,673	80	915	50	1,076	43	2,530	127	2,536	128	2,059	107
YES1	138	101	YES proto-oncogene 1, Src family tyrosine kinase [HGNC:12841]	106	89	99	86	117	89	77	76	106	89	132	97	131	99	90	67

each subset: a first model used the 978-landmark GES as input (referred as GES models), and the second one used the Morgan fingerprints of corresponding compounds to fairly compare model performances (referred as Morgan FP models). Models were trained using random forest classifiers (Breiman, 2001).

The training was performed using a 4-fold cross-validation on training set to tune the maximum depth of tree, before assessing prediction performances on the test set. The number of trees per model was set to 200. Models were built in Python 2.7 using the sklearn package: to account for unbalanced dataset, the "class_weight" parameter was set to "balanced_subsample," to increase the weight of the under-represented class samples when training the trees (Pedregosa et al., 2011). A first step of feature selection was performed using an initial random forest classifier, computing the feature importance (Breiman, 2001). Sum of importance of all feature was 1, with each feature importance between 0 (non-important) and 1 (important). This step was performed 5 times, feature importance was averaged by feature, and only the 20 most important features were selected to be loaded into a final random forest model. This whole modeling pipeline, from train-test split to final model was performed 10 times per task, to account for variable performances depending on the dataset split.

Models were evaluated by counting the numbers of true positive (TP), true negative (TN), false positive (FP) and false negative (FN). These parameters were combined in the following metrics in order to compare model performances:

- $Sensitivity = TP/(TP + FN)$
- $Specificity = TN/(TN + FP)$
- Balanced accuracy (BA) : $BA = (Sensitivity + Specificity)/2$
- Matthews correlation coefficient (MCC) :

$$\mathrm{MCC} = \frac{\mathrm{TP}^{*}\mathrm{TN} - \mathrm{FP}^{*}\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\,(\mathrm{TP}+\mathrm{FN})\,(\mathrm{TN}+\mathrm{FP})\,(\mathrm{TN}+\mathrm{FN})}}$$

Balanced accuracy allows for a fair evaluation of model performances when using unbalanced datasets, by averaging accuracy for each class (here active and inactives).

Quadrant Plots

Between each possible pair of compounds active on the same target and in each cell line, Dice distance between Morgan fingerprints, and cosine distance between GESs in given cell line were computed. These 2 distances were plotted in a 2D plot (referred as distance plot), Dice distance on X-axis and cosine distance on Y-axis. These plots were theoretically split in 4 quadrants.

Quadrant I in the top-right corner contains active compound pairs having different structures (Morgan fingerprints Dice distance > 0.5) and presenting different GESs (cosine distance > 0.5); quadrant II in top-left corner contains active compound pairs having similar structures (Morgan fingerprints Dice distance < 0.5) and presenting different GESs (cosine distance > 0.5); quadrant III in bottom-left corner contains active compound pairs having similar structures (Morgan fingerprints Dice distance < 0.5) and presenting similar GESs (cosine distance < 0.5); quadrant IV in bottom-right corner contains active compound pairs having different structures (Morgan fingerprints Dice distance > 0.5) and presenting similar GESs (cosine distance < 0.5). Number of active compound pairs in each quadrant were counted for each distance plot. Similar calculations were made using not only active compounds, but all compounds having an annotation (active or inactive) for considered target and profiled in the same cell line.

RESULTS

In the present work, we investigated the link between compound structure information ($n = 9{,}035$) and their corresponding induced biological responses captured by GESs ($n = 39{,}544$) in human tumor cell lines and evaluated the potential of machine learning approaches to infer about molecular targets involved in the compound bioactivity. In addition, we compared these machine learning models using GESs with counterpart models using Morgan fingerprints.

Exploration of Chemical and Biological Spaces

As a first step, to observe the diversity of the 9,035 compounds profiled in the 10 μM/24 h L1000 signature dataset, the corresponding chemical space was visualized. **Figure 2A** is a 2-dimensional t-SNE representation of the chemical space, illustrating the variability in terms of Morgan fingerprints. The 9,035 compounds form a broad chemical space, with a mean Dice distance between compound pairs of 0.81 (ChEMBL has a mean pairwise Dice distance of 0.82). The center of the chemical space is mostly composed by small molecules having on average a molecular weight lower than 500 Da whereas the outer part is populated by clusters of compounds with higher molecular weights (>500 Da). Overall, we were able to retrieve, in the public domain, at least one activity information for 7,837 compounds, from which 4,872 were active in at least one target. The majority of those compounds were found active in a low number of targets, on average 6 per compound, with a median of 2. Not surprisingly, a set of 23 kinase inhibitors were found to be active in more than 100 targets.

Figure 2B shows a t-SNE plot created using all GESs induced by the 9,035 compounds in the different cell lines to examine the complete biological space. This t-SNE is color coded by the different cell lines used to generate the gene signatures. Each cell line is represented by a set of 4 to 5 main clusters of GESs differing in size and some overlap of the cluster indicates similar GESs derived from different cell lines. In order to better appreciated the differences and communalities in GESs obtained with the selected compounds, t-SNE plots were created highlighting the clusters derived for cell lines originating from the same tumor type namely prostate tumor (VCAP and PC3 in **Figure 2C**) and lung tumor (A549 and HCC515 in **Figure 2D**). GESs derived from cell lines coming

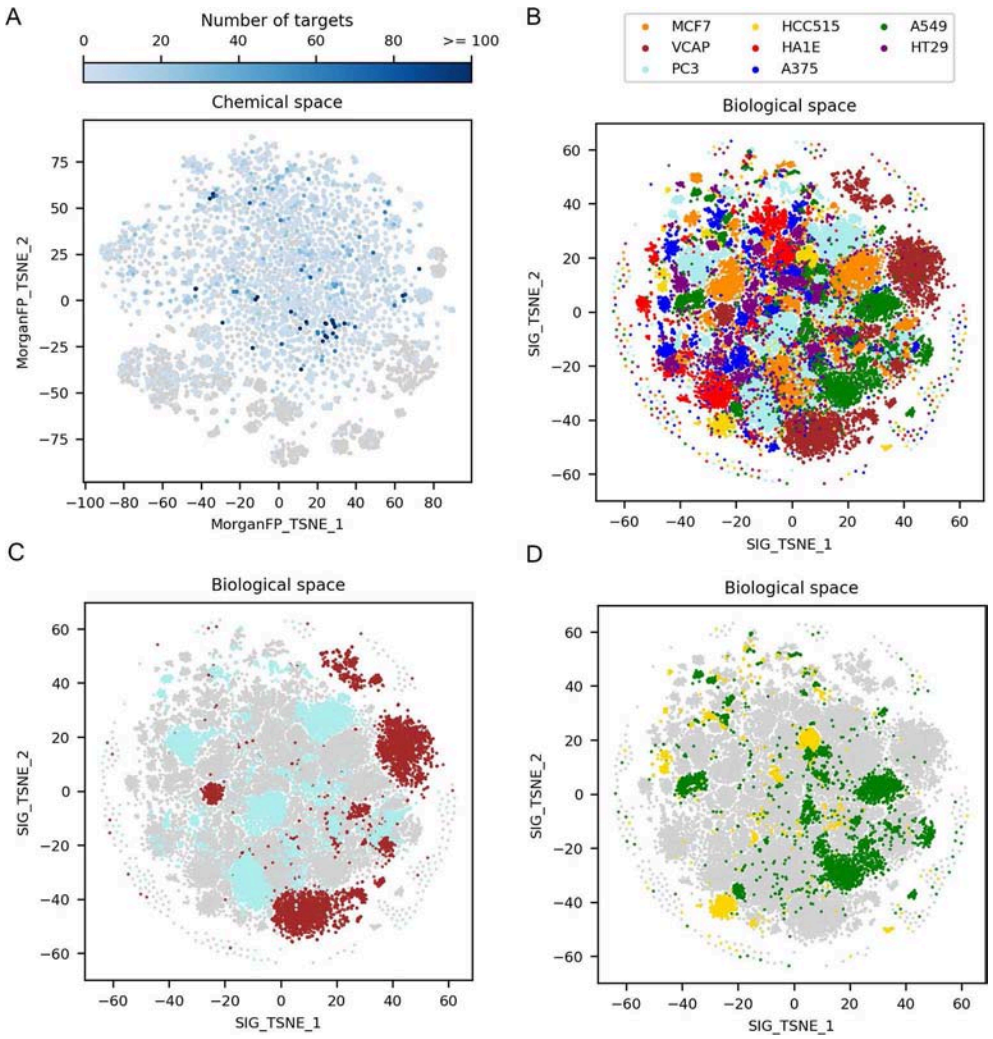


FIGURE 2 | Exploration of the 2D chemical space, along with the corresponding 2D biological space formed by all GES. **(A)** t-SNE on Morgan fingerprints from the 9,035 compounds in working dataset, representing the chemical space. Points corresponding to compounds for which there is no known target are represented by gray points (n = 4,163). Points corresponding to compounds for which there is at least one known target are in blue (n = 4,872), with darker blue depending on the increasing number of targets. **(B)** t-SNE on all GESs in the working dataset, representing the biological (transcriptomic response) space. Points corresponding to GESs are colored by cell line. **(C)** Biological space highlighting only PC3 and VCAP signatures, 2 cell lines originating from prostate cancer. **(D)** Biological space highlighting only A549 and HCC515 signatures, 2 cell lines originating from lung cancer.

from the same tissue present very little overlap as can be observed in **Figures 2C,D**.

These results illustrate the variability in the cellular modifications occurring during carcinogenesis (Hanahan and Weinberg, 2011) and show that each cell line represent a distinct biological space even if the cell lines are derived from the same tissue type. Interestingly, when comparing, for a set of compounds showing GESs in a single cluster in VCAP, GESs of these compounds in PC3 are spread across various clusters from the PC3 biological space (data not shown). This shows that each cell line explores different biological responses to compounds.

After having described the global variability of GESs in the different cell lines, we explored the chemical and biological spaces corresponding to active and inactive compounds on different targets. Since each compound-induced GES obtained in each cell line was shown to represent a unique biological space, t-SNE plots were computed per cell line in order to further explore the link between the different biological spaces and the corresponding chemical ones. For this, we decided to focus on three cell lines derived from different tissues and among the largest GES dataset generated that is to say A549 (lung cancer), MCF7 (breast cancer) and PC3 (prostate cancer). In addition, we selected 3 representative molecular targets showing different chemical and biological space profiles: compounds active on the glucocorticoid receptor (NR3C1) have similar structures, and similar GESs in some cell lines (**Figures 3A–D**); tubulin beta I (TUBB) actives have more diverse structures but show similar GESs in each

cell line considered in this work (**Figures 4A–D**); and dopamine receptor D1 (DRD1) actives have diverse structure and GESs in every used cell line (**Figures 5A–D**).

NR3C1 actives compounds are mostly grouped together in the chemical space, as shown in **Figure 3A** ($n = 54$; mean Dice distance = 0.67). Most of NR3C1 active GESs are also grouped in a cluster in the A549 biological space ($n = 38$; mean cosine distance = 0.76), visible in **Figure 3B**, and remaining NR3C1 active GESs are spread across this biological space. Following the similarity principle, we could conjecture that other GESs that are close to this cluster are responses from other NR3C1 actives, especially in the PC3 biological space where the cluster contains compounds known to be active. The same phenomenon can be observed in the biological space of PC3 (**Figure 3D**), HCC515, HA1E and VCAP (data not shown). Surprisingly, the GES clusters populated by numerous known NR3C1 actives in the biological spaces of A549 and PC3 also contain some known inactive compounds. In the biological spaces of MCF7, A375 and HT29, there is no such clustering, like shown in **Figure 3C** representing the MCF7 biological space ($n = 54$; mean cosine distance = 0.92). Overall, these results show that compounds that are known to be active on the NR3C1 target can show a similar response in only certain cellular contexts.

TUBB actives compounds are spread in the chemical space (represented in **Figure 4A**), indicating that they have diverse chemical structures ($n = 51$; mean Dice distance = 0.76). Most importantly, in each cell lines used in this work, GESs induced by TUBB actives compounds were similar (as illustrated in **Figures 4B–D**), with a rather low mean cosine distance between active compounds ranging between 0.61 and 0.75 depending on the cell line dataset. Moreover, GESs of TUBB actives tend to be similar across all cell lines used in this work (highlighted in **Supplementary Figure 1**). This conserved pattern in GESs induced by tubulin binding compounds likely illustrate certainly the ubiquitous role of tubulin polymerization of the eukaryotic cytoskeleton (Chaaban and Brouhard, 2017).

Finally, DRD1 actives compounds, that are represented in the chemical space t-SNE, have diverse chemical structures ($n = 99$; mean Dice distance = 0.81), associated with diverse GESs for the 3 cell lines presented (mean cosine distance between 0.88 and 0.92 depending on the cell line), as illustrated in **Figures 5A–D**. Since GESs of active compounds in any cell lines are not similar, nor their chemical structures, actives cannot be easily discriminated from inactives using these two types of descriptors, as opposed to what was observed with NR3C1 actives that have similar structures, or TUBB actives having similar GESs in every cell line used in this study.

Model Performances: GES Vs. Morgan Fingerprints

Based on the observed GES similarity of compounds sharing target activity in appropriate cellular contexts, we tested building predictive machine learning models using GESs as descriptors and compare their performances with the ones of the models using Morgan fingerprints.

In order to avoid building models with too unbalanced datasets and to ensure a minimum of active compounds when testing model performances, we first pre-selected targets having at least 50 active compounds in the total dataset (representing between 1 and 63% of active compounds per target). We obtained one dataset per cell line—target combination (restricted to compounds having signatures in the considered cell line, as shown in **Figure 1**) and carried out a second selection by performing prediction tasks using datasets containing at least 20 active compounds for the considered target (representing between 1 and 69% of active compounds per dataset). For each selected cell line—target dataset, one model using GESs (referred as GES model) was computed. In order to perform a fair comparison per task, one counterpart model using corresponding compound Morgan fingerprints (referred as Morgan FP model) was built using the same set of compounds as the one used in the corresponding GES models. Performances of models were evaluated using the balanced accuracy (BA) metric on a test set, to account for class imbalance in datasets. In total, 990 models were built for a total of 69 different targets. BAs of all built models are presented in **Table 3**. MCC of all built models are presented in **Supplementary Table 1**.

Overall, GES model performances appeared to be variable depending on the predicted target and on the cell line that was used to generate the GESs, with a BA ranging from 0.49 to 0.88. Counterpart models trained with Morgan fingerprints also had variable performances, with a BA ranging from 0.50 to 0.98. On average, Morgan FP models (mean BA = 0.65) yielded better performances for the target activity prediction than their counterpart GES models (mean BA = 0.58). On the 495 cell line—target combinations, BA of GES models was higher than BA of counterpart Morgan FP models for 124 combinations (25%).

On the 990 models, 208 models reached a BA higher than 0.7 (21%) for 40 targets (59 GES models for 18 targets; 138 Morgan FP models for 28 targets), and 33 models reached a BA higher than 0.8 (3%) for 10 targets (10 GES models for 4 targets; 21 Morgan FP models for 7 targets). For all 138 Morgan FP models reaching BA higher than 0.7, BA was superior to counterpart GES models, and for the 59 GES models reaching BA higher than 0.7, only 6 had counterpart Morgan FP model with higher BA.

For NR3C1 activity prediction, Morgan FP models yielded a BA between 0.93 and 0.98 depending on the cell line dataset. It is not surprising considering that that NR3C1 actives have similar structure as shown in **Figure 3A**. On the GES models, a BA of 0.77 was reached using the A549 signature dataset, correlating to similar GESs that were observed in the A549 biological space (**Figure 3B**), whereas a BA of 0.6 was obtained using the MCF7 signature dataset (no GES cluster in MCF7 biological space, shown in **Figure 3C**). A549 and MCF7 signature model performances cannot be fairly compared because they were built using different sets of compounds. In fact, performances of different GES models cannot be compared across cell lines nor across targets, performances can only be compared to observed similarity between active compounds in either chemical and biological space plots for a given target.

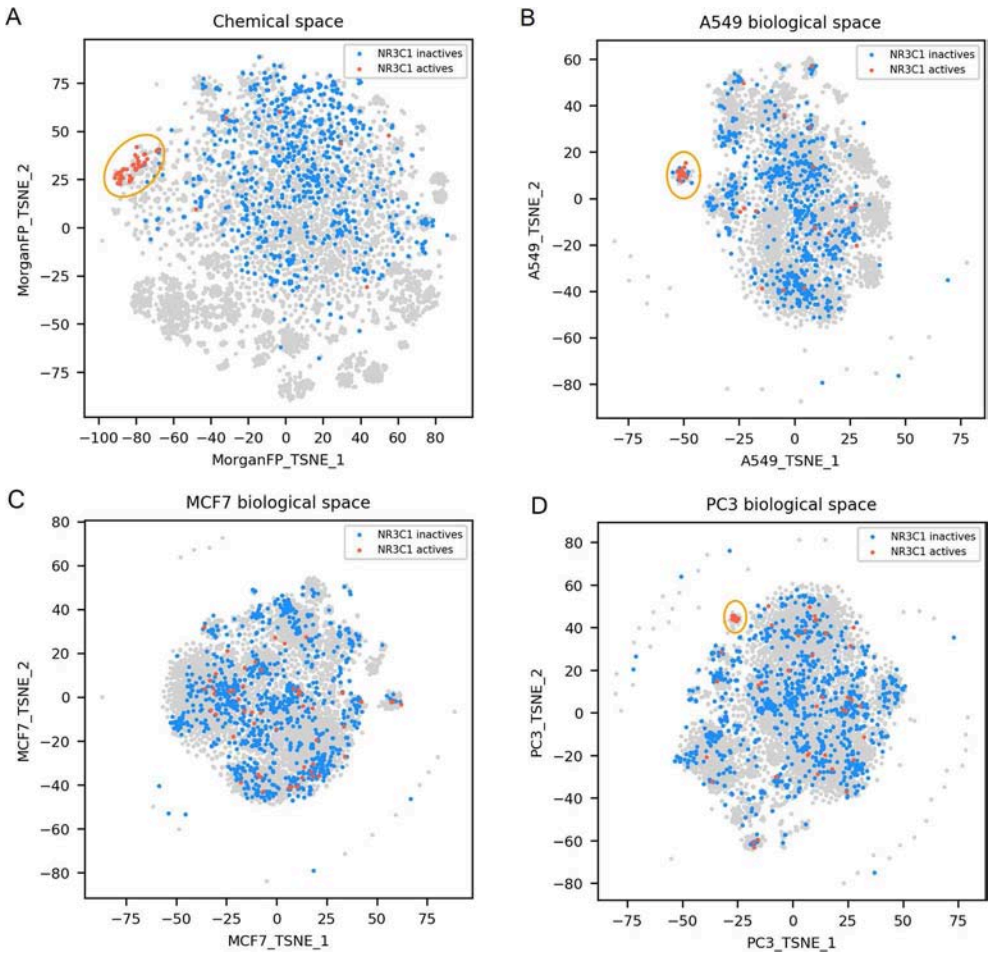


FIGURE 3 | NR3C1 active and inactives compounds in the chemical space and the different biological spaces formed by GES produced in a single cell line. **(A)** Chemical space; **(B)** t-SNE on all A549 signatures (A549 biological space); **(C)** t-SNE on all MCF7 signatures (MCF7 biological space); **(D)** t-SNE on all PC3 signatures (PC3 biological space). Points corresponding to NR3C1 actives are red ($n = 54$), NR3C1 inactives ($n = 925$) are blue, gray points have no available label concerning NR3C1 activity. Orange circles point out clustering of active compounds.

For TUBB activity prediction, GES models yielded BA between 0.81 and 0.88 depending on the cell line dataset, which was among the 10 best GES models. Interestingly, their counterpart Morgan FP models were not significantly underperforming (BA ranging from 0.78 to 0.82). Even though the TUBB active structures are diverse, the models still managed to identify structural fragments that could produce such predictive performance.

For DRD1 activity prediction, Morgan FP models yielded BA between 0.68 and 0.74 depending on the cell line dataset, and were always better than their counterpart GES model, with a BA ranging from 0.58 to 0.64.

Overall, we conclude that it was possible to build GES models with acceptable performances, performing similarly or better than their counterpart Morgan FP models in 25% of the target prediction tasks. Moreover, we see an important advantage in the GES models: they are theoretically performing independently of the chemical space considered, allowing target identification of new compounds even if their corresponding structural diversity is not represented in the training set.

Rationalizing Model Performances Using Distance Plots

To further describe and understand the reasons for the differences in performances between GES model and Morgan FP model, for every dataset used in each cell line—target combinations, Morgan fingerprints Dice distance was plotted against GES cosine distance between each pair of active compounds in the given dataset.

Generated distance plots were split in 4 quadrants separated by a 0.5 threshold for Dice distance (dotted vertical line) and a 0.5 threshold for cosine distance (dotted horizontal line). Data points in top right (Quadrant I) represent pairs of active compounds showing diverse structures and different GESs in the considered cell line and contains most of compound pairs (average of 95.1%). Data points in top left (Quadrant II) represent pairs of active

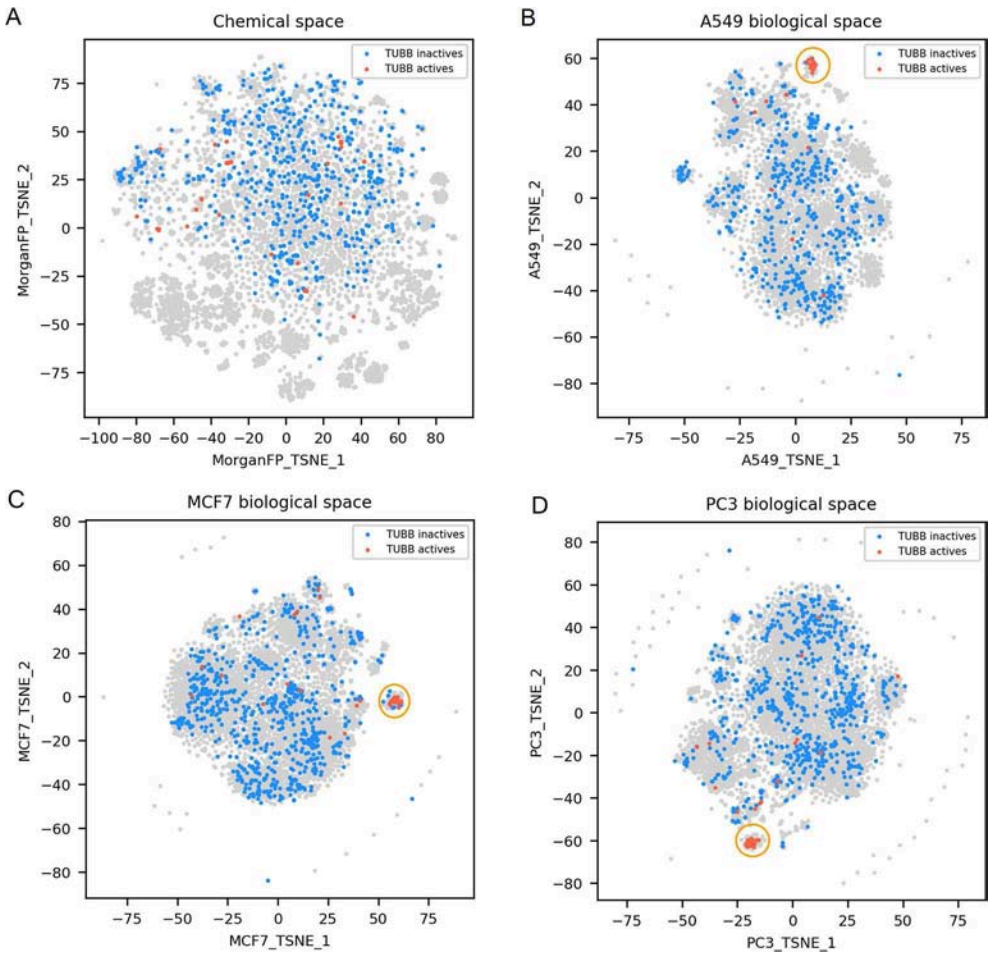


FIGURE 4 | TUBB active and inactives compounds in the chemical space and the different biological spaces formed by GES produced in a single cell line. **(A)** Chemical space; **(B)** A549 biological space; **(C)** MCF7 biological space; **(D)** PC3 biological space. Points corresponding to TUBB actives ($n = 51$) are red, TUBB inactives ($n = 697$) are blue, gray points have no available label concerning TUBB activity. Orange circles point out clustering of active compounds.

compounds showing similar structures while showing diverse GESs in the considered cell line (average of 1.9%). Data points in bottom left (Quadrant III) represents pairs of active compounds showing similar structures and similar GESs in the considered cell line and contains least compound pairs (average of 0.5%). Data points in bottom right (Quadrant IV) represents pairs of active compounds showing similar GESs while having different structures (average of 2.5%). Intuitively, we think that sample similarity within the same class (here: actives) is a good indicator to know if a machine learning model will be able to properly predict samples from this class.

Overall, the mean percentage of compound pairs (active and inactives) were 99.3, 0.3, 0.01 and 0.4% for quadrants I, II, III and IV respectively. Based on this dataset, compounds active toward a molecular target have on average more similar structures and GESs than the totality of the compounds.

We expected to reach good Morgan FP model prediction for combinations having a high proportion of points in quadrants II and III (similar structures), and good GES model prediction for combinations having a high proportion of points in quadrants III and IV (similar GESs). We evaluated the use of distance plots on the three targets and three cell lines used in previous space plots (**Figures 3**, **5**). Similar work was performed using not only active compounds, but all compounds having at least one annotation for each of the three previously described targets, shown in **Supplementary Figure 2**.

For NR3C1 distance plots, there are approximately 10% of compound pairs in quadrants II and III of the 3 plots (**Figures 6A–C**), coherent with good Morgan FP model performances. However, depending on which cell line the GESs were generated from, there were different proportions of compound pairs in quadrants III and IV: there are 20% of pairs for A549, and only 1% of pairs in MCF7. This is in agreement with what was observed in model performances: performance of GES models using the A549 dataset (BA = 0.77) was much better than performances using MCF7 dataset (BA = 0.60). Surprisingly, prediction using GESs from the PC3 dataset showed good performances (BA = 0.73), even though the proportion

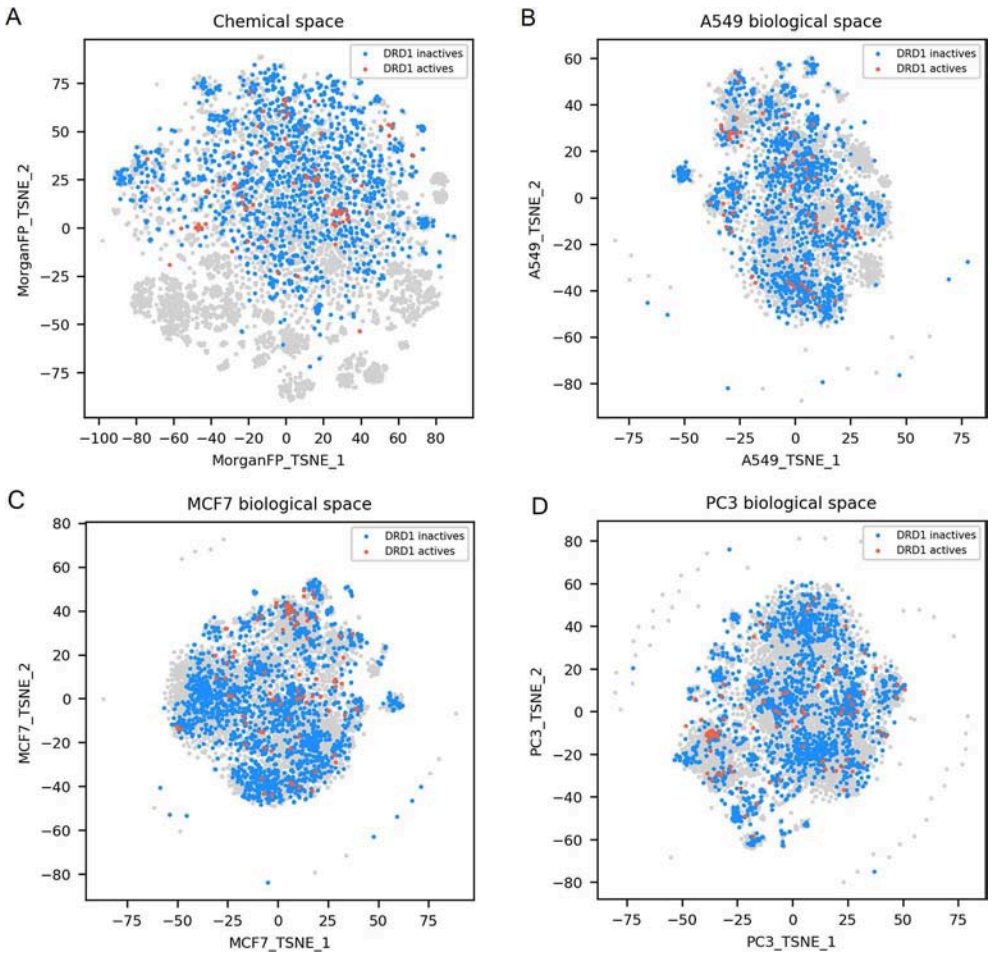


FIGURE 5 | DRD1 active and inactives compounds in the chemical space and the different biological spaces formed by GES produced in a single cell line. **(A)** Chemical space; **(B)** A549 biological space; **(C)** MCF7 biological space; **(D)** PC3 biological space. Points corresponding to DRD1 actives ($n = 99$) are red, DRD1 inactives ($n = 1843$) are blue, gray points have no available label concerning DRD1 activity.

of active compound pairs in quadrants III and IV was around 1% (similar to the proportion observed for the MCF7 dataset that showed worse performances). This suggests that the GES model built with PC3 was able to capture a subset of genes to discriminate active compounds from inactives, even with active compounds showing different GESs.

For TUBB distance plots (**Figures 6D–F**), between 7 and 10% of compound pairs was observed in quadrants II and III, matching the good Morgan FP model performances with the 3 cell line datasets (BA ranging from 0.80 to 0.82). Most importantly, there are between 26 and 40% of compound pairs in quadrants III and IV, echoing the better performances of the GES models in these cell lines (BA ranging from 0.81 to 0.88).

For DRD1 distance plots (**Figures 6G–I**), 98% of compound pairs are in quadrant I, leaving low number of active compound pairs in the other quadrants (with similar Morgan fingerprints and/or GESs). This is coherent with the average performances of GES (BA ranging from 0.68 to 0.74) and Morgan FP models (BA ranging from 0.58 to 0.64) built for this target.

Among the 50 best GES models, the mean percentage of active pairs in quadrants III and IV was 5.2% (vs. 2.3% in quadrants II and III). For the 50 best Morgan FP models, the mean percentage of active pairs in quadrants II and III was 4.0% (vs. 2.2% in quadrants III and IV). This suggests a positive relationship between sample similarity between active compounds using a given set of descriptors for active compounds and performances of models using these descriptors.

In the current work, GESs were shown to be effective descriptors to predict compound activity toward molecular targets. In 25% of target prediction tasks, GES models outperformed their counterpart Morgan FP models, especially when using GES produced in a cell line exhibiting similar GESs between compounds active toward the target of interest. Such GES models performs independently of the structural diversity of compounds that were used to produce GESs, offering a great opportunity to escape the classical chemical space limitations associated with QSAR models. In addition, t-SNE plots, along with 2D distance plots, can give insights to assess the predictive

TABLE 3 | Mean BAs of models (mean per condition).

Target class	Target name	Cell line	A375		A549		HA1E		HCC515		HT29		MCF7		PC3		VCAP	
		Target/ Descriptor	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP
Enzyme	15-Hydroxyprostaglandin dehydrogenase	HPGD	0.5	0.65	0.51	0.7	0.52	0.67	0.5	0.72	0.51	0.68	0.5	0.71	0.52	0.67	0.52	0.64
	Arachidonate 15-lipoxygenase	ALOX15	0.54	0.74	0.57	0.75	0.54	0.67	0.5	0.63	0.5	0.73	0.56	0.74	0.65	0.75	0.53	0.75
	ATP binding cassette B1	ABCB1	0.54	0.62	0.51	0.61	0.5	0.54	0.5	0.5	0.51	0.5	0.56	0.62	0.59	0.63	0.51	0.63
	BRCA1, dna repair associated	BRCA1	0.74	0.69	0.69	0.62	0.67	0.57	0.6	0.52	0.71	0.6	0.78	0.64	0.75	0.64	0.67	0.65
	Cytochrome P450 1A2	CYP1A2	0.54	0.6	0.56	0.6	0.57	0.61	0.59	0.58	0.5	0.59	0.59	0.63	0.6	0.63	0.56	0.61
	Cytochrome P450 2C19	CYP2C19	0.52	0.57	0.55	0.59	0.5	0.55	0.52	0.55	0.51	0.57	0.56	0.58	0.56	0.59	0.53	0.58
	Cytochrome P450 2C9	CYP2C9	0.55	0.57	0.52	0.61	0.52	0.6	0.51	0.6	0.53	0.58	0.55	0.58	0.6	0.58	0.54	0.6
	Cytochrome P450 3A4	CYP3A4	0.51	0.55	0.52	0.59	0.52	0.58	0.49	0.6	0.51	0.54	0.52	0.6	0.54	0.6	0.53	0.57
	DNA polymerase beta	POLB	0.5	0.66	0.5	0.69	0.51	0.82	–	–	0.5	0.69	0.5	0.69	0.53	0.73	0.5	0.69
	DNA polymerase eta	POLH	0.55	0.71	0.51	0.74	0.66	0.82	–	–	0.5	0.76	0.51	0.73	0.55	0.72	0.52	0.69
	DNA polymerase iota	POLI	0.5	0.7	0.5	0.71	0.54	0.71	0.5	0.6	0.51	0.65	0.51	0.72	0.51	0.73	0.5	0.68
	DNA polymerase kappa	POLK	0.58	0.77	0.51	0.78	0.56	0.83	0.52	0.79	0.5	0.76	0.54	0.81	0.56	0.83	0.5	0.83
	Flap structure-specific endonuclease 1	FEN1	0.5	0.72	0.5	0.68	0.5	0.77	–	–	0.5	0.74	0.5	0.77	0.51	0.75	0.5	0.7
	Glutaminase	GLS	0.5	0.64	0.51	0.64	0.53	0.58	–	–	–	–	0.51	0.74	0.51	0.69	0.51	0.7
	Growth factor, augmenter of liver regeneration	GFER	0.51	0.72	0.5	0.72	0.59	0.74	0.5	0.71	0.5	0.72	0.52	0.72	0.56	0.71	0.5	0.7
	Hydroxysteroid 17-beta dehydrogenase 10	HSD17B10	0.51	0.57	0.54	0.62	0.55	0.63	0.51	0.63	0.49	0.58	0.54	0.65	0.53	0.64	0.5	0.62
	Janus kinase 2	JAK2	0.71	0.57	0.73	0.63	0.69	0.56	0.63	0.56	0.68	0.56	0.8	0.63	0.73	0.59	0.65	0.58
	MDM2 proto-oncogene	MDM2	0.77	0.69	0.71	0.61	0.7	0.62	0.76	0.53	0.67	0.59	0.83	0.65	0.81	0.64	0.76	0.66
	Phosphatidylinositol-5-phosphate 4-kinase type 2 alpha	PIP4K2A	0.51	0.69	0.54	0.75	0.53	0.66	–	–	0.5	0.58	0.53	0.76	0.5	0.76	0.52	0.74
	Phospholipase A2 group VII	PLA2G7	0.5	0.71	0.51	0.65	0.57	0.72	–	–	0.5	0.69	0.54	0.69	0.57	0.71	0.51	0.69
	Polo like kinase 1	PLK1	0.5	0.62	0.52	0.64	0.54	0.52	0.49	0.56	0.5	0.52	0.52	0.68	0.52	0.67	0.49	0.68
	Serine/threonine kinase 33	STK33	0.78	0.58	0.68	0.62	0.78	0.59	0.74	0.6	0.7	0.56	0.72	0.61	0.71	0.64	0.66	0.65
	Ubiquitin specific peptidase 1	USP1	0.5	0.54	0.51	0.58	0.51	0.53	0.51	0.53	0.5	0.55	0.52	0.57	0.57	0.59	0.5	0.57
	YES proto-oncogene 1, Src family tyrosine kinase	YES1	0.7	0.72	0.67	0.7	0.71	0.72	0.63	0.69	0.66	0.71	0.66	0.75	0.7	0.72	0.54	0.68
Epigenetic regulator	Bromodomain adjacent to zinc finger domain 2B	BAZ2B	0.6	0.66	0.54	0.68	0.63	0.69	0.52	0.6	0.5	0.65	0.55	0.67	0.6	0.66	0.53	0.68
	Chromobox 1	CBX1	0.56	0.57	0.55	0.62	0.6	0.6	0.57	0.61	0.54	0.55	0.61	0.6	0.57	0.62	0.55	0.61
	Lysine demethylase 4A	KDM4A	0.64	0.62	0.56	0.67	0.65	0.7	0.57	0.63	0.6	0.65	0.58	0.68	0.6	0.65	0.53	0.63
	Lysine demethylase 4E	KDM4E	0.59	0.76	0.52	0.75	0.61	0.75	0.53	0.75	0.5	0.72	0.53	0.73	0.57	0.74	0.52	0.72
	M-phase phosphoprotein 8	MPHOSPH8	0.52	0.53	0.52	0.64	0.54	0.64	0.51	0.65	0.5	0.54	0.63	0.63	0.54	0.64	0.51	0.65
	Protein arginine methyltransferase 1	PRMT1	0.5	0.65	0.51	0.71	0.5	0.54	–	–	–	–	0.52	0.7	0.51	0.71	0.51	0.72
	Sirtuin 5	SIRT5	–	–	0.51	0.65	–	–	–	–	–	–	0.51	0.62	0.5	0.61	0.5	0.6

(Continued)

TABLE 3 | Continued

Target class	Target name	Cell line	A375		A549		HA1E		HCC515		HT29		MCF7		PC3		VCAP	
		Target/ Descriptor	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP
	Survival of motor neuron 2, centromeric	SMN2	–	–	0.5	0.62	0.51	0.56	–	–	–	–	0.52	0.63	0.52	0.6	0.5	0.62
Ion channel	Potassium voltage-gated channel H2	KCNH2	0.72	0.82	0.64	0.8	0.66	0.79	0.65	0.76	0.74	0.8	0.65	0.84	0.65	0.8	0.67	0.8
Membrane receptor	5-hydroxytryptamine receptor 1A	HTR1A	0.5	0.72	0.51	0.75	0.54	0.75	0.53	0.75	0.51	0.7	0.55	0.77	0.52	0.76	0.55	0.79
	Cholinergic receptor muscarinic 1	CHRM1	0.6	0.69	0.64	0.69	0.66	0.71	0.7	0.69	0.67	0.62	0.59	0.71	0.6	0.73	0.65	0.72
	Cholinergic receptor muscarinic 4	CHRM4	0.63	0.71	0.67	0.75	0.64	0.72	0.7	0.68	0.63	0.66	0.66	0.7	0.62	0.73	0.62	0.72
	Cholinergic receptor muscarinic 5	CHRM5	0.6	0.69	0.62	0.76	0.64	0.72	0.69	0.72	0.63	0.68	0.64	0.73	0.59	0.71	0.61	0.68
	Dopamine receptor D1	DRD1	0.64	0.68	0.6	0.73	0.63	0.7	0.58	0.72	0.62	0.69	0.62	0.74	0.61	0.72	0.59	0.74
	Dopamine receptor D2	DRD2	0.61	0.74	0.61	0.79	0.61	0.79	0.6	0.79	0.61	0.73	0.65	0.79	0.62	0.79	0.63	0.8
	Dopamine receptor D3	DRD3	0.6	0.66	0.58	0.72	0.56	0.71	0.58	0.71	0.58	0.66	0.63	0.72	0.57	0.72	0.58	0.73
	Neuropeptide S receptor 1	NPSR1	–	–	0.59	0.64	0.5	0.55	–	–	–	–	0.63	0.66	0.57	0.66	0.58	0.63
	Opioid receptor kappa 1	OPRK1	0.5	0.61	0.52	0.65	0.5	0.61	–	–	0.54	0.63	0.57	0.65	0.53	0.64	0.55	0.68
	Thyroid stimulating hormone receptor	TSHR	0.51	0.64	0.5	0.56	0.56	0.52	0.57	0.5	–	–	0.56	0.61	0.56	0.6	0.55	0.61
	TNF receptor superfamily member 10b	TNFRSF10B	–	–	0.69	0.61	0.56	0.52	–	–	0.7	0.56	0.71	0.61	0.78	0.58	0.65	0.56
Other cytosolic protein	Heat shock protein 90 alpha A1	HSP90AA1	0.53	0.65	0.5	0.67	0.59	0.73	–	–	0.53	0.67	0.55	0.67	0.59	0.65	0.51	0.64
	Heat shock protein family B1	HSPB1	0.58	0.58	0.54	0.53	0.63	0.58	0.54	0.51	0.5	0.55	0.66	0.59	0.66	0.61	0.55	0.57
Secreted protein	Interleukin 1 beta	IL1B	0.62	0.55	0.65	0.6	0.65	0.55	0.65	0.55	0.66	0.57	0.68	0.63	0.69	0.62	0.63	0.62
Structural protein	Tubulin beta class I	TUBB	–	–	0.81	0.8	0.82	0.78	–	–	–	–	0.88	0.82	0.84	0.8	0.82	0.8
Transcription factor	Androgen receptor	AR	0.51	0.63	0.58	0.75	0.55	0.71	0.53	0.75	0.51	0.62	0.61	0.77	0.55	0.74	0.67	0.76
	Jun proto-oncogene, AP-1 transcription factor subunit	JUN	0.6	0.69	0.54	0.63	0.58	0.65	0.56	0.61	0.59	0.67	0.6	0.65	0.57	0.67	0.6	0.63
	Melanogenesis associated transcription factor	MITF	0.81	0.64	0.7	0.61	0.73	0.57	0.73	0.56	0.68	0.57	0.82	0.65	0.79	0.68	0.69	0.61
	Nuclear factor kappa B1	NFKB1	0.51	0.51	0.5	0.66	0.51	0.5	0.5	0.51	0.55	0.5	0.5	0.63	0.51	0.64	0.51	0.63
	Nuclear receptor 3C1	NR3C1	–	–	0.77	0.96	0.67	0.94	0.76	0.98	–	–	0.6	0.93	0.73	0.95	0.69	0.95
	Nuclear receptor 5A1	NR5A1	0.55	0.53	0.65	0.56	0.64	0.57	–	–	–	–	0.72	0.62	0.73	0.62	0.65	0.6
	Tumor protein p53	TP53	0.72	0.57	0.62	0.55	0.65	0.55	0.7	0.57	0.62	0.58	0.71	0.58	0.7	0.57	0.6	0.56
	Vitamin D receptor	VDR	0.5	0.57	0.5	0.6	0.52	0.6	0.51	0.54	0.53	0.53	0.58	0.62	0.55	0.59	0.53	0.59
Transporter	Abhydrolase domain containing 5	ABHD5	0.51	0.57	0.51	0.66	–	–	–	–	–	–	0.55	0.68	0.54	0.68	0.53	0.69
	Solute carrier family 6 member 3	SLC6A3	0.64	0.65	0.62	0.66	0.65	0.65	0.67	0.62	0.61	0.65	0.66	0.66	0.66	0.67	0.64	0.68
Unclassified protein	Ataxin 2	ATXN2	0.78	0.5	0.7	0.62	0.74	0.52	0.7	0.53	0.69	0.58	0.72	0.62	0.72	0.61	0.7	0.61
	ATPase family, AAA domain containing 5	ATAD5	0.58	0.56	0.52	0.67	0.59	0.62	0.55	0.62	0.52	0.6	0.6	0.65	0.64	0.65	0.52	0.68

(Continued)

TABLE 3 | Continued

Target class	Target name	Cell line	A375		A549		HA1E		HCC515		HT29		MCF7		PC3		VCAP	
		Target/ Descriptor	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP	GES	Morgan FP
	Endothelial PAS domain protein 1	EPAS1	–	–	0.63	0.62	–	–	–	–	–	–	0.73	0.68	0.76	0.65	0.69	0.67
	Geminin, DNA replication inhibitor	GMNN	0.7	0.59	0.71	0.58	0.69	0.55	0.65	0.59	0.67	0.55	0.76	0.62	0.75	0.59	0.68	0.6
	MLLT3, super elongation complex subunit	MLLT3	–	–	0.5	0.54	–	–	–	–	–	–	0.55	0.55	0.51	0.63	0.51	0.67
	MYC proto-oncogene, bHLH transcription factor	MYC	–	–	0.73	0.65	–	–	–	–	–	–	0.69	0.65	0.63	0.66	0.77	0.64
	Nuclear factor, erythroid 2 like 2	NFE2L2	0.57	0.51	0.57	0.6	0.55	0.58	0.57	0.59	0.58	0.53	0.56	0.61	0.61	0.6	0.59	0.6
	Nucleotide binding oligomerization domain containing 1	NOD1	–	–	0.58	0.66	–	–	–	–	0.6	0.58	0.66	0.69	0.65	0.68	0.56	0.7
	Nucleotide binding oligomerization domain containing 2	NOD2	0.5	0.53	0.53	0.61	0.56	0.54	–	–	0.52	0.53	0.58	0.68	0.61	0.68	0.56	0.64
	RAD52 homolog, DNA repair protein	RAD52	–	–	0.54	0.57	–	–	–	–	–	–	0.52	0.61	0.53	0.63	0.53	0.69
	TAR DNA binding protein	TARDBP	–	–	0.5	0.56	0.51	0.54	–	–	–	–	0.54	0.55	0.54	0.54	0.5	0.57

Target in lines, and cell line and used descriptor in columns. GES, model using gene expression signature. Morgan FP, model using chemical fingerprints from counterpart GES model dataset. Cells containing "–" corresponds to models that were not computed in cause of a too low number of actives (< 20) in the dataset to perform appropriate classification. Model presenting BA between 0.7 and 0.8 are highlighted in orange whereas those with BA > 0.8 are highlighted in red.

power of GESs and Morgan fingerprints for target prediction, based on a limited dataset depending on biological (GES and bioactivity assay) data availability.

DISCUSSION

Our results show that compound-induced transcriptomic responses derived from cell lines have the potential to support target prediction of unknown compounds with large structural diversity. Interestingly, we observed that compound induced biological responses are mostly cell line specific even when cell lines are derived from the same tissue. Nevertheless, machine learning models using GESs were shown to perform well as long as the appropriate cell line was used. Exploring biological spaces can help to overcome the limitations derived from a restricted chemical space when using traditional QSAR. To improve the predictivity of GES models, we have identified several limitations, and discuss possible improvements.

Data Acquisition

First limitations come from gene expression data preprocessing. Gene expression values were obtained through multiple preprocessing steps from the initial generated raw data. For instance, there is a first peak deconvolution step to determine the gene expression levels, that as well as the plate-normalized z-scoring to obtain the normalized ("Level 5") can still be improved as already stated by Li et al. (2017). Using GESs obtained with different preprocessing methods could potentially give more accurate normalized values leading to increased performances in machine learning models.

Secondly, the CMAP L1000 technology relies on the measurement of 978 landmark genes, representing about 5% of the human transcriptome (Pertea, 2012). The gene values of the remaining transcriptome can be inferred through different computational methods (Subramanian et al. (2017) method reached good prediction for 81% of inferred genes), that are still under improvement (Blasco et al., 2019). We decided to only use the 978 landmarks as input data for the machine models generated, to reflect real measured gene expression. Doing so, we might have missed some valuable information captured by a change of expression of the non-measured genes. Therefore, it would be interesting to explore the potential added value of expanding the number of descriptors by adding the inferred gene information to the target prediction models.

Data Restrictions

Another limitation is also coming with the activity dataset that was used. Since compound activity is a selective interaction, there is for each target a low number of active compounds compared to the number of inactive compounds. As a consequence, the training sets used for model building were highly unbalanced favorizing the prediction of the category inactive. Moreover, not every compound was tested for activity in all targets, leading to a sparse dataset (5% of total compound target interactions are known).

On top of this activity data limitation, not all available compounds were profiled in all the 8 cell lines used in this work.

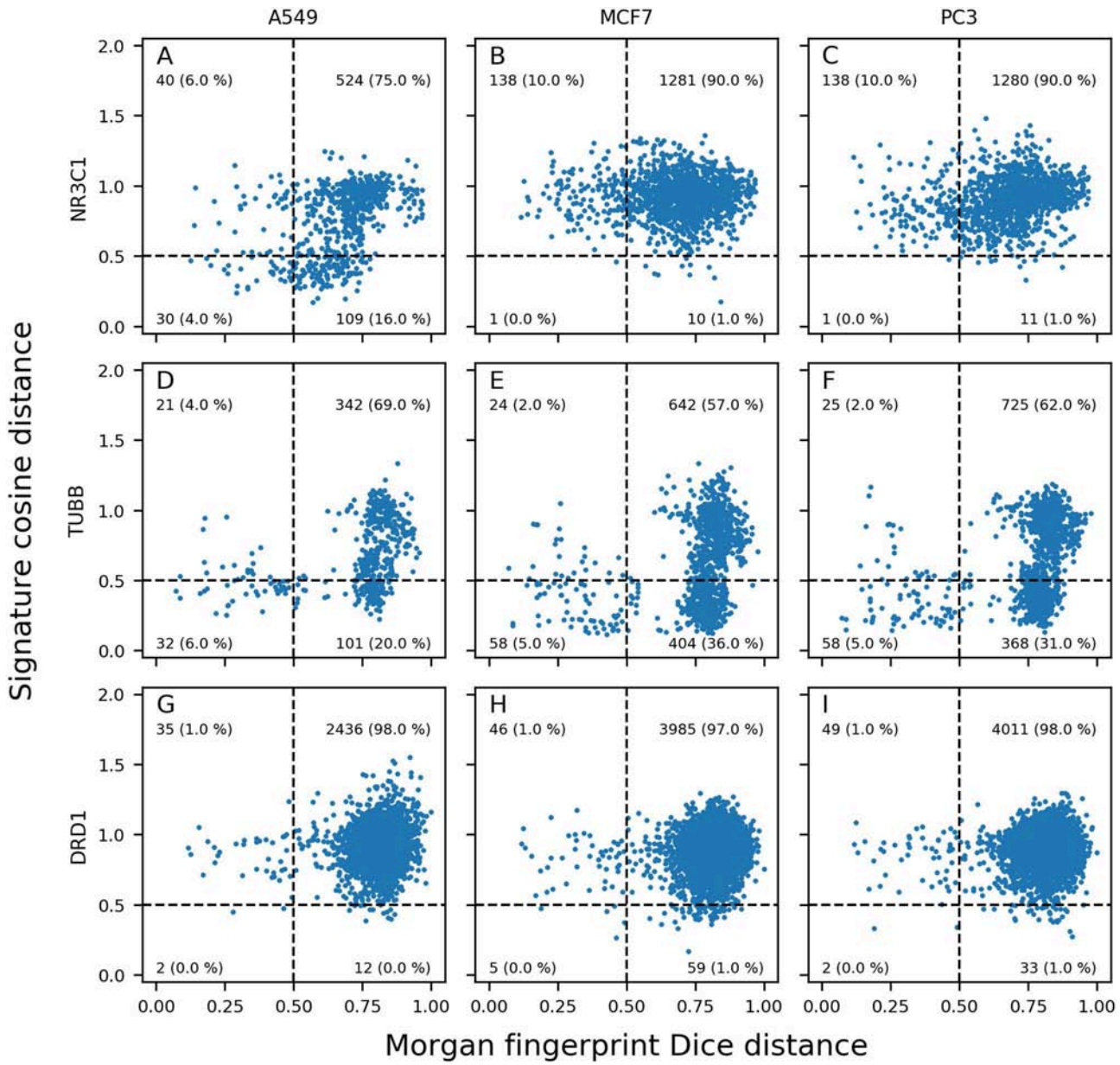


FIGURE 6 | Morgan fingerprints Dice distance vs. GES cosine distance (distance plots). Different panels show information for pairs of NR3C1 **(A–C)**, TUBB **(D–F)**, and DRD1 **(G–I)**; active compounds in A549 **(A,D,G)**, MCF7 **(B,E,H)**, and PC3 **(C,F,I)** cell lines.

There were only about 600 compounds profiled in all the cell lines, which is too limited to build predictive models, with regard to available activity data. Consequently, one dataset per cell line was created, formed by compounds profiled in this cell line and resulting GESs. For each target prediction, the cell line datasets were restricted to compounds having a known label for the target of interest. Since each task used a different dataset, performances of models across cell lines or targets the comparison across GES models was not possible. The difference in dataset sizes is explaining at least partly the variation of performances of GES models across targets, ranging from models close to a random predictor (BA = 0.50) to good GES models (BA = 0.88), as well as the variation of performances of counterpart Morgan FP models (BA ranging from 0.50 to 0.98).

Biological Response Constraints

Biologically, variation of GES model performances can also be caused by the difference in the pathway representation in the cell lines and consequently to compound induced signatures. Compounds active on a given target might show GESs with different degree of similarity or no similarity among the considered cell lines. as illustrated by the cases of NR3C1, TUBB and DRD1. Gene expression responses depend on the cellular context as shown in this work and elsewhere (Chen et al., 2013;

Yu et al., 2019). Thus, the biological system in which the GESs are generated is of utmost importance for target prediction.

Due to practical aspects (scalability, low price, etc.), biological systems such as *in vitro* immortalized cell cultures (like cancer cell lines used in this work) are widely used, but they come with some disadvantages: they show limited physiological representativity and have been shown to drift along passages (Hughes et al., 2007). Even within the same cell line, it was shown that strains show different responses to the same compounds, indicating a reduced reproducibility between generated GESs (Ben-David et al., 2018). Ideally, the GESs should be derived from biological systems mimicking as much as possible the biological responses observed in the corresponding target organ.

The advantage of transcriptomic evaluations over single endpoint assays is that in theory they have the potential to capture integrative responses from compound treatments, ranging from on target activity at high potency to off-target activities at lower potencies, depending on the tested concentrations. GESs responses are also known to be variable depending on time exposition (Aguayo-Orozco et al., 2018). That is the reason why we selected data sets originating from the same study design. GESs measured at a concentration of 10 μM after 24 h of treatment of the cell lines were extracted, as this is the most represented experimental condition (De Wolf et al., 2016; Lv et al., 2017).

GES Models Versus Morgan FP Models

We showed that using GES datasets produced by the Broad Institute with the CMAP L1000 technique (Subramanian et al., 2017), random forest models outperformed counterpart Morgan FP models for target prediction in 25% of the cases. Evidently, the outcome of this comparison is depending on the available data for the different targets to build the models as illustrated by the wide range of differences of BA between the two types of predictive models. Practically, both QSAR and transcriptomic descriptors represent good opportunities for target prediction, but each come with advantages and constraints that needs to be considered when building predictive models.

QSAR models for target prediction are widely used because of the wide dataset available, with existing databases like PubChem or ChEMBL. Most QSAR descriptors are discrete unambiguous values extracted from the chemical formula of compounds, thus easily computed. In the context of hit discovery, a major drawback of QSAR models is that they show significant error rate when trying to predict activity for compounds that are too structurally different from the training set (Cherkasov et al., 2014). Using a new set of descriptors, like compound bioactivity such as GESs extracted from *in vitro* experiments, can help in target prediction while escaping from the classical chemical space limitation observed in QSAR approaches.

On the other hand, GESs represents a number of changes on a certain number of genes (the 978 landmarks), capturing the effect of compounds. These data could be used to make inference about biology (i.e., finding targets or biomarkers). Each cell line shows a unique biological space that can be explored. However, these biological experiment data are prone to technical and biological variability like discussed earlier. Gene expression can be measured in different dose and time conditions, adding dimensions to explore in order to find the conditions reaching best performances in GES models. Finally, the gene expression measurements are more and more cost effective, making the use of such data at a large scale possible.

When exploring a new chemical class in hit discovery, evaluating chemical-induced biological responses in appropriate cell-lines using transcriptomic profiling can support chemical prioritization. This biologically-based approach present the advantage in a given biological space of being in principle chemical space independent as opposed to QSAR modeling that is constrained by the chemical space of the training set. Furthermore, during lead optimization, biological spaces inform about the direct activity of candidates, which can help fine-tuning their desired activity profile, by optimizing the on-target activity. It has been recently shown that this type of data can be used for *de novo* chemical design fulfilling a specific GES (Méndez-Lucio et al., 2020). In a chemical safety approach, it can be used to detect compound interaction with off-targets. However, a difference between these 2 applications would be the conditions in which the GESs are generated: on-target effects are observable at low concentrations (Kd often in the nanomolar range), while off-target effect are known to typically appear at higher concentration as illustrated by Li et al. (2019).

In conclusion, in this work, we evaluated the use of a large public dataset of compound-induced transcriptomic data, to predict compound activity on 69 molecular targets. We compared machine learning models built with transcriptomics data with counterpart models built using Morgan fingerprints. Active compounds on a given target could exhibit similar signatures in one or multiple cell lines, independent from the chemical structure similarity between these active compounds. For 25% of the tasks, random forest models using transcriptomics signatures performed similarly or better than counterpart models built with Morgan fingerprints, occurring mostly using signatures produced in cell lines that showed similar signatures for active compounds on a given target. Compound-induced transcriptomic data could offer a great opportunity for target prediction based on cell response similarity and allows to circumvent the applicability domain limitation of QSAR models.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found in the Gene Expression Omnibus https://www.ncbi.nlm.nih.gov/gds (GEO IDs GSE92742 and GSE70138).

AUTHOR CONTRIBUTIONS

BB, OM-L, and DR contributed to the conception of the work. BB performed the workflow with the help of OM-L. BB wrote the first draft of the manuscript. JW, OM-L, and DR provided guidance and helped with the manuscript preparation and contributed to manuscript revision, read and approved the submitted version. DR and OM-L wrote sections of the manuscript.

ACKNOWLEDGMENTS

Authors thank Karen Tilmant for proof reading the text and for her useful comments. We are also grateful to Arwa Al-Dilaimi, Djork-Arne Clevert and Angela Becker for supporting the project and for insightful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2020.00296/full#supplementary-material

REFERENCES

Aguayo-Orozco, A., Bois, F. Y., Brunak, S., and Taboureau, O. (2018). Analysis of time-series gene expression data to explore mechanisms of chemical-induced hepatic steatosis toxicity. *Front. Genet.* 9:396. doi: 10.3389/fgene.2018.00396

Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., and Zhavoronkov, A. (2016). Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharm.* 13, 2524–2530. doi: 10.1021/acs.molpharmaceut.6b00248

Ben-David, U., Siranosian, B., Ha, G., Tang, H., Oren, Y., Hinohara, K., et al. (2018). Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* 560, 325–330. doi: 10.1038/s41586-018-0409-3

Blasco, A., Endres, M. G., Sergeev, R. A., Jonchhe, A., Macaluso, N. J. M., Narayan, R., et al. (2019). Advancing computational biology and bioinformatics research through open innovation competitions. *PLoS ONE* 14:e0222165. doi: 10.1371/journal.pone.0222165

Breiman, L. (2001). Randomforest2001. *Mach. Learn.* 45, 5–32. doi: 10.1017/CBO9781107415324.004

Chaaban, S., and Brouhard, G. J. (2017). A microtubule bestiary: structural diversity in tubulin polymers. *Mol. Biol. Cell* 28, 2924–2931. doi: 10.1091/mbc.E16-05-0271

Chen, M., Xiao, J., Zhang, Z., Liu, J., Wu, J., and Yu, J. (2013). Identification of human HK genes and gene expression regulation study in cancer from transcriptomics data analysis. *PLoS ONE* 8:e54082. doi: 10.1371/journal.pone.0054082

Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., et al. (2016). Drug-target interaction prediction: databases, web servers and computational models. *Brief Bioinform.* 17, 696–712. doi: 10.1093/bib/bbv066

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., et al. (2014). QSAR modeling: where have you been? where are you going to?. *J. Med. Chem.* 57, 4977–5010. doi: 10.1021/jm4004285

Corsello, S. M., Bittker, J. A., Liu, Z., Gould, J., McCarren, P., Hirschman, J. E., et al. (2017). The drug repurposing hub: a next-generation drug library and information resource. *Nat. Med.* 23, 405–408. doi: 10.1038/nm.4306

Cruz-Monteagudo, M., Medina-Franco, J. L., Pérez-Castillo, Y., Nicolotti, O., Cordeiro, M. N. D. S., and Borges, F. (2014). Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discov. Today* 19, 109–1080. doi: 10.1016/j.drudis.2014.02.003

De Wolf, H., Cougnaud, L., Van Hoorde, K., De Bondt, A., Wegner, J. K., Ceulemans, H., et al. (2018). High-throughput gene expression profiles to define drug similarity and predict compound activity. *Assay Drug Dev. Technol.* 16, 162–176. doi: 10.1089/adt.2018.845

De Wolf, H., De Bondt, A., Turner, H., and Göhlmann, H. W. H. (2016). Transcriptional characterization of compounds: lessons learned from the public LINCS data. *Assay Drug Dev. Technol.* 14, 252–260. doi: 10.1089/adt.2016.715

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Hieronymus, H., Lamb, J., Ross, K. N., Peng, X. P., Clement, C., Rodina, A., et al. (2006). Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell* 10, 321–330. doi: 10.1016/j.ccr.2006.09.005

Hofmarcher, M., Rumetshofer, E., Clevert, D. A., Hochreiter, S., and Klambauer, G. (2019). Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. *J. Chem. Inf. Model* 59, 1163–1171. doi: 10.1021/acs.jcim.8b00670

Hughes, P., Marshall, D., Reid, Y., Parkes, H., and Gelber, C. (2007). The costs of using unauthenticated, over-passaged cell lines: how much more data do we need? *BioTechniques*. 43, 575, 577–578, 581–582. doi: 10.2144/000112598

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., et al. (2000). Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126. doi: 10.1016/S0092-8674(00)00015-5

Kuthuru, S., Szafran, A. T., Stossi, F., Mancini, M. A., and Rao, A. (2019). Leveraging image-derived phenotypic measurements for drug-target interaction predictions. *Cancer Inform.* 18:1176935119856595. doi: 10.1177/1176935119856595

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., et al. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935. doi: 10.1126/science.1132939

Lapins, M., and Spjuth, O. (2019). Evaluation of gene expression and phenotypic profiling data as quantitative descriptors for predicting drug targets and mechanisms of action. *Biorxiv*. doi: 10.1101/580654. [Epub ahead of print].

Laufkötter, O., Sturm, N., Bajorath, J., Chen, H., and Engkvist, O. (2019). Combining structural and bioactivity-based fingerprints improves prediction performance and scaffold hopping capability. *J. Cheminform.* 11:54. doi: 10.1186/s13321-019-0376-1

Lee, H., Kang, S., and Kim, W. (2016). Drug repositioning for cancer therapy based on large-scale drug-induced transcriptional signatures. *PLoS ONE* 11:e0150460. doi: 10.1371/journal.pone.0150460

Li, A., Lu, X., Natoli, T., Bittker, J., Sipes, N. S., Subramanian, A., et al. (2019). The carcinogenome project: *in vitro* gene expression profiling of chemical perturbations to predict long-term carcinogenicity. *Environ. Health Perspect.* 127:47002. doi: 10.1289/EHP3986

Li, Z., Li, J., and Yu, P. (2017). l1kdeconv: an R package for peak calling analysis with LINCS L1000 data. *BMC Bioinformatics* 18:356. doi: 10.1186/s12859-017-1767-9

Lv, C., Wu, X., Wang, X., Su, J., Zeng, H., Zhao, J., et al. (2017). The gene expression profiles in response to 102 traditional Chinese medicine (TCM) components: a general template for research on TCMs. *Sci. Rep.* 7:352. doi: 10.1038/s41598-017-00535-8

Méndez-Lucio, O., Baillif, B., Clevert, D. A., Rouquié, D., and Wichard, J. (2020). *De novo* generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat. Commun.* 11:10. doi: 10.1038/s41467-019-13807-w

Pabon, N. A., Xia, Y., Estabrooks, S. K., Ye, Z., Herbrand, A. K., Süß, E., et al. (2018). Predicting protein targets for drug-like compounds using transcriptomics. *PLoS Comput. Biol.* 14:e1006651. doi: 10.1371/journal.pcbi.1006651

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://jmlr.org/papers/v12/pedregosa11a.html

Pertea, M. (2012). The human transcriptome: an unfinished story. *Genes* 3, 344–360. doi: 10.3390/genes3030344

Petrone, P. M., Simms, B., Nigsch, F., Lounkine, E., Kutchukian, P., Cornett, A., et al. (2012). Rethinking molecular similarity: Comparing compounds on the basis of biological activity. *ACS Chem. Biol.* 7, 1399–1409. doi: 10.1021/cb3001028

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi: 10.1021/ci100050t

Rouquié, D., Heneweer, M., Botham, J., Ketelslegers, H., Markell, L., Pfister, T., et al. (2015). Contribution of new technologies to characterization and prediction of adverse effects. *Crit. Rev. Toxicol.* 45, 172–183. doi: 10.3109/10408444.2014.986054

Scheeder, C., Heigwer, F., and Boutros, M. (2018). Machine learning and image-based profiling in drug discovery. *Curr. Opin. Syst. Biol.* 10, 43–52. doi: 10.1016/j.coisb.2018.05.004

Simm, J., Klambauer, G., Arany, A., Steijaert, M., Wegner, J. K., Gustin, E., et al. (2018). Repurposing high-throughput image assays enables biological activity prediction for drug discovery. *Cell Chem. Biol.* 25, 611-618. doi: 10.1016/j.chembiol.2018.01.015

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 171, 1437–1452 e17. doi: 10.1016/j.cell.2017.10.049

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 18, 463–477. doi: 10.1038/s41573-019-0024-5

Van Der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605. Available online at: http://www.jmlr.org/papers/v9/vandermaaten08a.html

Wang, Y., Suzek, T., Zhang, J., Wang, J., He, S., Cheng, T., et al. (2014). PubChem Bioassay: 2014 update. *Nucleic Acids Res.* 1, D1075–D1082. doi: 10.1093/nar/gkt978

Wei, G., Twomey, D., Lamb, J., Schlis, K., Agarwal, J., Stam, R. W., et al. (2006). Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer Cell* 10, 331–342. doi: 10.1016/j.ccr.2006.09.006

Yu, K., Chen, B., Aran, D., Charalel, J., Yau, C., Wolf, D. M., et al. (2019). Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nat. Commun.* 10:3574. doi: 10.1038/s41467-019-11415-2

Conflict of Interest: JW is employee of Bayer AG. OM-L, BB, and DR work directly or indirectly for Bayer SAS.

Copyright © 2020 Baillif, Wichard, Méndez-Lucio and Rouquié. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

REVIEW
published: 28 April 2020
doi: 10.3389/fchem.2020.00343

Structure-Based Virtual Screening: From Classical to Artificial Intelligence

Eduardo Habib Bechelane Maia[1,2]*, *Letícia Cristina Assis*[1], *Tiago Alves de Oliveira*[2], *Alisson Marques da Silva*[2] and *Alex Gutterres Taranto*[1]

[1] *Laboratory of Pharmaceutical Medicinal Chemistry, Federal University of São João Del Rei, Divinópolis, Brazil,* [2] *Federal Center for Technological Education of Minas Gerais—CEFET-MG, Belo Horizonte, Brazil*

The drug development process is a major challenge in the pharmaceutical industry since it takes a substantial amount of time and money to move through all the phases of developing of a new drug. One extensively used method to minimize the cost and time for the drug development process is computer-aided drug design (CADD). CADD allows better focusing on experiments, which can reduce the time and cost involved in researching new drugs. In this context, structure-based virtual screening (SBVS) is robust and useful and is one of the most promising *in silico* techniques for drug design. SBVS attempts to predict the best interaction mode between two molecules to form a stable complex, and it uses scoring functions to estimate the force of non-covalent interactions between a ligand and molecular target. Thus, scoring functions are the main reason for the success or failure of SBVS software. Many software programs are used to perform SBVS, and since they use different algorithms, it is possible to obtain different results from different software using the same input. In the last decade, a new technique of SBVS called consensus virtual screening (CVS) has been used in some studies to increase the accuracy of SBVS and to reduce the false positives obtained in these experiments. An indispensable condition to be able to utilize SBVS is the availability of a 3D structure of the target protein. Some virtual databases, such as the Protein Data Bank, have been created to store the 3D structures of molecules. However, sometimes it is not possible to experimentally obtain the 3D structure. In this situation, the homology modeling methodology allows the prediction of the 3D structure of a protein from its amino acid sequence. This review presents an overview of the challenges involved in the use of CADD to perform SBVS, the areas where CADD tools support SBVS, a comparison between the most commonly used tools, and the techniques currently used in an attempt to reduce the time and cost in the drug development process. Finally, the final considerations demonstrate the importance of using SBVS in the drug development process.

Keywords: SBVS, homology modeling, consensus virtual screening, scoring functions, computer-aided drug design

OPEN ACCESS

Edited by:
Teodorico Castro Ramalho, Universidade Federal de Lavras, Brazil

Reviewed by:
Manoelito Coelho Santos Junior, State University of Feira de Santana, Brazil
Daniel Henriques Soares Leal, Federal University of Itajubá, Brazil

***Correspondence:**
Eduardo Habib Bechelane Maia
habib@cefetmg.br

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *27 June 2019*
Accepted: *01 April 2020*
Published: *28 April 2020*

Citation:
Maia EHB, Assis LC, Oliveira TA, Silva AM and Taranto AG (2020) Structure-Based Virtual Screening: From Classical to Artificial Intelligence. Front. Chem. 8:343. doi: 10.3389/fchem.2020.00343

INTRODUCTION

In the past, the discovery of new drugs was made through random screening and empirical observations of the effects of natural products for known diseases.

This random screening process, although inefficient, led to the identification of several important compounds until the 1980s. Currently, this process is improved by high-throughput screening (HTS), which is suitable for automating the screening process of many thousands of compounds against a molecular target or cellular assay very quickly. The milestone of HTS was used in the identification of cyclosporine A as a immunosuppressant (von Wartburg and Traber, 1988). Subsequently, several drugs such as nevirapine (Merluzzi et al., 1990), gefitinib (Ward et al., 1994), and maraviroc (Wood and Armour, 2005) have reached the market. Notably, gefitinib was discovered by computational methods through a collection of 1500 compounds by ALLADIN (Martin, 1992) software. In addition, computational methods have been used to search successful compounds against malaria disease (Nunes et al., 2019). The structures of these molecules are in **Figure 1**.

Alternatively, the increased cost and evolution of medicines available in the last century have led to an improvement in the quality of life of the world population. However, while the average quality of life has been improved, a third of the population is still without access to essential medicines, which means that more than 2 billion people cannot afford to buy basic medicines (Leisinger et al., 2012). This problem is even worse in some places in Africa and Asia, where more than 50% of the people face problems obtaining medicines (Leisinger et al., 2012). Moreover, throughout the world, more than 18 million deaths that occur every year could be avoided, as well as tens of millions of deaths related to poverty and lack of access to essential medicines (Sridhar, 2008). The price of many medicines is inaccessible to limited-income populations and middle-income countries (Stevens and Huys, 2017).

While there is a need to increase the population's access to medicines, the pharmaceutical industry is facing unprecedented challenges in its business model (Paul et al., 2010). The current process of developing new drugs began to mature only in the second half of the twentieth century. The process evolved from observations made in the correlation of certain physical-chemical properties of organic molecules with biological potency. Optimization of these compounds by the incorporation of more favorable substituents resulted in more potent drugs. X-ray crystallography and nuclear magnetic resonance (NMR) techniques have provided information on the structures of enzymes and drug receptors. Many drugs, such as angiotensin-converting-enzyme (ACE) inhibitors, have been introduced to the clinical practice from this structural information.

The drug development process aims to identify bioactive compounds to assist in the treatment of diseases. In summary (**Figure 2**), the process starts with the identification of molecular targets for a given compound (natural or synthetic) and is followed by their validation. Then, virtual screening (VS) can be used to assist in hit identification (identification of active drug candidates) and lead optimization (biologically active compounds are transformed into appropriate drugs by improving their physicochemical properties). Finally these optimized leads will undergo preclinical and clinical trials to ultimately be approved by regulatory bodies (Lima et al., 2016).

In general, this process is time-consuming, laborious and expensive. The development of a new drug has an average cost between 1 and 2 billion USD and could take 10–17 years (Leelananda and Lindert, 2016), since it has to move through all phases for new drug development, from target discovery to drug registration. Even so, Arrowsmith (2012) showed that the probability of a drug candidate reaching the market after entering Phase I clinical trials fell from 10% in the 2002–2004 period to approximately 5% between 2006 and 2008, which represents a 50% decrease in just 4 years.

Thus, researchers are constantly investing in the development of new methods to increase the efficiency of the drug discovery process (Hillisch et al., 2004). The computer-aided drug design (CADD) approach, which employs molecular modeling techniques, has been used by researchers to increase the efficacy in the development of new drugs since it uses *in silico* simulations. Molecular modeling allows the analysis of many molecules in a short period of time, demonstrating how they interact with targets of pharmacological interest even before their synthesis. The technique allows the simulation and prediction of several essential factors, such as toxicity, activity, bioavailability and efficacy, even before the compound undergoes *in vitro* testing, thus allowing better planning and direction of the research (Ferreira et al., 2011). Better planning of the research means, in this case, fewer *in vitro* and *in vivo* experiments. Therefore, it reduces the run time and overall research costs.

In this context, virtual screening (VS) is a promising *in silico* technique used in the drug discovery process. An indispensable condition in performing virtual screening is the availability of a 3D structure of the target protein (Cavasotto, 2011). Therefore, some virtual databases were created to store 3D structures of molecules. Virtual screening is now widely applied in the development of new drugs and has already contributed to compounds on the market. Examples of drugs that came to the market with the assistance of VS include captopril (antihypertensive drug), saquinavir, ritonavir, and indinavir (three drugs for the treatment of human immunodeficiency virus), tirofiban (fibrinogen antagonist), dorzolamide (used to treat glaucoma), zanamivir (a selective antiviral for influenza virus), aliskiren (antihypertensive drug), boceprevir (protease inhibitor used for the treatment of hepatitis C), nolatrexed (in phase III clinical trial for the treatment of liver cancer) (Talele et al., 2010; Sliwoski et al., 2013; Devi et al., 2015; Nunes et al., 2019). The structures of these molecules are in **Figures 3**, **4**.

This review will present an overview of the challenges involved in the development of new drugs. Section Computer-aided drug design (CADD) will describe CADD while section 3 will demonstrate how VS has been used as an agent in the process of developing of new drugs. Section Virtual screening (VS), in turn, will explain the main scoring functions used in recent scientific research. Section Consensus docking will explain consensus docking, which is a relatively unexplored

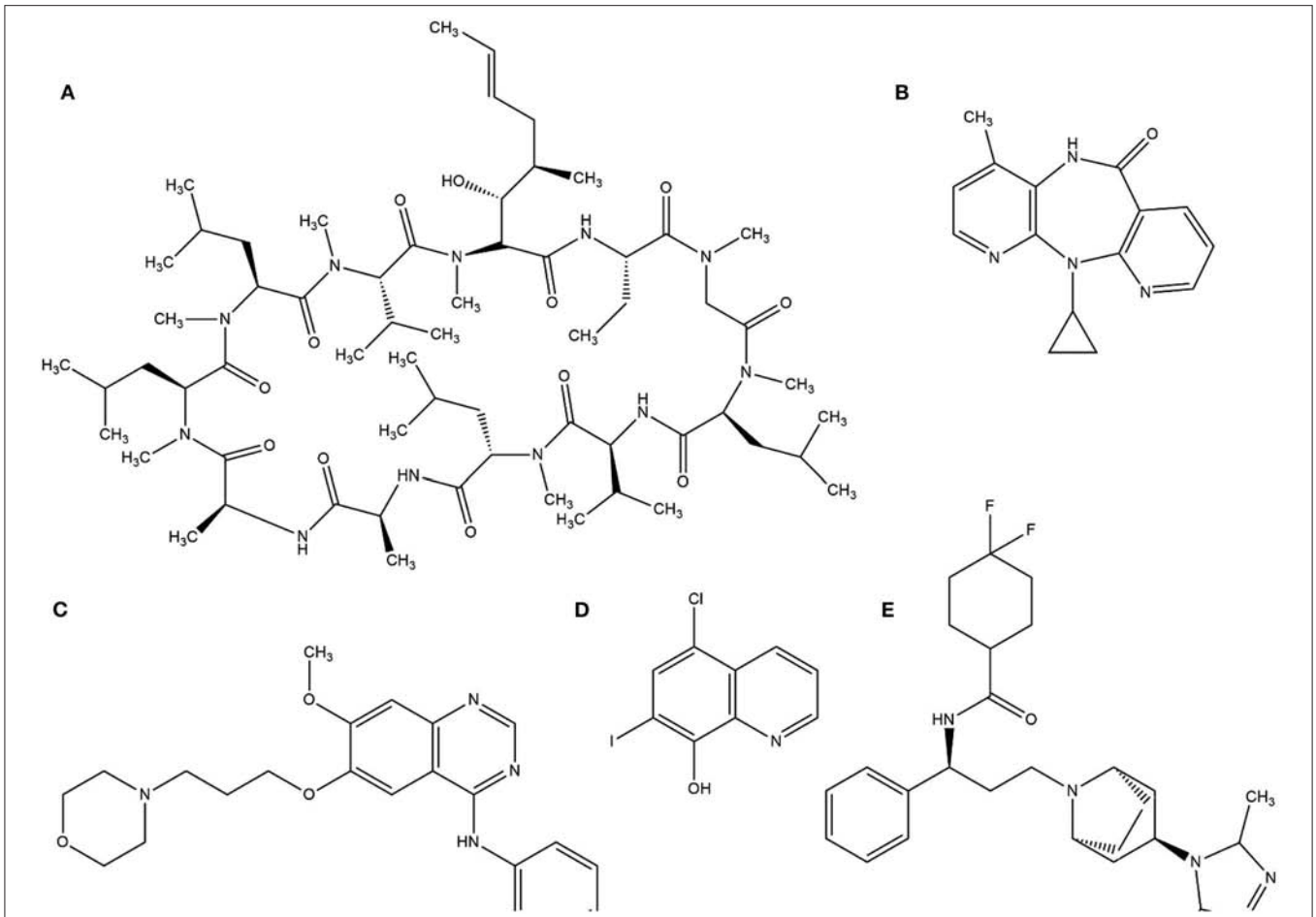


FIGURE 1 | Examples of structures identified by HTS. **(A)** cyclosporine A, **(B)** Neviparine, **(C)** Gefitinib, **(D)** Clioquinol, and **(E)** Maraviroc.

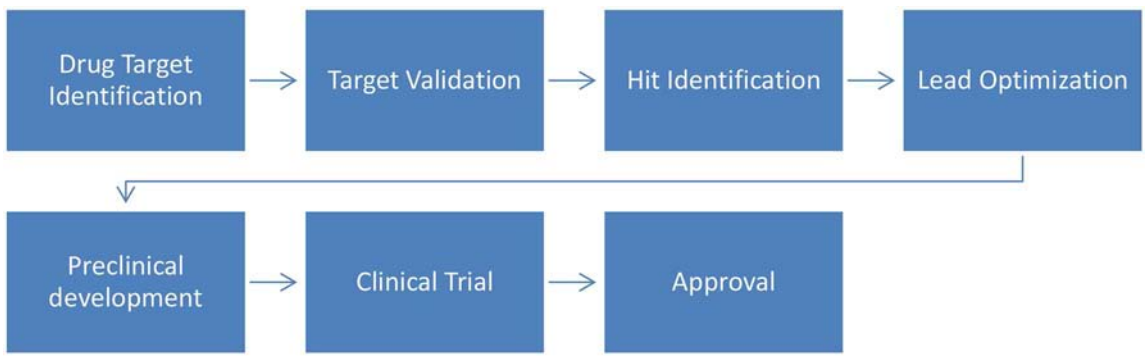


FIGURE 2 | Drug development timeline.

topic in the virtual screening process. Section Virtual Databases will list the main virtual databases used in this task. Section Virtual screening algorithms presents the main VS algorithms used. Section Methods of evaluating the quality of a simulation will present some evaluation methods used to verify if the quality of the performed model/simulation is good. Section VS software programs, in turn, will present the main VS software currently used. Section Final considerations will present final considerations.

COMPUTER-AIDED DRUG DESIGN (CADD)

One approach used to increase the effectiveness in the development of new drugs is the use of computer-aided

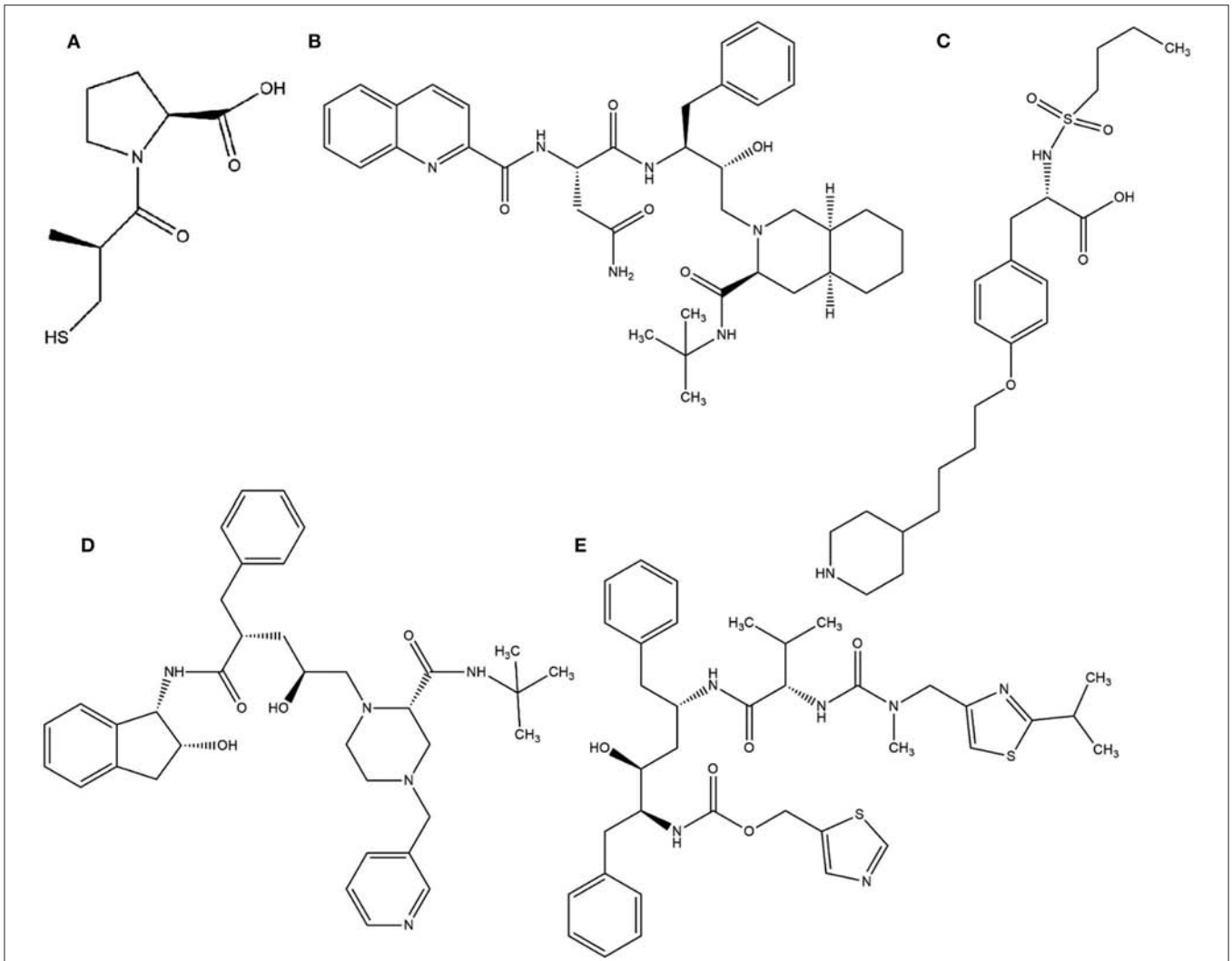


FIGURE 3 | Drugs that came to the market with the assistance of VS: **(A)** Captopril, **(B)** Saquinavir, **(C)** Tirofiban, **(D)** Indinavir, **(E)** Ritonavir.

drug design (CADD, well known as an *in silico* method) techniques, which uses a computational chemistry approach for the drug discovery process. CADD is a cyclic process for developing new drugs, in which all stages of design and analysis are performed by computer programs, operated by medicinal chemists (Oglic et al., 2018).

Strategies for CADD may vary, depending on what information about the target and ligand are available. In the early stage of the drug development process, it is normal for little or no information to exist about the target, ligands, or their structures. CADD techniques are able to obtain this information, such as which proteins can be targeted in pathogenesis and what are the possible active ligands that can inhibit these proteins. Kapetanovic (2008) briefly notes that CADD comprises (i) making the drug discovery and development process faster with the contribution of *in silico* simulations; (ii) optimizing and identifying new drugs using the computational approach to discover chemical and biological information about possible ligands and/or molecular targets; and (iii) using simulations to eliminate compounds with undesirable properties and selecting candidates with more chances for success. Recent software uses empirical molecular mechanics, quantum mechanics and, more recently, statistical mechanics. This last advancement allows the explicit effects of solvents to be incorporated (Das and Saha, 2017).

CADD gained prominence, as it allows obtaining information about the specific properties of a molecule, which can influence its interaction with the receptor. Thus, it has been considered a useful tool in rational planning and the discovery of new bioactive compounds. Alternatively, CADD simulations require a high computational cost, taking up to weeks if long jobs are used for molecular dynamics simulations. Therefore, it is a continuous challenge to find viable solutions that reduce the simulation runtime and simultaneously increase the accuracy of the

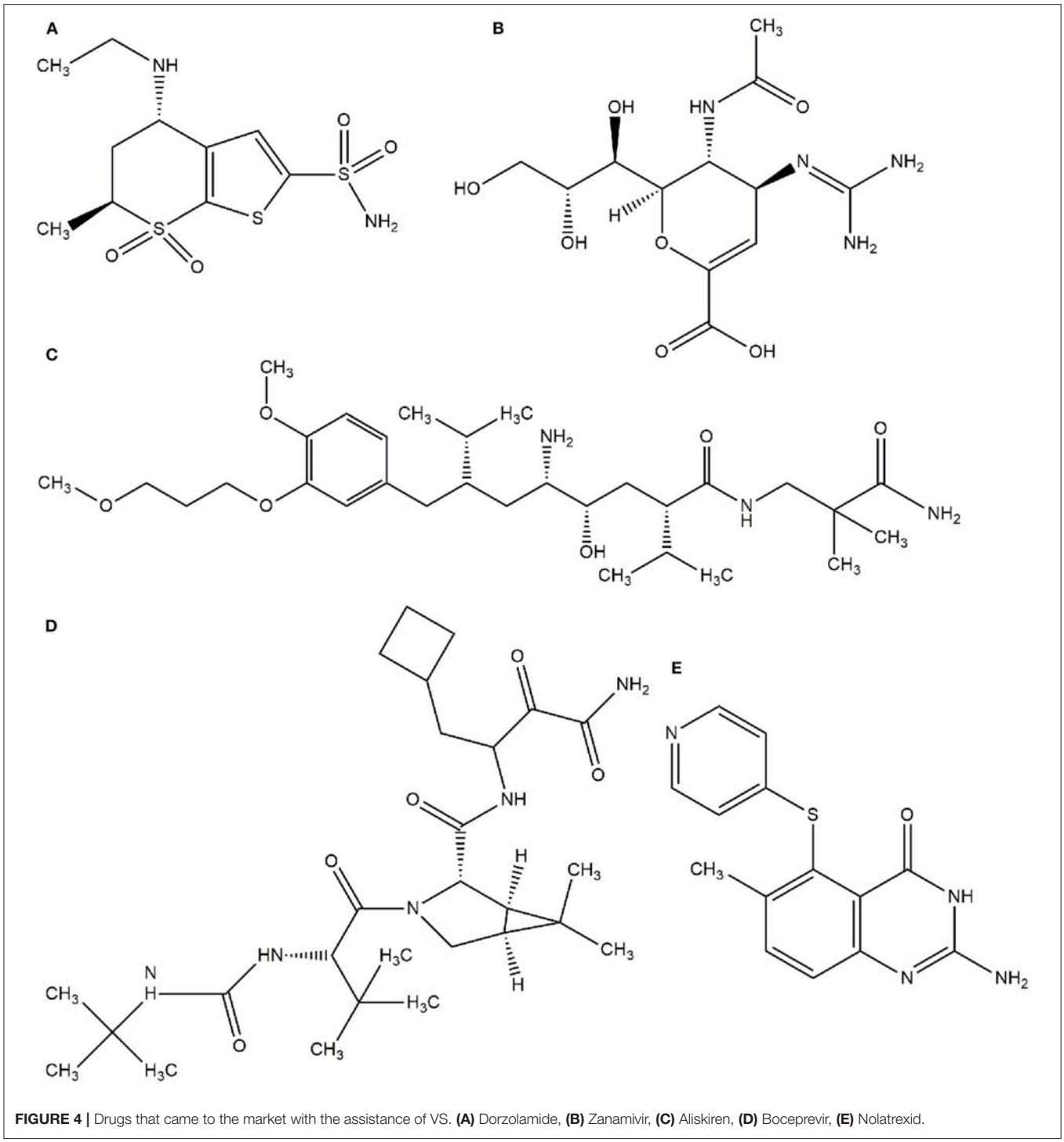


FIGURE 4 | Drugs that came to the market with the assistance of VS. **(A)** Dorzolamide, **(B)** Zanamivir, **(C)** Aliskiren, **(D)** Boceprevir, **(E)** Nolatrexid.

simulations (Ripphausen et al., 2011). In this context, VS is a promising approach.

VIRTUAL SCREENING (VS)

Popular VS techniques originated in the 1980s, but the first publication about VS appeared in 1997 (Horvath, 1997). In recent times, the use of VS techniques has been shown to be an excellent alternative to high throughput screening, especially in terms of cost-effectiveness and probability of finding the most appropriate result through a large virtual database (Surabhi and Singh, 2018).

VS is an *in silico* technique used in the drug discovery process. During VS, large databases of known 3D structures are automatically evaluated using computational methods (Maia et al., 2017). VS works like a funnel, by selecting

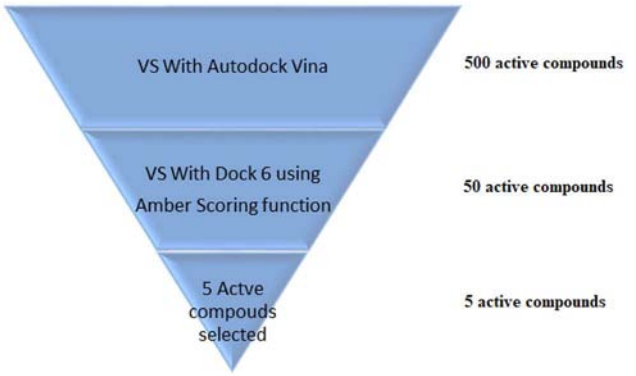


FIGURE 5 | VS scheme.

more promising molecules for *in vitro* assays to be performed. In the example shown in **Figure 5**, it is assumed that a virtual screening will be performed on 500 possible active ligands for a target. Then, VS with AutoDock Vina (Trott and Olson, 2009) was carried out and the top 50 ligands were selected. Then, a VS using DOCK 6 (Allen et al., 2015) with the Amber scoring function was performed. DOCK 6 with Amber scoring function takes longer, because it performs molecular dynamics, but it promises better results. Finally, after VS with DOCK 6, the top 5 active compounds are selected to be purchased and then tested *in vitro*. With the use of VS, it is expected that those identified molecules are more susceptible to binding to the molecular target, which is typically a protein or enzyme receptor. Therefore, VS assists in identifying the most promising hits able to bind to the target protein or enzyme receptor, and only the most promising molecules are synthesized. In addition, VS identifies compounds that may be toxic or have unfavorable pharmacodynamic (for example, potency, affinity, selectivity) and pharmacokinetic (for example, absorption, metabolism, bioavailability) properties. Thus, VS techniques play a prominent role among strategies for the identification of new bioactive substances (Berman et al., 2013).

VS for drug discovery is becoming an essential tool to assist in fast and cost-effective lead discovery and drug optimization (Maia et al., 2017). This technique can aid in the discovery of bioactive molecules, since they allow the selection of compounds in a structure database that are most likely to show biological activity against a target of interest. After identification, these bioactive molecules undergo biological assays. In addition, there are VS techniques using machine learning methods that predict compounds with specific pharmacodynamic, pharmacokinetic or toxicological properties based on their structural and physicochemical properties that are derived from the ligand structure (Ma et al., 2009). Hence, VS tools play a prominent role among the strategies used for the identification of new bioactive substances, since they increase the speed of the drug discovery process as long as they automatically evaluate large compound libraries through computational simulations (Maithri and Narendra, 2016).

Structure based virtual screening (SBVS) is a robust, useful and promising *in silico* technique for drug design (Lionta et al., 2014). Therefore, this review will address SBVS, although there are other types of VS such as ligand-based virtual screening (Banegas-Luna et al., 2018) and fragment-based virtual screening (Wang et al., 2015).

4-Structure-Based Virtual Screening (SBVS)

Structure-based virtual screening (SBVS), also known as target-based virtual screening (TBVS), attempts to predict the best interaction between ligands against a molecular target to form a complex. As a result, the ligands are ranked according to their affinity to the target, and the most promising compounds are shown at the top of the list. SBVS methods require that the 3D structure of the target protein be known so that the interactions between the target and each chemical compound can be predicted *in silico* (Liu et al., 2018). In this strategy, the compounds are selected from a database and classified according to their affinity for the receptor site.

Among the techniques of SBVS, molecular docking is noteworthy due to its low computational cost and good results achieved (Meng et al., 2011). This technique emerged in the 1980s, when Kuntz et al. (1982) designed and tested a set of algorithms that could explore the geometrically feasible alignments of a ligand and target. However, although the approach was promising, it was only in the 1990s that it became widely used after there was an improvement in the techniques used in conjunction with an increase in the computational power and a greater access to the structural data of target molecules. During the execution of SBVS, the evaluated molecules are sorted according to their affinity to the receptor site. Hence, it is possible to identify ligands that are more likely to present some pharmacological activity with the molecular target. Score functions are used to verify the likelihood of a binding site describing the affinity between the ligand and target. In this process, a reliable scoring function is the critical component of the docking process (Leelananda and Lindert, 2016).

The use of SBVS has advantages and disadvantages. Among the advantages are the following:

I There is a decrease in the time and cost involved in the screening of millions of small molecules.
II There is no need for the physical existence of the molecule, so it can be tested computationally even before being synthesized.
III There are several tools available to assist SBVS.

The disadvantages can be highlighted as the following:

I Some tools work best in specific cases, but not in more general cases (Lionta et al., 2014).
II It is difficult to accurately predict the correct binding position and classification of compounds due to the difficulty of parameterizing the complexity of ligand-receptor binding interactions.
III It can generate false positives and false negatives.

Despite the disadvantages noted above, many studies using SBVS have been developed in recent years (Carregal et al., 2017; Mugumbate et al., 2017; Wójcikowski et al., 2017; Carpenter et al., 2018; Dutkiewicz and Mikstacka, 2018; Surabhi and Singh, 2018; Nunes et al., 2019), which shows that although SBVS has disadvantages, it is still wide used for developing drugs due to the reduction of time and cost. However, docking protocols are essential for achieving accurate SBVS. These protocols are composed of two main components: the search algorithm and the score function.

Search Algorithms

Search algorithms are used to systematically search for ligand orientations and conformations at the binding site. A good docking protocol will achieve the most viable ligand conformations, in addition the most realistic position of the ligand at the binding site.

Thus, the search algorithm explores different positions of ligands at the active binding site using translational and rotational degrees of freedom in the case of rigid docking, while flexible docking adds conformational degrees of freedom to translations and rotations of the ligands. To predict the correct conformation of ligands, search algorithms adopt various techniques, such as checking the chemistry and geometry of the atoms involved [DOCK 6 (Allen et al., 2015), FLEXX (Rarey et al., 1996)], genetic algorithm [GOLD (Verdonk et al., 2003)] and incremental construction (Friesner et al., 2004). Algorithms that consider ligand flexibility can be divided into three types: systematic, stochastic and deterministic (Ruiz-Tagle et al., 2018). Some software uses more than one of these approaches to obtain better results.

Systematic search algorithms exploit the degrees of freedom of the molecules, usually through their incremental construction at the binding site. Increasing the degree of freedom (rotatable bonds) increases the number of evaluations needed to be performed by the algorithm. Increasing the degree of freedom (rotary links) increases the number of evaluations required to be performed by the algorithm, causing an increase in the time required for its execution. To reduce the time it takes to execute, termination criteria are inserted that prevent the algorithm from trying solutions that are in the space known to lead to wrong solutions. DOCK 6 (Allen et al., 2015), FLEXX (Rarey et al., 1996), and Glide (Friesner et al., 2004) are examples of software that uses systematic search algorithms.

Stochastic search algorithms perform random changes in the spatial conformation of the ligand, usually changing one system degree of freedom at a time, which leads to the exploration of several possible conformations (Ruiz-Tagle et al., 2018). The main problem of stochastic algorithms is the uncertainty of converging to a good solution. For this reason, to minimize this problem, several independent executions of stochastic algorithms are usually performed. Examples of stochastic research algorithms are Monte Carlo (MC) methods used by Glide (Friesner et al., 2004) and MOE (Vilar et al., 2008) and genetic algorithms used by GOLD (Verdonk et al., 2003) and AutoDock4 (Morris et al., 2009).

During the execution of a deterministic search algorithm, the initial state is responsible for determining the movement that can be made to generate the next state, which generally must be equal to or less in the energy from the initial state. One problem with deterministic algorithms is that they are often trapped in local minima because they cannot cross barriers; there are approaches, such as increasing the simulation temperature, that can be implemented to circumvent this problem. Energy minimization methods are an example of deterministic algorithms. Molecular dynamics (MD) is also an example of a deterministic search algorithm and is used by DOCK 6 (Allen et al., 2015). However, MD computational demands are very high, and while MD promises to have better results and ensures full-system flexibility, the runtime becomes a limiting factor for simulations because structure databases can have millions of ligands and targets.

Scoring Functions

Molecular docking software uses scoring functions to estimate the force of non-covalent interactions between a ligand and molecular target using mathematical methods. A scoring function is one of the most important components in SBVS (Huang et al., 2010) as it is primarily responsible for predicting the binding affinity between a target and its ligand candidate. Thus, the scoring functions are the main reason for the success or failure of docking tools (ten Brink and Exner, 2009). Therefore, despite the wide use, the estimation of the interaction force between a ligand and molecular target remains a major challenge in VS. **Figure 6** illustrates docking using Autodock Vina between cyclooxygenase-2 (PDB ID: 4PH9) and two ligands (a) an inactive ligand and (b) celecoxib (an anti-inflammatory). Compared to the inactive ligand, celecoxib is observed to have much more interactions with the protein, which causes celecoxib to form a more stable binding in the VS. This result causes the AutoDock Vina scoring function to see a binding energy of −10.4 kcal/mol for celecoxib and −5.4 kcal/mol for the inactive compound. The ligand with the highest binding affinity to the target can be selected for further testing. Therefore, in this case celecoxib would be chosen.

In general, there are three important applications of scoring functions in molecular docking. First, they can be used to determine the ligand binding site and the conformation between a target and ligand. This approach can be used to search for allosteric sites. Second, they can be used to predict the binding affinity between a protein and ligand. Third, they can also be used in lead optimization (Li et al., 2013).

Most authors define the scoring functions as three types (Huang et al., 2010; Ferreira et al., 2015; Haga et al., 2016): force field (FF), empirical and knowledge-based. Liu and Wang (2015) define two more types of scoring functions as: machine-learning-based and hybrid methods.

The force field scoring functions are based on the intermolecular interactions between the ligand and target atoms, such as the van der Waals, electrostatic and bond stretching/bending/torsional force interactions, obtained from experimental data and in accordance with the principles of

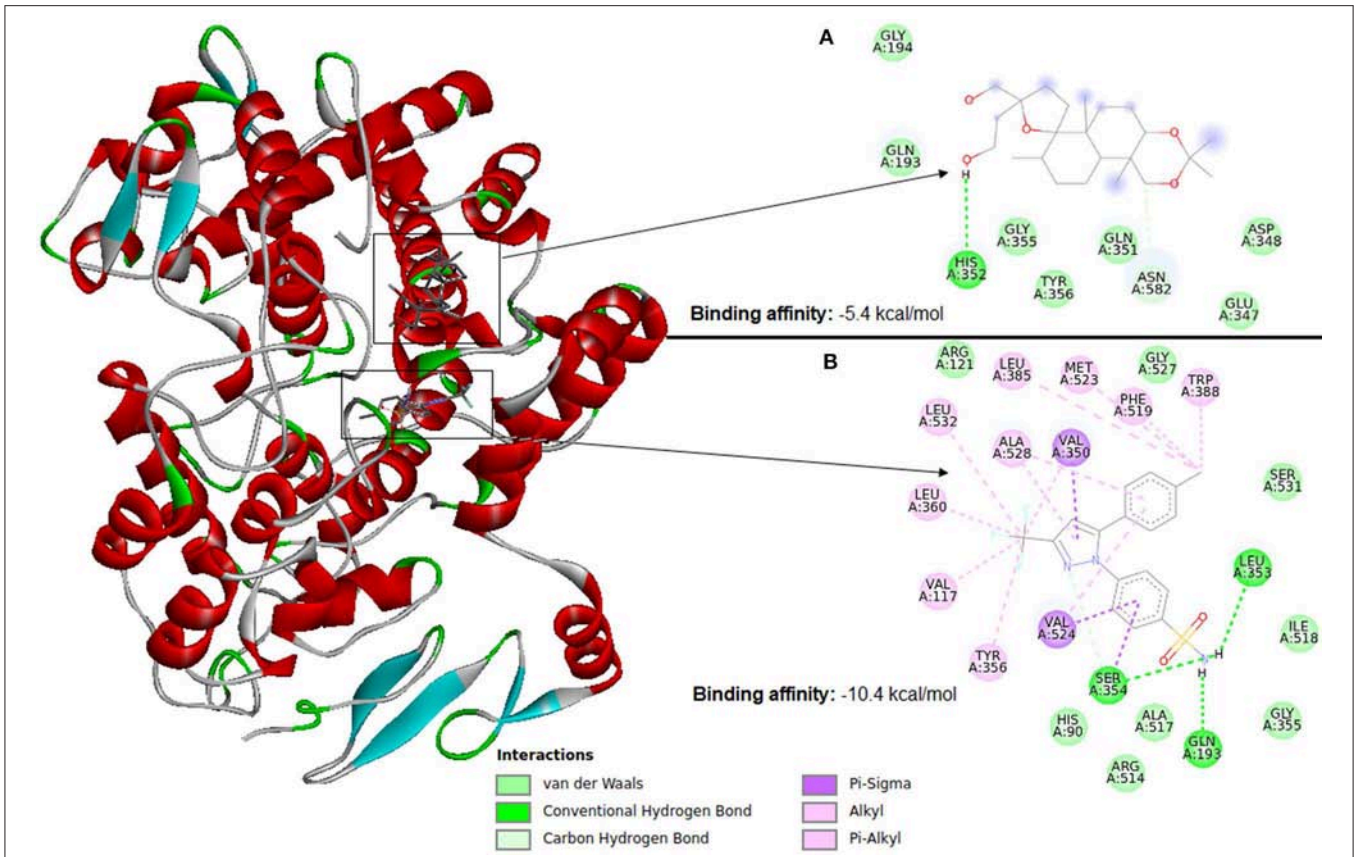


FIGURE 6 | Identification of a ligand candidate by using a typical scoring function. The hydrogens were omitted for better visualization. **(A)** Inactive ligand, **(B)** celecoxib.

molecular mechanics (Ferreira et al., 2015). Some published force-field scoring functions include the ones described in Li et al. (2015), Goldscore (Verdonk et al., 2003), and Sybyl/D-Score (Ash et al., 1997).

Empirical scoring functions estimate the binding free energy based on weighted structural parameters by adjusting the scoring functions to experimentally determine the binding constants of a set of complexes (Ferreira et al., 2015). To create an empirical scoring function, a set of data from protein-binding complexes whose affinities are known is initially used for training. A linear regression is then performed as a way of predicting the values of some variables (Huang et al., 2010). The weight constants generated by the empirical function are used as coefficients to adjust the equation terms. Each term of the function describes a type of physical event involved in the formation of the ligand-receptor complex. Thus, hydrogen bonding, ionic bonding, non-polar interactions, desolvation and entropic effects are considered. Some popular empirical, scoring functions include Glide-Score (Friesner et al., 2004), Sybyl-X/F-score (Certara, 2016) and DOCK 6 empirical force field (Allen et al., 2015).

In the knowledge-based scoring functions, the binding affinity is calculated by summing the binding interactions of the atoms of a protein and the molecular target (Ferreira et al., 2015). These functions consider statistical observations performed on large databases (Ferreira et al., 2015). The method uses pairwise energy potentials extracted from known ligand-receptor complexes to obtain a general scoring function. These methods assume that intermolecular interactions occurring near certain types of atoms or functional groups that occur more frequently are more likely to contribute favorably to the binding affinity. The final score is given as a sum of the score of all individual interactions. One example of software that uses a knowledge-based scoring function is ParaDockS (Meier et al., 2010).

In addition, machine-learning-based methods (Liu and Wang, 2015) have been considered as a fourth type of scoring function. Machine learning-based methods have gained attention for their reliable prediction (Pereira et al., 2016; Chen et al., 2018). Many researchers have used machine learning to improve SBVS algorithms, but we do not know any drugs developed after combining SBVS with machine learning. However, some researchers applied machine learning techniques to discover a new antibiotic capable of inhibiting the growth of *E. coli* bacteria (Stokes et al., 2020). These techniques have been used in quantitative structure-activity relationship (QSAR) analysis to predict various physical-chemical (for example, hydrophobicity, and stereochemistry of the molecule), biological (for example, activity and selectivity), and pharmaceutical (for example, absorption, and metabolism) properties of small molecule

compounds. In these types of scoring functions, modern QSAR analyses can be applied to derive statistical models that calculate protein-ligand binding scores. Some scoring functions of this type are NNScore 2.0 (Durrant and McCammon, 2011), RF-Score-VS (Wójcikowski et al., 2017), SFCscoreRF (Zilian and Sotriffer, 2013), SVR-KB (Li et al., 2011), SVR-EP (Li et al., 2011), ID-Score (Li et al., 2013) and CScore (Ouyang et al., 2011).

There are some hybridized scoring functions that cannot easily be classified into any of the categories listed above because they combine two or more of the previously defined scoring function types [force field (FF), empirical, knowledge based and machine-learning-based] into one scoring function. Therefore, they are called hybrid scoring functions. In general, the hybrid scoring function is a linear combination of the two or more scoring function components derived from a multiple linear regression fitting procedure (Tanchuk et al., 2016). For example, the GalaxyDock score function is a hybrid of physics-based, empirical, and knowledge-based score terms that has the advantages of each component. As a result, the performance was improved in decoy pose discrimination tests (Baek et al., 2017). A few recently published examples of this type of scoring function include the hybrid scoring function developed by Tanchuk et al. (Tanchuk et al., 2016), which combines force field machine learning scoring functions; SMoG2016 (Geng et al., 2019), which combines knowledge-based and an empirical scoring functions; GalaxyDock BP2 (Baek et al., 2017), which combines force field, empirical, and knowledge-based scoring functions and iScore (Geng et al., 2019), which combines empirical and force-field scoring functions.

Consensus Docking

In the last decade, a new technique of VS called consensus docking (CD) has been used in some studies (Park et al., 2014; Tuccinardi et al., 2014; Chermak et al., 2016; Poli et al., 2016; Aliebrahimi et al., 2017) to increase the accuracy of VS studies and to reduce the false positives obtained in VS experiments (Aliebrahimi et al., 2017).

This technique is a combination of two different approaches, in which the resultant combination is better than a single approach alone. However, Poli et al. (2016) reported that there are few studies that evaluate the possibility of combining the results from different VS methods to achieve higher success rates in VS studies.

Houston and Walkinshaw (2013) described the main reason for using this combination: the individual program may present incorrect results and these errors are mostly random. Therefore, even when two programs present different results, the combination of these results may, in principle, be much closer to the correct answer than even the best program alone. Houston and Walkinshaw also suggest that CD approaches using two different docking programs improve the precision of the predicted binding mode for any VS study. The same study also verified that a greater level of consensus in a given pose indicates a greater reliability in this result. Finally, the results presented by the authors suggest that the CD approach works as well as the best VS approaches available in the literature.

Park et al. (2014) use an approach in which they used a combination of the programs AutoDock 4.2 (Morris et al., 2009) and FlexX (Rarey et al., 1996) programs. These programs were chosen because both use different types of score functions (force field in AutoDock and empirical in FlexX). In this study, they achieved superior performance with the application of consensus docking than using each of the programs alone.

Alternatively, when using two different VS programs, there is extra time to run the two different tools and combine the results. However, Houston and Walkinshaw (2013) showed that the increased runtime may be advantageous; using AutoDock Vina (Trott and Olson, 2009) in a VS approach along with AutoDock4 (Morris et al., 2009) increased the final runtime by ∼10%. This combination is interesting given the potential gains from its use.

Therefore, the use of consensus docking is a recent technique, and although there are few papers in the literature on the subject, it seems to be a promising approach for further VS studies.

VIRTUAL DATABASES

An indispensable condition in performing VS is the availability of a 3D structure of the target protein (Cavasotto, 2011) and ligands to be docked. Some databases were created to store 3D structures of molecules. Some of the free databases include Protein Data Bank (PDB) (Berman et al., 2013), PubChem (Kim et al., 2016), ChEMBL (Bento et al., 2014), ChemSpider (Pence and Williams, 2010), Zinc (Sterling and Irwin, 2015), Brazilian Malaria Molecular Targets (BraMMT) (Nunes et al., 2019), Drugbank (Wishart et al., 2018), and Our Own Molecular Targets (OOMT) (Carregal et al., 2013). In addition, there are some commercially available databases such as the MDL Drug Data Report[1] Below we are going to present a brief explanation of each of these databases:

- Protein Data Bank (PDB) (Berman et al., 2013): PDB is the public database where three-dimensional structures of proteins, nucleic acids, and complex molecules have been deposited since 1971. The worldwide PDB organization ensures that PDB files are publicly available to the global community. It is widely used by the academic community and has grown consistently in recent years. In the last 10 years, the number of 3D structures of the PDB increased from 48,169 at the end of 2008 to 147,604 in the end of 2018, an increase of nearly 207%. This implies that in the last 10 years, almost 9,943 new structures have been added to the PDB every year, just over 27 structures per day, on average. The pace of this growth has increased. At the beginning of this decade approximately 25 new entries were added per day on average. In 2018, over 31 new structures were added per day, an average daily growth of 24% compared to 2010.
- PubChem (Kim et al., 2016): PubChem is a public database, aggregating information from smaller, more specific databases. It has more than 97 million compounds available.

[1] http://accelrys.com/products/collaborative-science/databases/bioactivity-databases/mddr.html

- ChEMBL (Bento et al., 2014): ChEMBL is a database of bioactive molecules with medicinal properties maintained by the European Institute of Bioinformatics (EBI) of the European Molecular Biology Laboratory (EMBL). Currently, it has almost 2.3 million compounds and 15.2 million known biological activities.
- Zinc (Sterling and Irwin, 2015): Zinc is a free database of commercially available compounds for VS. Zinc has more than 230 million commercially available compounds in the 3D format. Zinc is maintained by Irwin and Shoichet Laboratories of the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF).
- NatProDB (Paixão and Pita, 2016): The State University of Feira de Santana has made NatProDB available. This database stores 3D structures of the semiarid biome. The pharmacological profile of compounds from the semiarid flora have not yet been studied, which has motivated our research group to deepen the research by their molecular targets (Taranto et al., 2015).
- Our Own Molecular Target (OOMT) (Carregal et al., 2013): OOMT is a special molecular target database because it has the biological assay for all its molecular targets, and includes specific targets for cancer, dengue, and malaria. OOMT was created by a group of researchers from Federal University of São João del-Rei (UFSJ).
- Brazilian Malaria Molecular Targets (BraMMT) (Nunes et al., 2019): The BRAMMT database comprises thirty-five molecular targets for *Plasmodium falciparum* retrieved from the PDB database. This database allows *in silico* virtual high throughput screening (vHTS) experiments against a pool of *P. falciparum* molecular targets.
- Drugbank (Wishart et al., 2018): DrugBank is a database that contains comprehensive molecular information about drugs, their mechanisms, their interactions, and their targets. The database contains more than 11,900 drug entries, including nearly 2,538 FDA-approved small molecule drugs, 1,670 biotechnology (protein / peptide) drugs approved by the FDA, 129 nutraceuticals and nearly 6,000 investigational drugs.

Commercially available Databases:

- MDL Drug Data Report (MDDR) (Sci Tegic Accelrys Inc, 2019): MDDR is a commercial database built from patent databases, publications and congresses. It has more than 260,000 biologically relevant compounds and approximately 10,000 compounds are added every year.
- ChemSpider (Pence and Williams, 2010): ChemSpider is a database of chemical substances owned by the Royal Society of Chemistry. It has more than 71 million chemical structures from over 250 data sources. ChemSpider allows downloading up to 1000 structures per day. Previous contact is needed for the download of more structures, and ChemSpider is therefore not a totally free database.

VIRTUAL SCREENING ALGORITHMS

In VS, we are targeting proteins in the human body to find novel ligands that will bind to them. VS can be divided into two classes: structure-based and ligand-based. In structure-based virtual screening, a 3D structure of the target protein is known, and the goal is to identify ligands from a database of candidates that will have better affinity with the 3D structure of the target. VS can be performed using molecular docking, a computational process where ligands are moved in 3D space to find a configuration of the target and ligand that maximizes the scoring function. The ligands in the database are ranked according to their maximum score, and the best ones can be investigated further, e.g., by examining the mode and type of interaction that occurs. Additionally, VS techniques can be divided according to the algorithms used as follows:

- Machine Learning-based Algorithms
  - Artificial neural networks (ANNs) (Ashtawy and Mahapatra, 2018);
  - Support vector machines (Sengupta and Bandyopadhyay, 2014);
  - Bayesian techniques (Abdo et al., 2010);
  - Decision tree (Ho, 1998);
  - k-nearest neighbors (kNN) (Peterson et al., 2009);
  - Kohonen's SOMs and counterpropagation ANNs (Schneider et al., 2009);
  - Ensemble methods using machine learning (Korkmaz et al., 2015);
- Evolutionary Algorithms
  - Genetic algorithms (Xia et al., 2017);
  - Differential evolution (Friesner et al., 2004), Gold (Verdonk et al., 2003), Surflex (Spitzer and Jain, 2012) and FlexX (Hui-fang et al., 2010);
  - Ant colony optimization (Korb et al., 2009);
  - Tabu search (Baxter et al., 1998);
  - Particle swarm optimization (Gowthaman et al., 2015) and PSOVina (Ng et al., 2015);
- Local search such as Autodock Vina (Trott and Olson, 2009), SwissDock/EADock (Grosdidier et al., 2011) and GlamDock (Tietze and Apostolakis, 2007);
- Exhaustive search such as eHiTS (Zsoldos et al., 2007);
- Linear programming methods such as Simplex Method (Ruiz-Carmona et al., 2014);
- Systematic methods such as incremental construction used by FlexX (Rarey et al., 1996), Surflex (Spitzer and Jain, 2012), and Sybyl-X (Certara, 2016);
- Statistical methods
  - Monte Carlo (Harrison, 2010);
  - Simulated annealing (SA) (Doucet and Pelletier, 2007), Hatmal and Taha (Hatmal and Taha, 2017);
  - Conformational space annealing (CSA) (Shin et al., 2011);
- Similarity-based algorithms
  - Based on substructures (Tresadern et al., 2009);
  - Pharmacochemical (Cruz-Monteagudo et al., 2014);
  - Overlapping volumes (Leach et al., 2010);
  - Molecular interaction fields (MIFs) (Willett, 2006);

- Hybrid approach (Morris et al., 2009; Haga et al., 2016);

After performing a VS simulation, it is necessary to verify whether the quality of the generated protein-ligand complexes can represent a complex that could be reproduced in experiments. There are several methods that can perform this assessment, which will be explained in the next section.

METHODS OF EVALUATING THE QUALITY OF A SIMULATION

To verify the quality of a docking approach, some methods are used to evaluate generated complexes and to verify if the protein generated by the docking can reproduce the experimental data results of the ligand-receptor complex. The most common evaluation methods are root mean square deviation (RMSD) (Hawkins et al., 2008), receiver operating characteristic (ROC), area under the curve ROC (AUC-ROC) (Flach and Wu, 2005; Trott and Olson, 2009) enrichment factors (EFs) (Truchon and Bayly, 2007) and Boltzmann-enhanced discrimination of ROC (BEDROC) (Truchon and Bayly, 2007).

Root-Mean-Square Deviation (RMSD)

One of the aspects evaluated in docking programs is the accuracy of the generated geometry (Jain, 2008). Docking programs attempt to reproduce the conformation of the ligand-receptor complex in a crystallographic structure. The metric root-mean-square deviation (RMSD) of atomic coordinates after the ideal superposition of rigid bodies of two structures is popular. Its popularity is because it allows the quantification of the differences between two structures, and these can be structures with the same and different amino acid sequences (Sargsyan et al., 2017). RMSD is widely used to evaluate the quality of a docking process performed by a program (Ding et al., 2016). The RMSD between two structures can be calculated according to the following equation (Sargsyan et al., 2017):

$$RMSD\,(A,B) = \frac{1}{N}\sum_{i=1}^{n} d_i^2$$

where d is the distance between atom i in the two structures and N is the total number of equivalent atoms. Since the calculation of RMSD requires the same number of atoms in both structures, it is often used in the calculation of only the heavy atoms or backbone of each amino acid residue.

Using the RMSD calculation, it is possible to evaluate if a program was able to reliably reproduce a known crystallographic conformation, as well as their respective intramolecular interactions. To verify if a given program can accomplish this task, ligand-targets complexes are subjected to a redocking process. After redocking, the overlap of the crystallographic ligand with the conformation of the ligand obtained with the docking program is then performed. Then, the RMSD calculation is used to check the average distance between the corresponding atoms (usually backbone atoms).

Generally, the RMSD threshold value is 2.0 Å (Jain, 2008; Meier et al., 2010; Gowthaman et al., 2015). However, for ligands with several dihedral angles, an RMSD value of 2.5 Å is considered acceptable (De Magalhães et al., 2004). In the case of binding a large ligand, some authors generally relax this criterion (Méndez et al., 2003; Verschueren et al., 2013). For a model generated by homology modeling, evaluating the RMSD value is important, although visual inspection of the generated model is also essential.

However, RMSD has some important limitations:

- RMSD can only compare structures with the same number of atoms;
- A small perturbation in just one part of the structure can create large RMSD values, suggesting that the two structures are very different, although they are not (Carugo, 2007);
- It has also been observed that RMSD values depend on the resolution of structures that are compared (Carugo, 2003);
- RMSD does not distinguish between a structure with some very rigid regions and some very flexible regions from a molecule in which all regions are semiflexible (Sargsyan et al., 2017);

Comparing the RMSD value of large structures may be significantly distorted from the commonly used 2Å threshold (Méndez et al., 2003). Despite these limitations, RMSD remains one of the most commonly used metrics to quantify differences between structures (Sargsyan et al., 2017).

Figure 7 shows the visualization of the FCP ligand superposed with its conformation after redocking to a protein (PDB ID: 1VZK, A Thiophene Based Diamidine Forms a "Super" AT Binding Minor Groove Agent). The RMSD between the crystallographic ligand and the same ligand after the redocking using DOCK6 is 0.97 Å. In the figure below, red represents the crystallographic ligand FCP and yellow represents FCP ligand after redocking using DOCK 6.

ROC Curve and AUC

One of the great challenges of VS methods is the ability to differentiate true positive compounds (TPCs) against the target from false positive compounds (FPCs) (Awuni and Mu, 2015). Thus, it is important that VS tools have ways to assist their users in distinguishing TPCs from FPCs. The ROC curve and the area under the ROC curve (AUC-ROC) (Lätti et al., 2016) are widely used methodologies for this purpose.

TPC and decoys are used to create a ROC curve and AUC-ROC. TPCs are those with known biological activity for the molecular target of interest. Some databases, such as ChEMBL (Gaulton et al., 2012; Bento et al., 2014), allows users to search for these compounds. Alternatively, decoys are compounds that, although possessing physical properties similar to a TPC (such as molecular mass, number of rotatable bonds, and logP), have different chemical structures that make them inactive. They are generated from random molecular modifications in the structure of a TPC (Huang et al., 2006). Some databases, such as DUD-E (Mysinger et al., 2012) and Zinc (Sterling and Irwin, 2015), provide decoys for compounds of interest. DUD-E generates 50 different decoys for each TPC. The idea of using DUD-E decoys in VS is that the result of VS is more reliable if the program can

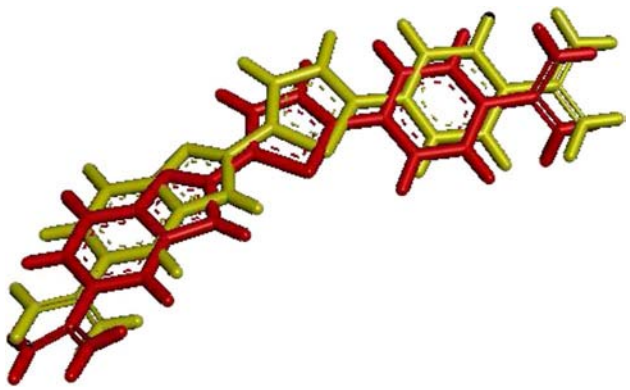


FIGURE 7 | RMSD between the ligand FCP with a protein (PDB ID: 1VZK) after redocking using DOCK6.

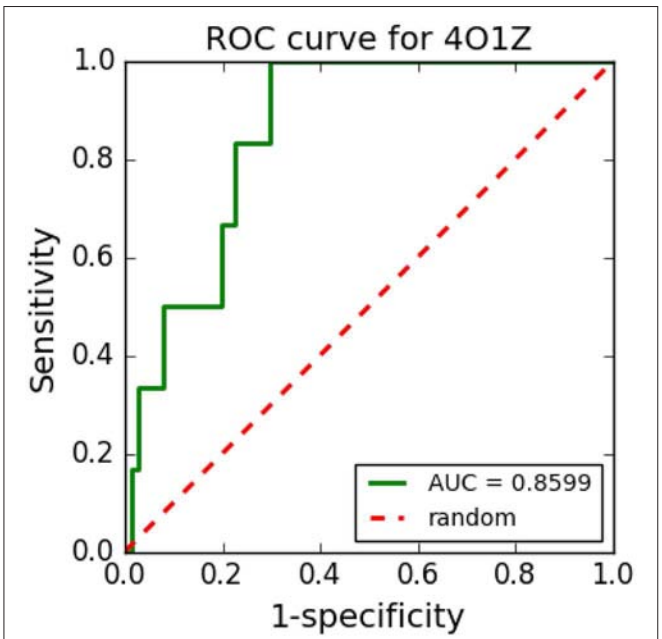


FIGURE 8 | ROC curve example.

separate TPCs from FPCs generated by DUD-E because FPCs have many TPC-like physical properties but are known to be inactive. A small number (>2) of known TPCs have to be used to calculate an AUC-ROC (Lätti et al., 2016).

After generating decoys, a VS process is performed using known TPCs and decoys against a target of interest (Yuriev and Ramsland, 2013). For each ligand-target complex, an affinity energy is then calculated. TPCs are expected to have lower affinity energy than inactive compounds. The ROC curve plots the distribution of true and false results on a graph, while AUC-ROC allows the evaluation of the probability of a result to be false. Hence, AUC-ROC reflects the probability of recovering an active compound preferentially to inactive compounds (Triballeau et al., 2005; Zhao et al., 2009), allowing verification of the sensitivity of a VS experiment in relation to its specificity. The larger the area under the curve, the better the ability to have a TPC and fewer FPC.

The AUC value can vary between 0 and 1. Hamza (Hamza et al., 2012) showed a practical way of interpreting the AUC values:

- AUC between 0.90 and 1.00: Excellent
- AUC between 0.80 and 0.90: Good
- AUC between 0.70 and 0.80: Fair
- AUC between 0.60 and 0.70: Poor
- AUC between 0.50 and 0.60: Failure

Therefore, the closer the AUC is to 1, the greater the ability of the VS tool to separate between TPCs and FPCs. AUC-ROC values close to 0.5 indicate a random process (Ogrizek et al., 2015). Acceptable values should be >0.7.

Figure 8 shows an example of an ROC curve generated in a VS performed with cyclooxygenase-1 complexed with meloxicam (PDB ID: 4O1Z) protein using five TPCs and 250 decoys. The VS tool was able to distinguish well between TPCs and FPCs with the generated ROC curve and its respective AUC, which was 0.8628.

Boltzmann-Enhanced Discrimination of ROC (BEDROC)

There is much criticism in the use of the ROC curve as a method to measure virtual screening performance because it does not highlight the best ranked active compounds that would be used in *in vitro* experiments, which is called early recognition. Thus, Tuchon and Bayly (Truchon and Bayly, 2007) proposed Boltzmann-Enhanced Discrimination of ROC (BEDROC), which uses exponential weighting to give early rankings of active compounds more weight than late rankings of active compounds. However, Nicholls (Nicholls, 2008) say that AUC-ROC and BEDROC correlate when considering virtual screening simulations, and therefore, the ROC curve is a sufficient metric for performance measurements.

Enrichment Factors (EFs)

The enrichment factor (EF) consists of the number of active compounds found in a fraction of $0 < \chi < 1$ in relation to the number of active compounds that would be found after a random search (Truchon and Bayly, 2007). EFs are often calculated against a given percentage of the database. For example, EF10% represents the value obtained when 10% of the database is screened. EFs can be defined by the following formula (1):

$$EF = \frac{\sum_1^n \delta_i}{\chi^n} \qquad \text{where } \delta_i = \begin{cases} 1, & r_i \leq \chi N \\ 0, & r_i > \chi N \end{cases} \tag{1}$$

r_i is the rank of the *i*th active compound in the list, N is the total number of compounds and n is the number corresponding to the selected compounds. The maximum value of EF is 1 / χ if $x \geq n / N$ and N / n if $\chi < n / N$. The minimum value for EF is 0.

TABLE 1 | Virtual screening software.

Software	License	Platform	Protein flexibility	Docking algorithm	Scoring function
AutoDock4 (Morris et al., 2009)	Free for academic use	Windows, Linux and Mac	Yes	Genetic algorithm Simulated annealing	Hybrid (Force-field and empirical)
Autodock Vina (Trott and Olson, 2009)	Open- source	Windows, Linux and Mac	Yes	Genetic algorithm Simulated annealing Local search Particle swarm optimization	Hybrid (Empirical and knowledge-based)
DOCK 6 (Allen et al., 2015)	Free for academic use	Windows, Linux and Mac	Yes	Shape fitting (sphere sets) Lowest energy binding	Force-Field Empirical
SwissDock/EADock DSS (Grosdidier et al., 2011)	Free for academic use	Web	No	Stochastic (Tabu search based) Local search Combination of broad and local search of the conformational space	Force-field
eHiTS (Zsoldos et al., 2007)	Freeware for academic use	Unix	No	Exhaustive search	Hybrid (Empirical and knowledge-based)
FITTED (Corbeil et al., 2007, 2008)	Commercial	Linux, Windows and Mac	Yes	Genetic algorithm	Force-field
FlexX (Rarey et al., 1996)	Commercial	Windows and Linux	No	Incremental construction	Empirical
FLIPDock (Zhao and Sanner, 2007)	Freeware for academic Use	Linux e Windows	Yes	Genetic algorithm	Force-field
Fred (McGann, 2011)	Free for academic use	Windows, Linux and Mac	No	Exhaustive search algorithm	Hybrid
GalaxyDock2 (Shin et al., 2013)	Freeware	Linux	Yes	Conformational analysis Genetic algorithm	Force-field
GeauxDock (Fang et al., 2016)	Open-source	Linux	Yes	Monte Carlo	Hybrid (Empirical and knowledge-based)
GlamDock (Tietze and Apostolakis, 2007)	Freeware	Windows, Linux and Mac	No	Monte Carlo Simulated annealing Local search Conformational analysis	Empirical
Glide (Friesner et al., 2004)	Commercial	Windows, Linux	Yes	Conformational analysis Monte Carlo sampling	Empirical
GOLD (Verdonk et al., 2003)	Commercial	Linux and Windows	Yes	Genetic algorithm	Force-field
ICM (Abagyan et al., 1994)	Commercial	Windows, Linux and Mac	Yes	Monte Carlo minimization	Force-field
iGEMDOCK/GEMDOCK (Hsu et al., 2011)	Freeware	Windows and Linux	Yes	Genetic algorithm	Empirical
LigandFit (Montes et al., 2007)	Commercial	Linux	Yes	Monte Carlo	Force-field
LigDockCSA (Shin et al., 2011)	–	–	Yes	Conformational space annealing Global optimization	Hybrid (Empirical and Force-field)
MOE (Vilar et al., 2008)	Commercial	Windows, Linux and Mac	Yes	Conformational analysis	Empirical, Force-field
ParaDockS (Meier et al., 2010)	Freeware	Linux	No	Genetic algorithm	Hybrid (Knowledge-based and empirical)
rDOCK (Ruiz-Carmona et al., 2014)	Open-source	Linux	Yes	Genetic algorithm, Monte CarloSimplex minimization	Hybrid (Empirical and force-field)
SLIDE (Schnecke and Kuhn, 2000)	Free for academic use	Linux	Yes	Conformational analysis	Empirical
Surflex (Spitzer and Jain, 2012)	Commercial	Windows, Linux and Mac	Yes	Incremental xonstruction	Empirical
Sybyl-X (Certara, 2016)	Commercial	Windows	Yes	Incremental construction	Force field
vLifeDock (Chopade, 2015)	Commercial	Windows, Linux and Mac	Yes	Genetic algorithm	Empirical

EF is quite simple, but it has some disadvantages. The EF, in addition to depending on the value set for χ, depends on the number of true positives and true negatives, which makes it another measure of experiment performance rather than measuring method performance (Nicholls, 2011). Another disadvantage of EF is that it weighs active compounds equally within the cutoff, so it is not possible to distinguish the best ranking algorithm in which all active compounds are ranked at the beginning of the ordered list of a worse algorithm and they are sorted immediately before the cutoff value [saturation effect (Lopes et al., 2017)].

The relative enrichment factor (REF) proposed by von Korff et al. (2009) eliminates the problem associated with the saturation effect by normalizing the EF by the maximum possible enrichment. Consequently, REF has well-defined boundaries and is less subject to the saturation effect.

VS SOFTWARE PROGRAMS

There are several VS software programs using different docking algorithms that make a VS process easier for the researchers to execute by avoiding the need to have advanced knowledge of computer science and on how to implement the algorithms used in this task. In this regard, VS software can act as a possible cost reducer, since they function as filters that select from a database with thousands of molecules that are more likely to present biological activity against a target of interest. VS programs measure the affinity energy of a small molecule (ligand) to a molecular target of interest to determine the interaction energy of the resulting complex (Carregal et al., 2017).

Table 1 summarizes the main characteristics of the most used software in VS. The first column contains the software used and its reference. The second column contains the type of software license: free for academic use, freeware, open-source, or commercial. The free for academic use license indicates that the software in question can be used for teaching and research in the academic world without a fee. However, it implies that the software has restrictions for commercial use. A freeware license indicates that the software is free. Thus, users can use it without a fee, and all the functions of the program are available to be used without any restrictions. An open-source license indicates that the software source code is accessible so users can study, change, and distribute the software to anyone and for any purpose. Software developed under a commercial license indicates that it is designed and developed for a commercial purpose. Thus, in general, it is necessary to pay some licensing fee for its use. The third column indicates on which platforms the software can be used (Windows, Linux, or Mac). The next column indicates whether or not the software may consider protein flexibility during anchoring. The docking algorithm column lists the algorithms used by the software to perform the docking. The sixth column, called the scoring function, indicates which scoring functions are used by the software.

FINAL CONSIDERATIONS

CADD has been used to improve the drug development process. In the past, the discovery of new drugs was often conducted through the empirical observation of the effect of natural products in known diseases. Thus, several possible drug candidates were tested without efficacy, and thereby wasted resources. The use of CADD allows for improving the development of new biologically active compounds and decreasing the time and cost for the development of a new drug. Thus, the emergence of SBVS has improved the drug discovery process and was established as one of the most promising *in silico* techniques for drug design.

This review verified that CADD approaches can contribute to many stages of the drug discovery process, notably to perform a search for active compounds by VS.

The use of techniques, such as SBVS, has limitations, such as the possibility of generating false positives and correct ranking of ligands docked. Moreover, there are several CADD methods and it is possible to obtain different results for the same input in different software. However, reducing the time and cost of the new drug development process as well as the constant improvement of existing docking tools indicates that CADD techniques will be one of the most promising techniques in the drug discovery process over the next years.

In the last decade, many studies have applied artificial intelligence in CADD to obtain more accurate models. Thus, most studies and future innovations will benefit from the application of AI in CADD.

Finally, the use of CADD tools requires a variety of expertise of researchers to perform all of the steps of the process, such as selecting and preparing targets and ligands, analyzing the results and having broad knowledge of computation, chemistry and biology. Thus, the researcher's background is important for the selection of new hits and to enrich high throughput experiments.

AUTHOR CONTRIBUTIONS

All authors of this review have made a great contribution to the work. All authors wrote the paper and approved the final version.

FUNDING

Funding sources for this project include FAPEMIG (APQ-02742-17 and APQ-00557-14), CNPq (449984/2014-1), CNPq Universal (426261/2018-6), and UFSJ/PPGBiotec. AT and LA are grateful to CNPq (305117/2017-3) and CAPES for their research fellowships.

ACKNOWLEDGMENTS

The authors would like to thank the Federal Univesity of São João del-Rei (UFSJ) and the Federal Center for Technological Education of Minas Gerais (CEFET-MG) for providing the physical infrastructure.

REFERENCES

Abagyan, R., Totrov, M., and Kuznetsov, D. (1994). ICM - a new method for protein modeling and design: applications to docking and structure prediction from distorted native conformation. *J. Comput. Chem.* 15, 488–506. doi: 10.1002/jcc.540150503

Abdo, A., Chen, B., Mueller, C., Salim, N., and Willett, P. (2010). Ligand-based virtual screening using bayesian networks. *J. Chem. Inf. Model.* 50, 1012–1020. doi: 10.1021/ci100090p

Aliebrahimi, S., Montasser Kouhsari, S., Ostad, S. N., Arab, S. S., and Karami, L. (2017). Identification of phytochemicals targeting c-Met kinase domain using consensus docking and molecular dynamics simulation studies. *Cell Biochem. Biophys.* 76, 135–145. doi: 10.1007/s12013-017-0821-6

Allen, W. J., Balius, T. E., Mukherjee, S., Brozell, S. R., Moustakas, D. T., Lang, P. T., et al. (2015). DOCK 6: impact of new features and current docking performance. *J. Comput. Chem.* 36, 1132–1156. doi: 10.1002/jcc.23905

Arrowsmith, J. (2012). A decade of change. *Nat. Rev. Drug Discov.* 11, 17–18. doi: 10.1038/nrd3630

Ash, S., Cline, M. A., Homer, R. W., Hurst, T., and Smith, G. B. (1997). SYBYL line notation (SLN): a versatile language for chemical structure representation. *J. Chem. Inf. Model.* 37, 71–79. doi: 10.1021/ci960109j

Ashtawy, H. M., and Mahapatra, N. R. (2018). Task-specific scoring functions for predicting ligand binding poses and affinity and for screening enrichment. *J. Chem. Inf. Model.* 58, 119–133. doi: 10.1021/acs.jcim.7b00309

Awuni, Y., and Mu, Y. (2015). Reduction of false positives in structure-based virtual screening when receptor plasticity is considered. *Molecules* 20, 5152–5164. doi: 10.3390/molecules20035152

Baek, M., Shin, W-H., Chung, H. W., and Seok, C. (2017). GalaxyDock BP2 score: a hybrid scoring function for accurate protein–ligand docking. *J. Comput. Aided. Mol. Des.* 31, 653–666. doi: 10.1007/s10822-017-0030-9

Banegas-Luna, A.-J., Cerón-Carrasco, J. P., and Pérez-Sánchez, H. (2018). A review of ligand-based virtual screening web tools and screening algorithms in large molecular databases in the age of big data. *Future Med. Chem.* 10, 2641–2658. doi: 10.4155/fmc-2018-0076

Baxter, C. A., Murray, C. W., Clark, D. E., Westhead, D. R., and Eldridge, M. D. (1998). Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins Struct. Funct. Genet.* 33, 367–382. doi: 10.1002/(SICI)1097-0134(19981115)33:3<367::AID-PROT6>3.0.CO;2-W

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., et al. (2014). The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42, D1083–D1090. doi: 10.1093/nar/gkt1031

Berman, H. M., Kleywegt, G. J., Nakamura, H., and Markley, J. L. (2013). The future of the protein data bank. *Biopolymers* 99, 218–222. doi: 10.1002/bip.22132

Carpenter, K. A., Cohen, D. S., Jarrell, J. T., and Huang, X. (2018). Deep learning and virtual drug screening. *Future Med. Chem.* 10, 2557–2567. doi: 10.4155/fmc-2018-0314

Carregal, A. P., Maciel, F. V., Carregal, J. B., dos Reis Santos, B., da Silva, A. M., and Taranto, A. G. (2017). Docking-based virtual screening of Brazilian natural compounds using the OOMT as the pharmacological target database. *J. Mol. Model.* 23:111. doi: 10.1007/s00894-017-3253-8

Carregal, A. P. Jr., Comar, M., and Taranto, A. G. (2013). Our own molecular targets data bank (OOMT). *Biochem. Biotechnol. Reports* 2, 14–16. doi: 10.5433/2316-5200.2013v2n2espp14

Carugo, O. (2003). How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared. *J. Appl. Crystallogr.* 36, 125–128. doi: 10.1107/S0021889802020502

Carugo, O. (2007). Statistical validation of the root-mean-square-distance, a measure of protein structural proximity. *Protein Eng. Des. Sel.* 20, 33–37. doi: 10.1093/protein/gzl051

Cavasotto, C. N. (2011). Homology models in docking and high-throughput docking. *Curr. Top. Med. Chem.* 11, 1528–1534. doi: 10.2174/156802611795860951

Certara (2016). *SYBYL-X Suite 2.1*. Available online at: https://support.certara.com/software/molecular-modeling-and-simulation/sybyl-x/ (accessed August 23, 2017).

Chen, R., Liu, X., Jin, S., Lin, J., and Liu, J. (2018). Machine learning for drug-target interaction prediction. *Molecules* 23:2205. doi: 10.3390/molecules23092208

Chermak, E., De Donato, R., Lensink, M. F., Petta, A., Serra, L., Scarano, V., et al. (2016). Introducing a clustering step in a consensus approach for the scoring of protein-protein docking models. *PLoS ONE* 11:e0166460. doi: 10.1371/journal.pone.0166460

Chopade, A. R. (2015). Molecular docking studies of phytocompounds from phyllanthus species as potential chronic pain modulators. *Sci. Pharm.* 83, 243–267. doi: 10.3797/scipharm.1408-10

Corbeil, C. R., Englebienne, P., and Moitessier, N. (2007). Docking ligands into flexible and solvated macromolecules. 1. development and validation of FITTED 1.0. *J. Chem. Inf. Model.* 47, 435–449. doi: 10.1021/ci6002637

Corbeil, C. R., Englebienne, P., Yannopoulos, C. G., Chan, L., Das, S. K., Bilimoria, D., et al. (2008). Docking ligands into flexible and solvated macromolecules 2 development and application of F ITTED 1. 5 to the virtual screening of potential HCV polymerase inhibitors. *J. Chem. Inf. Model.*, 902–909. doi: 10.1021/ci700398h

Cruz-Monteagudo, M., Medina-Franco, J. L., Pérez-Castillo, Y., Nicolotti, O., Cordeiro, M. N. D. S., and Borges, F. (2014). Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discov. Today* 19, 1069–1080. doi: 10.1016/j.drudis.2014.02.003

Das, P. S., and Saha, P. (2017). A review on computer aided drug design in drug discovery. *World J. Pharm. Pharm. Sci.* 6, 279–291. doi: 10.20959/wjpps20177-9450

De Magalhães, C. S., Barbosa, H. J. C., and Dardenne, L. E. (2004). "Selection-insertion schemes in genetic algorithms for the flexible ligand docking problem," In: *Genetic and Evolutionary Computation – GECCO 2004. GECCO 2004. Lecture Notes in Computer Science*, ed K. Deb (Berlin, Heidelberg: Springer), 368–379. doi: 10.1007/978-3-540-24854-5_38

Devi, R. V., Sathya, S. S., and Coumar, M. S. (2015). Evolutionary algorithms for *de novo* drug design - a survey. *Appl. Soft Comput. J.* 27, 543–552. doi: 10.1016/j.asoc.2014.09.042

Ding, Y., Fang, Y., Moreno, J., Ramanujam, J., Jarrell, M., and Brylinski, M. (2016). Assessing the similarity of ligand binding conformations with the contact mode score. *Comput. Biol. Chem.* 64, 403–413. doi: 10.1016/j.compbiolchem.2016.08.007

Doucet, N., and Pelletier, J. N. (2007). Simulated annealing exploration of an active-site tyrosine in TEM-1β-lactamase suggests the existence of alternate conformations. *Proteins Struct. Funct. Bioinforma.* 69, 340–348. doi: 10.1002/prot.21485

Durrant, J. D., and McCammon, J. A. (2011). NNScore 2.0: a neural-network receptor-ligand scoring function. *J. Chem. Inf. Model.* 51, 2897–2903. doi: 10.1021/ci2003889

Dutkiewicz, Z., and Mikstacka, R. (2018). Structure-based drug design for cytochrome P450 family 1 inhibitors. *Bioinorg. Chem. Appl.* 2018:3924608. doi: 10.1155/2018/3924608

Fang, Y., Ding, Y., Feinstein, W. P., Koppelman, D. M., Moreno, J., Jarrell, M., et al. (2016). Geauxdock: accelerating structure-based virtual screening with heterogeneous computing. *PLoS ONE* 11:e0158898. doi: 10.1371/journal.pone.0158898

Ferreira, L., dos Santos, R., Oliva, G., and Andricopulo, A. (2015). Molecular docking and structure-based drug design strategies. *Molecules* 20, 13384–13421. doi: 10.3390/molecules200713384

Ferreira, R. S., Glaucius, O., and Andricopulo, A. D. (2011). Integração das técnicas de triagem virtual e triagem biológica automatizada em alta escala: oportunidades e desafios em P&D de fármacos. *Quim. Nova* 34, 1770–1778. doi: 10.1590/S0100-40422011001000010

Flach, P. A., and Wu, S. (2005). Repairing concavities in ROC curves. *IJCAI Int. Joint Conf. Arti. Intell.* 702–707. doi: 10.5555/1642293.1642406

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.* 47, 1739–1749. doi: 10.1021/jm0306430

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107. doi: 10.1093/nar/gkr777

Geng, C., Jung, Y., Renaud, N., Honavar, V., Bonvin, A. M. J. J., and Xue, L. C. (2019). iScore: a novel graph kernel-based function for scoring protein-protein docking models. *Bioinformatics* 36, 112–121. doi: 10.1101/498584

Gowthaman, R., Lyskov, S., and Karanicolas, J. (2015). DARC 2.0: improved docking and virtual screening at protein interaction sites. *PLoS ONE* 10:e0131612. doi: 10.1371/journal.pone.0131612

Grosdidier, A., Zoete, V., and Michielin, O. (2011). SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res.* 39, 270–277. doi: 10.1093/nar/gkr366

Haga, J. H., Ichikawa, K., and Date, S. (2016). Virtual screening techniques and current computational infrastructures. *Curr. Pharm. Des.* 22, 3576–3584. doi: 10.2174/1381612822666160414142530

Hamza, A., Wei, N-N., and Zhan, C-G. (2012). Ligand-based virtual screening approach using a new scoring function. *J. Chem. Inf. Model.* 52, 963–974. doi: 10.1021/ci200617d

Harrison, R. L. (2010). Introduction to monte carlo simulation. *AIP Conf. Proc.* 1204, 17–21. doi: 10.1063/1.3295638

Hatmal, M. M., and Taha, M. O. (2017). Simulated annealing molecular dynamics and ligand–receptor contacts analysis for pharmacophore modeling. *Future Med. Chem.* 9, 1141–1159. doi: 10.4155/fmc-2017-0061

Hawkins, P. C. D., Warren, G. L., Skillman, A. G., and Nicholls, A. (2008). How to do an evaluation: pitfalls and traps. *J. Comput. Aided. Mol. Des.* 22, 179–190. doi: 10.1007/s10822-007-9166-3

Hillisch, A., Pineda, L. F., and Hilgenfeld, R. (2004). Utility of homology models in the drug discovery process. *Drug Discov. Today* 9, 659–669. doi: 10.1016/S1359-6446(04)03196-4

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 832–844. doi: 10.1109/34.709601

Horvath, D. (1997). A virtual screening approach applied to the search for trypanothione reductase inhibitors. *J. Med. Chem.* 2623, 2412–2423. doi: 10.1021/jm9603781

Houston, D. R., and Walkinshaw, M. D. (2013). Consensus docking: improving the reliability of docking in a virtual screening context. *J. Chem. Inf. Model.* 53, 384–390. doi: 10.1021/ci300399w

Hsu, K-C., Chen, Y-F., Lin, S-R., and Yang, J-M. (2011). iGEMDOCK: a graphical environment of enhancing GEMDOCK using pharmacological interactions and post-screening analysis. *BMC Bioinformatics* 12(Suppl. 1):S33. doi: 10.1186/1471-2105-12-S1-S33

Huang, N., Shoichet, B. K., and Irwin, J. J. (2006). Benchmarking sets for molecular docking. *J. Med. Chem.* 49, 6789–6801. doi: 10.1021/jm0608356

Huang, S-Y., Grinter, S. Z., and Zou, X. (2010). Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.* 12, 12899–12908. doi: 10.1039/c0cp00151a

Hui-fang, L., Qing, S., Jian, Z., and Wei, F. (2010). Evaluation of various inverse docking schemes in multiple targets identification. *J. Mol. Graph. Model.* 29, 326–330. doi: 10.1016/j.jmgm.2010.09.004

Jain, A. N. (2008). Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput. Aided. Mol. Des.* 22, 201–212. doi: 10.1007/s10822-007-9151-x

Kapetanovic, I. M. (2008). Computer aided darug discovery and development: *in silico*-chemico-biological approach. *Chem. Biol. Interact.* 171, 165–176. doi: 10.1016/j.cbi.2006.12.006

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., et al. (2016). PubChem substance and compound databases. *Nucleic Acids Res.* 44, D1202–D1213. doi: 10.1093/nar/gkv951

Korb, O., Stutzle, T., and Exner, T. E. (2009). PLANTS: application of ant colony optimization to structure-based drug design. *Theor. Comput. Sci.* 49, 84–96. doi: 10.1007/11839088_22

Korkmaz, S., Zararsiz, G., and Goksuluk, D. (2015). MLViS: a web tool for machine learning-based virtual screening in early-phase of drug discovery and development. *PLoS ONE* 10:e0124600. doi: 10.1371/journal.pone.0124600

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* 161, 269–288. doi: 10.1016/0022-2836(82)90153-X

Lätti, S., Niinivehmas, S., and Pentikäinen, O. T. (2016). Rocker: open source, easy-to-use tool for AUC and enrichment calculations and ROC visualization. *J. Cheminform.* 8:45. doi: 10.1186/s13321-016-0158-y

Leach, A. R., Gillet, V. J., Lewis, R. A., and Taylor, R. (2010). Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.* 53, 539–558. doi: 10.1021/jm900817u

Leelananda, S. P., and Lindert, S. (2016). Computational methods in drug discovery. *Beilstein J. Org. Chem.* 12, 2694–2718. doi: 10.3762/bjoc.12.267

Leisinger, K. M., Garabedian, L. F., and Wagner, A. K. (2012). Improving access to medicines in low and middle income countries: corporate responsibilities in context. *South. Med Rev.* 5, 3–8.

Li, G. B., Yang, L. L., Wang, W. J., Li, L. L., and Yang, S. Y. (2013). ID-score: a new empirical scoring function based on a comprehensive set of descriptors related to protein-ligand interactions. *J. Chem. Inf. Model.* 53, 592–600. doi: 10.1021/ci300493w

Li, L., Wang, B., and Meroueh, S. O. (2011). Support vector regression scoring of receptor-ligand complexes for rank-ordering and virtual screening of chemical libraries. *J. Chem. Inf. Model.* 51, 2132–2138. doi: 10.1021/ci200078f

Li, Z., Gu, J., Zhuang, H., Kang, L., Zhao, X., and Guo, Q. (2015). Adaptive molecular docking method based on information entropy genetic algorithm. *Appl. Soft Comput.* 26, 299–302. doi: 10.1016/j.asoc.2014.10.008

Lima, A. N., Philot, E. A., Trossini, G. H. G., Scott, L. P. B., Maltarollo, V. G., and Honorio, K. M. (2016). Use of machine learning approaches for novel drug discovery. *Expert Opin. Drug Discov.* 11, 225–239. doi: 10.1517/17460441.2016.1146250

Lionta, E., Spyrou, G., Vassilatis, D., and Cournia, Z. (2014). Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr. Top. Med. Chem.* 14, 1923–1938. doi: 10.2174/1568026614666140929124445

Liu, J., and Wang, R. (2015). On classification of current scoring functions. *J. Chem. Inf. Model.* 55, 475–482. doi: 10.1021/ci500731a

Liu, S., Alnammi, M., Ericksen, S. S., Voter, A. F., Ananiev, G. E., Keck, J. L., et al. (2018). Practical model selection for prospective virtual screening. *J. Chem. Inf. Model. 59, 282–293.* doi: 10.1101/337956

Lopes, J. C. D., Dos Santos, F. M., Martins-José, A., Augustyns, K., and De Winter, H. (2017). The power metric: a new statistically robust enrichment-type metric for virtual screening applications with early recovery capability. *J. Cheminform.* 9:7. doi: 10.1186/s13321-016-0189-4

Ma, X., Jia, J., Zhu, F., Xue, Y., Li, Z., and Chen, Y. (2009). Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. *Comb. Chem. High Throughput Screen.* 12, 344–357. doi: 10.2174/138620709788167944

Maia, E. H. B., Campos, V. A., dos Reis Santos, B., Costa, M. S., Lima, I. G., Greco, S. J., et al. (2017). Octopus: a platform for the virtual high-throughput screening of a pool of compounds against a set of molecular targets. *J. Mol. Model.* 23:26. doi: 10.1007/s00894-016-3184-9

Maithri, G, Manasa, B, Vani, S. S., Narendra, A, and Harshita, T (2016). Computational drug design and molecular dynamic studies-a review. *Int. J. Biomed. Data Min.* 6:123. doi: 10.4172/2090-4924.1000123

Martin, Y. C. (1992). 3D database searching in drug design. *J. Med. Chem.* 35, 2145–2154. doi: 10.1021/jm00090a001

McGann, M. (2011). FRED pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* 51, 578–596. doi: 10.1021/ci100436p

Meier, R., Pippel, M., Brandt, F., Sippl, W., and Baldauf, C. (2010). ParaDockS: a framework for molecular docking with population-based metaheuristics. *J. Chem. Inf. Model.* 50, 879–889. doi: 10.1021/ci900467x

Méndez, R., Leplae, R., De Maria, L., and Wodak, S. J. (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins Struct. Funct. Genet.* 52, 51–67. doi: 10.1002/prot.10393

Meng, X-Y., Zhang, H-X., Mezei, M., and Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. *Curr. Comput. Aided. Drug Des.* 7, 146–157. doi: 10.2174/157340911795677602

Merluzzi, V., Hargrave, K., Labadia, M., Grozinger, K., Skoog, M., Wu, J., et al. (1990). Inhibition of HIV-1 replication by a nonnucleoside reverse transcriptase inhibitor. *Science* 250, 1411–1413. doi: 10.1126/science.1701568

Montes, M., Miteva, M. A., and Villoutreix, B. O. (2007). Structure-based virtual ligand screening with LigandFit: pose prediction and enrichment of compound collections. *Proteins Struct. Funct. Bioinforma* 68, 712–725. doi: 10.1002/prot.21405

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* 30, 2785–2791. doi: 10.1002/jcc.21256

Mugumbate, G., Mendes, V., Blaszczyk, M., Sabbah, M., Papadatos, G., Lelievre, J., et al. (2017). Target identification of *Mycobacterium tuberculosis* phenotypic hits using a concerted chemogenomic, biophysical, and structural approach. *Front. Pharmacol.* 8:681. doi: 10.3389/fphar.2017.00681

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55, 6582–6594. doi: 10.1021/jm300687e

Ng, M. C. K., Fong, S., and Siu, S. W. I. (2015). PSOVina: the hybrid particle swarm optimization algorithm for protein-ligand docking. *J. Bioinform. Comput. Biol.* 13:1541007. doi: 10.1142/S0219720015410073

Nicholls, A. (2008). What do we know and when do we know it? *J. Comput. Aided. Mol. Des.* 22, 239–255. doi: 10.1007/s10822-008-9170-2

Nicholls, A. (2011). What do we know?: Simple statistical techniques that help. *Chemoinform. Comput. Chem. Biol.* 672, 531–581. doi: 10.1007/978-1-60761-839-3_22

Nunes, R. R., da Fonseca, A. L., Pinto, A. C. D. S., Maia, E. H. B., da Silva, A. M., Varotti, F. de P., et al. (2019). Brazilian malaria molecular targets (BraMMT): selected receptors for virtual high-throughput screening experiments. *Mem. Inst. Oswaldo Cruz.* 114, 1–10. doi: 10.1590/0074-02760180465

Oglic, D., Oatley, S. A., Macdonald, S. J. F., Mcinally, T., Garnett, R., Hirst, J. D., et al. (2018). Active search for computer-aided drug design. *Mol. Inform.* 37:1700130. doi: 10.1002/minf.201700130

Ogrizek, M., Turk, S., Lešnik, S., Sosic, I., Hodošcek, M., Mirković, B., et al. (2015). Molecular dynamics to enhance structure-based virtual screening on cathepsin B. *J. Comput. Aided. Mol. Des.* 29, 707–712. doi: 10.1007/s10822-015-9847-2

Ouyang, X., Handoko, S. D., and Kwoh, C. K. (2011). Cscore : a simple yet effective scoring function for protein – ligand binding affinity prediction using modified cmac learning architecture. *J. Bioinform. Comput. Biol.* 9(Suppl. 1), 1–14. doi: 10.1142/S021972001100577X

Paixão, V. G., and Pita, S. S. R. (2016). Virtual Screening applied to search of inhibitors of trypanosoma cruzi trypanothione reductase employing the Natural Products Database from Bahia state (NatProDB). *Rev. Virtual Química* 8, 1289–1310. doi: 10.21577/1984-6835.20160093

Park, H., Eom, J. W., and Kim, Y. H. (2014). Consensus scoring approach to identify the inhibitors of AMP-activated protein kinase α2 with virtual screening. *J. Chem. Inf. Model.* 54, 2139–2146. doi: 10.1021/ci500214e

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., et al. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* 9, 203–214. doi: 10.1038/nrd3078

Pence, H. E., and Williams, A. (2010). ChemSpider: an online chemical information resource. *J. Chem. Educ.* 87, 1123–1124. doi: 10.1021/ed100697w

Pereira, J. C., Caffarena, E. R., and Dos Santos, C. N. (2016). Boosting docking-based virtual screening with deep learning. *J. Chem. Inf. Model.* 56, 2495–2506. doi: 10.1021/acs.jcim.6b00355

Peterson, Y. K., Wang, X. S., Casey, P. J., and Tropsha, A. (2009). The discovery of geranylgeranyltransferase-i inhibitors with novel scaffolds by the means of quantitative structure-activity relationship modeling, virtual screening, and experimental validation. *J. Med. Chem.* 52, 83–88. doi: 10.1021/jm8013772

Poli, G., Martinelli, A., and Tuccinardi, T. (2016). Reliability analysis and optimization of the consensus docking approach for the development of virtual screening studies. *J. Enzyme Inhib. Med. Chem.* 31, 167–173. doi: 10.1080/14756366.2016.1193736

Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* 261, 470–489. doi: 10.1006/jmbi.1996.0477

Ripphausen, P., Nisius, B., and Bajorath, J. J. (2011). State-of-the-art in ligand-based virtual screening. *Drug Discov. Today* 16, 372–376. doi: 10.1016/j.drudis.2011.02.011

Ruiz-Carmona, S., Alvarez-Garcia, D., Foloppe, N., Garmendia-Doval, A. B., Juhos, S., Schmidtke, P., et al. (2014). rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS Comput. Biol.* 10:e1003571. doi: 10.1371/journal.pcbi.1003571

Ruiz-Tagle, B., Villalobos-Cid, M., Dorn, M., and Inostroza-Ponta, M. (2018). "Evaluating the use of local search strategies for a memetic algorithm for the protein-ligand docking problem," *2017 36th International Conference of the Chilean Computer Science Society* (SCCC) (Arica), 1–12. doi: 10.1109/SCCC.2017.8405141

Sargsyan, K., Grauffel, C., and Lim, C. (2017). How molecular size impacts RMSD applications in molecular dynamics simulations. *J. Chem. Theory Comput.* 13, 1518–1524. doi: 10.1021/acs.jctc.7b00028

Schnecke, V., and Kuhn, L. A. (2000). Virtual screening with solvation and ligand-induced complementarity. *Perspect. Drug Discov. Des.* 20, 171–190. doi: 10.1023/A:1008737207775

Schneider, P., Tanrikulu, Y., and Schneider, G. (2009). Self-organizing maps in drug discovery: compound library design, scaffold-hopping, repurposing. *Curr. Med. Chem.* 16, 258–266. doi: 10.2174/092986709787002655

Sci Tegic Accelrys Inc (2019). *The MDL Drug Data Report (MDDR) database*. Available online at: http://accelrys.com/products/collaborative-science/databases/bioactivity-databases/mddr.html (accessed March 22, 2019).

Sengupta, S., and Bandyopadhyay, S. (2014). Application of support vector machines in virtual screening. *Int. J. Comput. Biol.* 1, 56–62. doi: 10.34040/IJCB.1.1.2012.20

Shin, W. H., Heo, L., Lee, J., Ko, J., Seok, C., and Lee, J. (2011). LigDockCSA: protein-ligand docking using conformational space annealing. *J. Comput. Chem.* 32, 3226–3232. doi: 10.1002/jcc.21905

Shin, W. H., Kim, J. K., Kim, D. S., and Seok, C. (2013). GalaxyDock2: protein-ligand docking using beta-complex and global optimization. *J. Comput. Chem.* 34, 2647–2656. doi: 10.1002/jcc.23438

Sliwoski, G., Kothiwale, S., Meiler, J., and Edward, W., Lowe, J. (2013). Computational methods in drug discovery. *Pharmacol. Rev.* 66, 334–395. doi: 10.1124/pr.112.007336

Spitzer, R., and Jain, A. N. (2012). Surflex-Dock: docking benchmarks and real-world application. *J. Comput. Aided. Mol. Des.* 26, 687–699. doi: 10.1007/s10822-011-9533-y

Sridhar, D. (2008). Improving access to essential medicines: how health concerns can be prioritised in the global governance system. *Public Health Ethics* 1, 83–88. doi: 10.1093/phe/phn012

Sterling, T., and Irwin, J. J. (2015). ZINC15–ligand discovery for everyone. *J. Chem. Inf. Model.* 55, 2324–2337. doi: 10.1021/acs.jcim.5b00559

Stevens, H., and Huys, I. (2017). Innovative approaches to increase access to medicines in developing countries. *Front. Med.* 4:218. doi: 10.3389/fmed.2017.00218

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., et al. (2020). A deep learning approach to antibiotic discovery. *Cell* 180, 688-702.e13. doi: 10.1016/j.cell.2020.01.021

Surabhi, S., and Singh, B. (2018). Computer aided drug design: an overview. *J. Drug Deliv. Ther.* 8, 504–509. doi: 10.22270/jddt.v8i5.1894

Talele, T. T., Khedkar, S. A., and Rigby, A. C. (2010). Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Curr. Top. Med. Chem.* 10, 127–41. doi: 10.2174/156802610790232251

Tanchuk, V. Y., Tanin, V. O., Vovk, A. I., and Poda, G. (2016). A new, improved hybrid scoring function for molecular docking and scoring based on AutoDock and AutoDock vina. *Chem. Biol. Drug Des.* 87, 618–625. doi: 10.1111/cbdd.12697

Taranto, A. G., dos R., Santos, B., Costa, M. S., Campos, V. A., Lima, I. G., et al. (2015). "Octopus: a virtual high thoughput screening plataform for multi-compouds and targets," in: *XVIIISimpósio Brasileiro de Química Teórica* (Pirenópolis - GO -Brasil: Editora da UnB), 266.

ten Brink, T., and Exner, T. E. (2009). Influence of protonation, tautomeric, and stereoisomeric states on protein–ligand docking results. *J. Chem. Inf. Model.* 49, 1535–1546. doi: 10.1021/ci800420z

Tietze, S., and Apostolakis, J. (2007). GlamDock: development and validation of a new docking tool on several thousand protein-ligand complexes. *J. Chem. Inf. Model.* 47, 1657–1672. doi: 10.1021/ci7001236

Tresadern, G., Bemporad, D., and Howe, T. (2009). A comparison of ligand based virtual screening methods and application to corticotropin releasing factor 1 receptor. *J. Mol. Graph. Model.* 27, 860–870. doi: 10.1016/j.jmgm.2009.01.003

Triballeau, N., Acher, F., Brabet, I., Pin, J., and Bertrand, H. (2005). Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* 48, 2534–2547. doi: 10.1021/jm049092j

Trott, O., and Olson, A. J. (2009). AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455–461. doi: 10.1002/jcc.21334

Truchon, J. F., and Bayly, C. I. (2007). Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* 47, 488–508. doi: 10.1021/ci600426e

Tuccinardi, T., Poli, G., Romboli, V., Giordano, A., and Martinelli, A. (2014). Extensive consensus docking evaluation for ligand pose prediction and virtual screening studies. *J. Chem. Inf. Model.* 54, 2980–2986. doi: 10.1021/ci500424n

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins Struct. Funct. Genet.* 52, 609–623. doi: 10.1002/prot.10465

Verschueren, E., Vanhee, P., Rousseau, F., Schymkowitz, J., and Serrano, L. (2013). Protein-peptide complex prediction through fragment interaction patterns. *Structure* 21, 789–797. doi: 10.1016/j.str.2013.02.023

Vilar, S., Cozza, G., and Moro, S. (2008). Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Curr. Top. Med. Chem.* 8, 1555–1572. doi: 10.2174/156802608786786624

von Korff, M., Freyss, J., and Sander, T. (2009). Comparison of ligand- and structure-based virtual screening on the DUD data set. *J. Chem. Inf. Model.* 49, 209–231. doi: 10.1021/ci800303k

von Wartburg, A., and Traber, R. (1988). 1 cyclosporins, fungal metabolites with immunosuppressive activities. *Prog. Med. Chem.* 25, 1–33. doi: 10.1016/S0079-6468(08)70276-5

Wang, T., Wu, M., Chen, Z., Chen, H., Lin, J., and Yang, L. (2015). Fragment-based drug discovery and molecular docking in drug design. *Curr. Pharm. Biotechnol.* 16, 11–25. doi: 10.2174/1389201015666141122204532

Ward, W. H. J., Cook, P. N., Slater, A. M., Davies, D. H., Holdgate, G. A., and Green, L. R. (1994). Epidermal growth factor receptor tyrosine kinase. Investigation of catalytic mechanism, structure-based searching and discovery of a potent inhibitor. *Biochem. Pharmacol.* 48, 659–666. doi: 10.1016/0006-2952(94)90042-6

Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* 11, 1046–1053. doi: 10.1016/j.drudis.2006.10.005

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, 1074–1082. doi: 10.1093/nar/gkx1037

Wójcikowski, M., Ballester, P. J., and Siedlecki, P. (2017). Performance of machine-learning scoring functions in structure-based virtual screening. *Sci. Rep.* 7:46710. doi: 10.1038/srep46710

Wood, A., and Armour, D. (2005). The discovery of the CCR5 receptor antagonist, UK-427,857, a new agent for the treatment of HIV infection and AIDS. *Progress Med Chem*. 43, 239–271. doi: 10.1016/S0079-6468(05)43007-6

Xia, J., Feng, B., Shao, Q., Yuan, Y., Wang, X. S., Chen, N., et al. (2017). Virtual screening against phosphoglycerate kinase 1 in quest of novel apoptosis inhibitors. *Molecules* 22:1029. doi: 10.3390/molecules22061029

Yuriev, E., and Ramsland, P. A. (2013). Latest developments in molecular docking: 2010-2011 in review. *J. Mol. Recognit.* 26, 215–239. doi: 10.1002/jmr.2266

Zhao, W., Hevener, K. E., White, S. W., Lee, R. E., and Boyett, J. M. (2009). A statistical framework to evaluate virtual screening. *BMC Bioinform.* 10:225. doi: 10.1186/1471-2105-10-225

Zhao, Y., and Sanner, M. F. (2007). FLIPDock: docking flexible ligands into flexible receptors. *Proteins Struct. Funct. Bioinforma* 68, 726–737. doi: 10.1002/prot.21423

Zilian, D., and Sotriffer, C. A. (2013). SFCscoreRF: a random forest-based scoring function for improved affinity prediction of protein-ligand complexes. *J. Chem. Inf. Model.* 53, 1923–1933. doi: 10.1021/ci400120b

Zsoldos, Z., Reid, D., Simon, A., Sadjad, S. B., and Johnson, A. P. (2007). eHiTS: a new fast, exhaustive flexible ligand docking system. *J. Mol. Graph. Model.* 26, 198–212. doi: 10.1016/j.jmgm.2006.06.002

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past co-authorship with the author LA.

Copyright © 2020 Maia, Assis, de Oliveira, da Silva and Taranto. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 29 April 2020
doi: 10.3389/fchem.2020.00243

An Innovative Sequence-to-Structure-Based Approach to Drug Resistance Interpretation and Prediction: The Use of Molecular Interaction Fields to Detect HIV-1 Protease Binding-Site Dissimilarities

Nuno G. Alves[1†], *Ana I. Mata*[1†], *João P. Luís*[1†], *Rui M. M. Brito*[1,2] *and Carlos J. V. Simões*[1,2*]

[1] *Department of Chemistry, Coimbra Chemistry Centre, University of Coimbra, Coimbra, Portugal,* [2] *BSIM Therapeutics, Instituto Pedro Nunes, Coimbra, Portugal*

OPEN ACCESS

Edited by:
Simone Brogi,
University of Pisa, Italy

Reviewed by:
Sinosh Skariyachan,
St. Pius X College, Rajapuram, India
Zhijun Li,
University of the Sciences, United States

***Correspondence:**
Carlos J. V. Simões
carlos.simoes@bsimtx.com

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

Received: *06 December 2019*
Accepted: *13 March 2020*
Published: *29 April 2020*

Citation:
Alves NG, Mata AI, Luís JP, Brito RMM and Simões CJV (2020) An Innovative Sequence-to-Structure-Based Approach to Drug Resistance Interpretation and Prediction: The Use of Molecular Interaction Fields to Detect HIV-1 Protease Binding-Site Dissimilarities. Front. Chem. 8:243. doi: 10.3389/fchem.2020.00243

In silico methodologies have opened new avenues of research to understanding and predicting drug resistance, a pressing health issue that keeps rising at alarming pace. Sequence-based interpretation systems are routinely applied in clinical context in an attempt to predict mutation-based drug resistance and thus aid the choice of the most adequate antibiotic and antiviral therapy. An important limitation of approaches based on genotypic data exclusively is that mutations are not considered in the context of the three-dimensional (3D) structure of the target. Structure-based *in silico* methodologies are inherently more suitable to interpreting and predicting the impact of mutations on target-drug interactions, at the cost of higher computational and time demands when compared with sequence-based approaches. Herein, we present a fast, computationally inexpensive, *sequence-to-structure*-based approach to drug resistance prediction, which makes use of 3D protein structures encoded by input target sequences to draw binding-site comparisons with susceptible templates. Rather than performing atom-by-atom comparisons between input target and template structures, our workflow generates and compares Molecular Interaction Fields (MIFs) that map the areas of energetically favorable interactions between several chemical probe types and the target binding site. Quantitative, pairwise dissimilarity measurements between the target and the template binding sites are thus produced. The method is particularly suited to understanding changes to the 3D structure and the physicochemical environment introduced by mutations into the target binding site. Furthermore, the workflow relies exclusively on freeware, making it accessible to anyone. Using four datasets of known HIV-1 protease sequences as a case-study, we show that our approach is capable of correctly classifying resistant and susceptible sequences given as input. Guided by ROC curve analyses, we fined-tuned a dissimilarity threshold of classification that results in remarkable discriminatory performance (accuracy ≈ ROC AUC ≈ 0.99), illustrating the high potential of *sequence-to-structure*-, MIF-based approaches in the context of drug

resistance prediction. We discuss the complementarity of the proposed methodology to existing prediction algorithms based on genotypic data. The present work represents a new step toward a more comprehensive and structurally-informed interpretation of the impact of genetic variability on the response to HIV-1 therapies.

Keywords: drug resistance prediction, Molecular Interaction Fields, *sequence-to-structure* algorithm, binding-site dissimilarities, HIV-1 protease

INTRODUCTION

Drug resistance is one of the greatest threats of the twenty first century. Fundamentally, the problem resides in the development and spread of resistance-conferring mechanisms among infectious pathogens such as viruses and other microbial targets (McKeegan et al., 2002). Importantly, the selection of random mutations stands out as one of the main mechanisms of acquiring resistance, particularly relevant in viruses which mutate at high frequencies. RNA viruses, for instance, have a mutation rate estimated at 10^{-4} per nucleotide per replication, while DNA viruses have a rate of 10^{-8} per nucleotide per replication (Vere Hodge and Field, 2011; Mason et al., 2018). The extreme variability and rapid mutational spectrum of viral genomes, ongoing viral replication, and prolonged drug exposure linked with the selection and widespread of new drug-resistant strains is still a matter of great concern and importance, particularly in immunocompromised populations (Strasfeld and Chou, 2010; Mason et al., 2018). While a limited number of antiviral drug classes are getting approved for human use, an increasing resistance to some of the most effective available antivirals for HIV/AIDS, herpes, influenza and hepatitis, is being observed. Furthermore, the unpredictability of viral evolution and drug resistance means that antiviral treatments remain costly to the health care systems and are still associated with a significant risk of mortality, particularly in low- and middle-income countries (Irwin et al., 2016). Hence, *a priori* understanding and prediction of resistance against drug targets is of paramount importance toward developing more effective and longer lasting treatment options and regimens.

Antiviral drug resistance has been extensively studied in the rapidly mutating human immunodeficiency virus (HIV). HIV-1, in particular, is one of the most studied virus and the increasingly affordable and accessible genotypic data from clinical HIV-1 strains, together with corresponding data on strain susceptibility or resistance toward several drugs, have sparked the development of several genotypic interpretation systems for prediction of phenotypic drug resistance and therapy response based on genotype (Bonet, 2015). Said systems include (a) rule-based algorithms, including the *Agence Nationale de Recherche sur le Sida* (ANRS) (Brun-Vézinet et al., 2003), the Stanford HIV Drug Resistance Database interface (HIVdb) (Tang et al., 2012), Rega (Van Laethem et al., 2002), and HIV-GRADE (Obermeier et al., 2012a), which heavily rely on the periodic update of mutation-resistance profile lists, and on the knowledge of expert panels; and (b) machine learning-based algorithms trained on large sets of genotype–phenotype pairs to predict the *in vitro* resistance to a specific drug, with renowned examples such as *geno2pheno* (Beerenwinkel et al., 2003) and SHIVA (Riemenschneider et al., 2016). These sequence-based methods are relatively fast and low cost, justifying their routine use to support medical decision in HIV pharmacotherapy (Vercauteren and Vandamme, 2006).

The most relevant computational predictors of antiviral drug resistance currently available share the shortcoming of being purely based on genotypic sequence data. By disregarding the three-dimensional structural context and enzymatic function of the mutated amino acid residues, these systems fail to capture the links between genetic viral mutations and the corresponding mutation-induced structural changes to the effector protein viral machinery (Cao et al., 2005; Weber and Harrison, 2016; Khalid and Sezerman, 2018). This means that such methods are limited in their predictive power and interpretability toward novel mutations and combinations of mutations that go beyond the information accessible for training, such as mutation patterns that are encountered in only a small number of patients.

In contrast, structure-based methods hold potential to help understanding and eventually predicting resistance mechanisms for previously unknown data, shedding light on the elusive link between novel mutations and drug resistance. This may be justified by the fact that such methods can take advantage of available structural information on protein-ligand complexes and structural modeling of point mutations in the protein structure (Hao et al., 2012). Reported examples of the use of structure-based methods include the application of molecular docking to predict resistance or susceptibility of HIV1-PR to different inhibitors (Jenwitheesuk and Samudrala, 2005; Toor et al., 2011), the use of molecular dynamics simulations to study the impact of mutations on enzyme dynamics, stability and binding affinity (Hou and Yu, 2007; Agniswamy et al., 2016; Sheik Amamuddy et al., 2018), and the use of computational mutation scanning protocols to extract insights on free energy and binding affinity changes resulting from active site and non-active site mutations (Hao et al., 2010). Even though these methods are constantly adding new pieces to the puzzle and opening opportunities in the understanding of drug resistance, they suffer from various drawbacks, such as being time-consuming and offering limited predictive accuracy. As a result of such limitations, the primary challenge facing structure-based drug resistance prediction is to achieve an acceptable balance between prediction accuracy and computational efficiency to become both reliable and fast tools to be used in clinic context (Hao et al., 2012). In fact, some of the most recent reports describe the use of machine learning strategies merging both sequence and structural data in attempt

to achieve such balance (Masso and Vaisman, 2013; Yu et al., 2014; Khalid and Sezerman, 2018).

In this contribution, we describe a fast, computationally inexpensive, *sequence-to-structure*-based approach to the prediction of drug resistance. The proposed workflow makes use of an archetypal GRID-based method (Goodford, 1985) involving the generation and comparison of Molecular Interaction Fields (MIFs). MIFs may be defined as the spatial variation of interaction energies between a molecular target structure and selected types of chemical probes laid out on a three-dimensional (3D) grid (Cruciani, 2005). The broad range of applications of MIFs extends from ligand-based methodologies, e.g., 3D Quantitative Structure-Activity Relationships (3D-QSAR) models, drug metabolism and pharmacokinetics (DMPK) predictions and pharmacophore elucidation, all the way to structure-based drug design, including binding site detection and molecular docking (Artese et al., 2013). Within the context of viral drug resistance, MIFs hold potential in capturing subtle, mutation-induced, chemical perturbations within the binding site of resistant or susceptible viral structures, thus representing a promising approach to anticipating the impact of mutations on the response to antiviral drugs with atomistic detail.

HIV-1 protease (HIV1-PR) is one of the most characterized viral enzymes, with extensive structural, inhibitor, and mutation data available (Weber and Agniswamy, 2009). As of late 2019, the RCSB Protein Data Bank (RCSB PDB, 2000) ranks HIV-1 as the virus holding the highest number of available structures (2,586), majorly obtained through X-ray crystallography. Of these, the PDB returns 662 entities with at least 90% identity to the HIV1-PR subtype B *consensus* sequence from a BLAST sequence search (Stanford University, 1998a). The search by *consensus* sequences of other HIV-1 subtype B enzymes (Stanford University, 1998a) returns 586 structures for reverse transcriptase and 190 for integrase. With such amount of structural information available, we have built the framework of the present work using HIV1-PR as our first case-study. Commercially available HIV-1 protease inhibitors (PIs) are competitive peptidomimetics with a core structural scaffold that mimics the tetrahedral transition state of HIV1-PR substrate. Although these drugs are chemically distinct, their active conformations are superimposable, and generally establish the same pharmacophoric interactions with their target (Wlodawer and Erickson, 1993; King et al., 2004; Qiu and Liu, 2011; Nayak et al., 2019). Many mutations in HIV1-PR translate into changes in the structure and binding site physicochemical environment, thus affecting the affinity of PIs and representing a hurdle to achieving long-term viral suppression (Irwin et al., 2016; Pawar et al., 2019; Wensing et al., 2019). A quantitative analysis of HIV1-PR drug-resistant mutation frequency, with particular focus on the binding site, was performed using public sequence datasets to support the potential of a MIF-based approach to capturing mutation-induced active site dissimilarities. From this perspective, the workflow proposed here encompasses the use of a conservative structural modeling step for the generation of a HIV1-PR structure from its respective amino acid sequence, and a MIF-based structural alignment and chemical dissimilarity detection step comparing the input *sequence-structure* pair with a carefully selected naïve, susceptible template *sequence-structure* pair. We demonstrate that the quantification of such dissimilarity, depicting the extent of structural, physicochemical and pharmacophoric alterations introduced by mutations, allows for an accurate prediction of HIV1-PR's resistance to PIs.

Compared with previous approaches reported in the literature, and to the best of our knowledge, this work stands out as a first implementation of a fast, *sequence-to-structure*-based algorithm capable of discriminating susceptible and resistant HIV1-PR sequences. Considering that the problem of mutation-induced resistance cuts across virtually all infectious diseases, we believe the approach reported herein may be extended to a wide range of microbial targets besides HIV-1, thus helping rationalize and personalize the therapeutic decision-making process.

MATERIALS AND METHODS

The availability of a public and curated database such as HIVDB (Stanford University, 1998c; Rhee et al., 2003) allows access to HIV1-PR sequences with known levels of resistance, and thus to establish datasets for the development of new methodologies to predict HIV1-PR resistance to protease inhibitors (PIs). This section describes the materials and methods employed in (1) the preparation of sequence datasets with various levels of resistance to PIs; (2) frequency analysis of *major* and *minor* mutations in the sequence datasets in (1); (3) the structural modeling of the reference structure used as template for subsequent modeling of HIV1-PR structures corresponding to each sequence in the datasets; (4) the core components of the proposed algorithm, including the calculation and comparison of pairwise Molecular Interaction Field points between the resulting structural models and the selected naïve template structure; and (5) the performance metrics used to test and evaluate the predictive power of the developed structure-based drug-resistance classification algorithm. A general workflow illustrating (4) and (5) is sketched (draw.io, 2005) in **Figure 1** and the complete script *HIV1predict.sh* for running the sequences is available at GitHub (Alves et al., 2019b). Calculations were run on a 64-bit CentOS 6 Linux server with an Intel Xeon CPU (E5620) at 2.40 GHz (further information as **Supplementary Table S1**).

Datasets of Resistant and Susceptible Sequences

A set of genotype-phenotype correlated HIV1-PR sequences was retrieved from HIVDB, version 8.7 (Stanford University, 1998b,c), and filtered by drug class for PIs. The considered PIs include darunavir, fosamprenavir, atazanavir, indinavir, lopinavir, nelfinavir, saquinavir, and tipranavir. Analyzing the subtype B HIV1-PR sequence of each isolate, i.e., a viral sample obtained from an infected individual, and considering positions with a mixture of amino acids, all possible mutation patterns were written to the FASTA format using a script written in-house (Alves et al., 2018f).

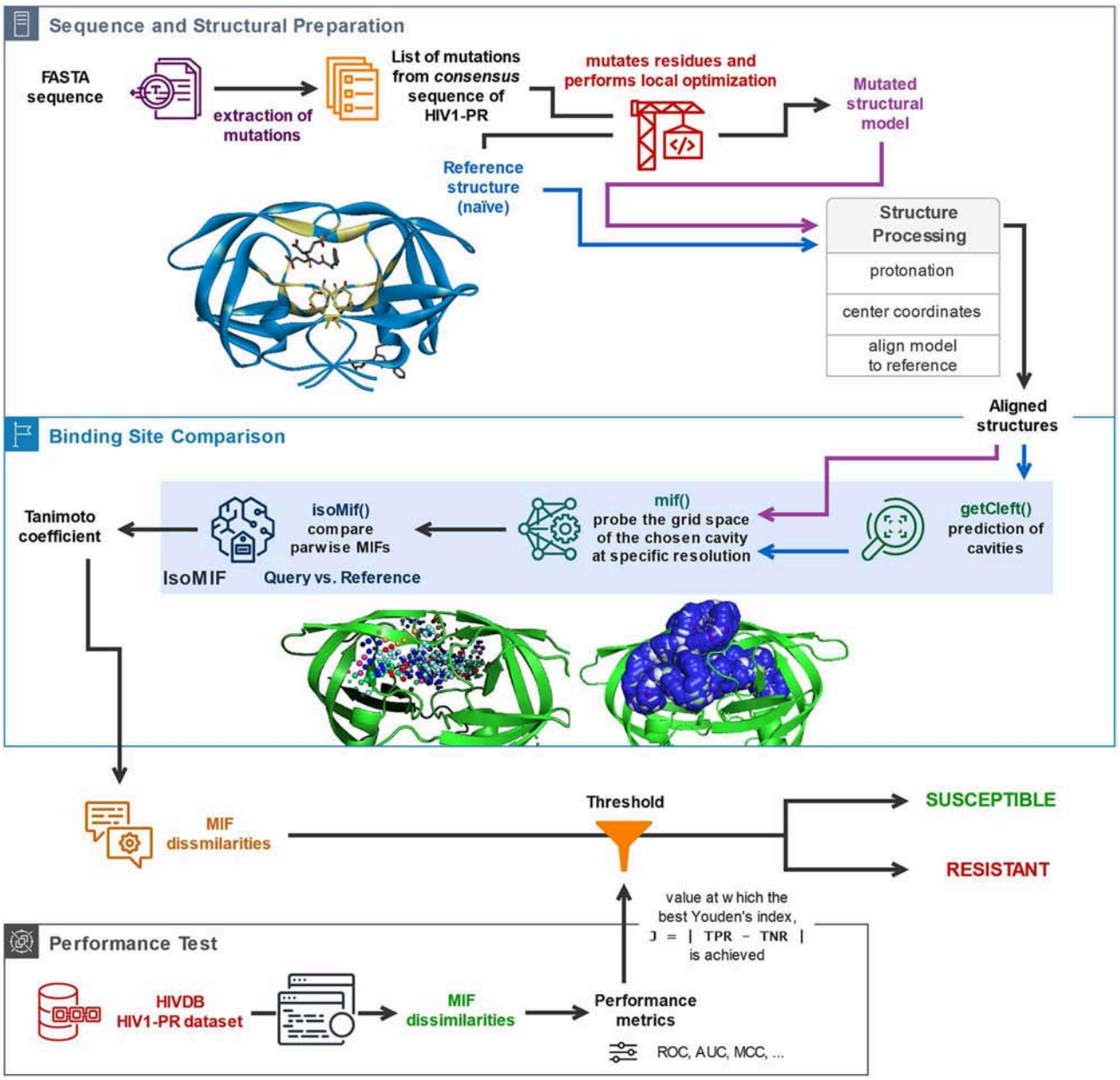


FIGURE 1 | General description of the workflow underlying the proposed *sequence-to-structure*-based drug resistance classifier, holding platform applications to multiple microbial targets. The depicted algorithm starts off with the sequential reading of an amino acid sequence in FASTA format, followed by identification of the present mutations by comparison with the consensus sequence, insertion of the identified mutations in the naïve template, and processing of the structural models for alignment and comparison. Upon structural alignment of the target or database (predicted) structure with the naïve template structure, six types of MIF probe points are computed for the former structure and superimposed with pre-calculated MIF probe points of the latter structure. Calculation of MIF dissimilarities by means of a Tanimoto coefficient proceeds. The bottom panel represents the process of performance evaluation of the proposed classifier based on its application to a large dataset of *sequence-structure* pairs generated for HIV1-PR sequences retrieved from HIVDB. Included are performance metrics such as accuracy, Matthews Correlation Coefficient (MCC), and the area under the receiver-operating-characteristic curve (ROC AUC).

The genotype-phenotype correlation results from the *in vitro* PhenoSense assay (Zhang et al., 2005), which measures the levels of resistance to a PI compared to the wild-type sequence. Following the categorization of susceptibility to PIs described by Rhee et al. (2006), the collected sequences were classified as follows:

- *Susceptible*. Sequences holding <3.0-fold resistance to all PIs in the dataset were considered susceptible (N = 7,768) [Susceptible].
- *Resistant*. Sequences holding more than 20.0-, or 15.0-, or 10.0-fold resistance to all PIs, resulting in three resistant subgroups of increasing degree of resistance: respectively, [Res_{20}] (N =

60) [Res20], [Res_{15}] ($N = 83$) [Res15], which encompasses [Res_{20}] plus 23 sequences holding between 15- and 20-fold resistance, and [Res_{15}] ($N = 873$) [Res10], which encompasses [Res_{20}] and [Res_{15}] plus 790 additional sequences holding between 15- and 10-fold resistance.

Counting of Mutations in HIV1-PR

The quantification of *major* and *minor* mutations (Weber and Agniswamy, 2009) in all datasets was carried out using scripts written in-house (Alves et al., 2018a,b, 2019a) that sequentially read the listing of mutations for each sequence, extract either the *major* or *minor* mutations, and count them for each sequence. Said script was applied to quantify *major* and *minor* mutations in the HIV1-PR binding site.

Preparation of HIV1-PR Structures

Using PDB's BLAST utility (Altschul et al., 1990) to guide the choice of a template for homology modeling, a sequence search, with a 10.0 E-value cut-off and at least 50% identity to the HIV1-PR subtype B *consensus* (Stanford University, 1998a), resulted in 784 entities available. With a more refined query of at least 95% identity to the HIV1-PR subtype B *consensus*, there were still 376 structures available to work with.

Out of these 376 structures, PDB entry 1NH0 for HIV1-PR was chosen as template structure for homology modeling by using PDB's BLAST utility (Altschul et al., 1990). It returned an E-value of 7.20281E-51, but since the intended work was heavily based on structure, our choice was also based on having the best resolution possible. The structure of 1NH0 holds 99% sequence identity (98/99) with the *consensus* B amino acid sequence of protease, HXB2 (henceforth referred to as *consensus sequence*), with one single mutation at position 37 (S37N), has 100% coverage of the sequence, and has been determined at 1.03 Å X-ray resolution. Importantly, this HIV1-PR sequence is known to be susceptible to all PIs.

In this work, Modeler version 9.19 (Šali and Blundell, 1993; Šali, 2019a) was used for predictive modeling of all HIV1-PR structures from their respective sequences. The listing of mutations present in each sequence was automated by scripting (Alves et al., 2018c) and followed by sequentially running the *mutate_model.py* script provided with Modeler (Šali, 2019b) to obtain the correct pattern of mutations and outputting the respective structural model. The procedure implemented in *mutate_model.py* performs local optimization of the mutated residues region and ensures that the obtained structural models are comparable to the template structure. The PDB structure itself (1NH0) was subjected to *mutate_model.py* in order to reverse the mutation present in the template with 99% identity (Asn37, on the outside of the protease) and keep on the *consensus sequence*, remove *HETATM* entries and *alt-locs*—thus yielding the reference template structure. This reference structure was used as template for the generation of the respective structural model of each input FASTA sequence present in the datasets.

All generated structural models were protonated using *Reduce*, version 3.23 (Word et al., 1999). The reference structure was centered to the origin of the axes of the cartesian coordinate system using *VMD*, version 1.9.3 (Humphrey et al., 1996). Structural alignment of all query models onto the centered reference structure was performed with *LovoAlign*, version 16.342 (Martínez et al., 2007).

Workflow for Detection and Scoring of Molecular Interaction Field Dissimilarities

The MIF module of the software package IsoMIF, version dated March 2015 (Chartier and Najmanovich, 2015), was used to generate Molecular Interaction Fields (MIFs) within the HIV1-PR binding sites. MIF-based alignment and calculation of pairwise MIF dissimilarities between reference and dataset binding sites proceeded using the IsoMIF module of the same package. The IsoMIF setup comprises three sequential modules: GetCleft, MIF, and IsoMIF.

Cavity Detection (GetCleft Module)

GetCleft (Gaudreault et al., 2015) was employed to predict cavities in the structure of the reference HIV1-PR (Alves et al., 2018e). This geometry-based method detects cavities by insertion of spheres of radius *r* between the non-hydrogen protein atoms, reducing such radius if they intersect with any neighboring atoms (clefts defined by the union of overlapping spheres). First, the top five largest cavities were searched at the same time, with a minimum and maximum sphere radius of 1.5 and 4.0 Å, respectively. The largest predicted cavity was visually confirmed to be completely enclosed within the HIV1-PR binding site, using *VMD*, version 1.9.3 (Humphrey et al., 1996). Next, such cavity volume represented by spheres was used to define the location of MIF interaction vectors to be calculated for the reference and all 3D HIV1-PR structural models.

Generation of Molecular Interaction Field (MIF) Probe Points (MIF Module)

The MIF module of IsoMIF was used to compute molecular interaction fields (MIFs) for six different chemical probe types (**Figure 2**): hydrophobic, aromatic, H-bond donor, H-bond acceptor, positive charge and negative charge. The pharmacophoric features shared by PIs (Wlodawer and Erickson, 1993; Nayak et al., 2019) highlight the importance of a conserved physicochemical environment in the binding site. Alterations of this environment are detected with the MIF probes (circled in **Figures 2A,B**) which allow for a quantification of changes caused by the presence of mutations. In this work, a grid resolution of 1.5 Å was defined to calculate the MIFs on the cleft covering the volume of the binding site. Such resolution was selected upon testing to achieve an adequate balance between speed and accuracy of IsoMIF pairwise field dissimilarity calculations.

Alignment of MIF Probe Points and Calculation of Dissimilarities (IsoMIF Module)

Field similarities were computed using the IsoMIF module, which employs a clique-based graph matching approach based on the Bron-Kerbosch algorithm (Bron and Kerbosch, 1973) to perform functional alignments between the probe points under comparison. A grid spacing of 1.5 Å, a geometric distance threshold of 1.0 Å and a maximum of 100 cliques were used as parameters for the calculation of similarities between the

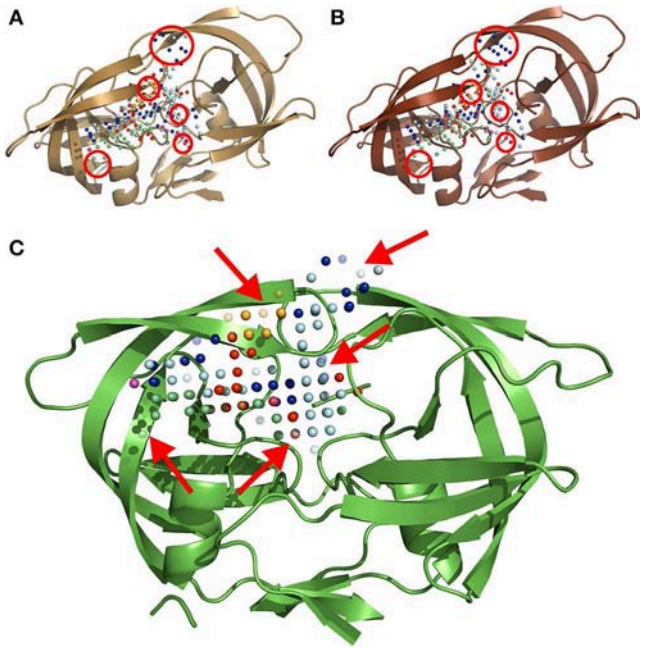


FIGURE 2 | A three-dimensional ribbon depiction of the quaternary structure HIV1-PR dimeric unit, including the six different types of MIF probe points calculated on the enzyme's binding site. (A) Structure and probe points of the reference naïve template. (B) Illustration of the structure and probe points of a mutated model resulting from a sequence holding over 20-fold resistance for PIs. (C) An exemplary output of IsoMIF, highlighting a MIF-based alignment and comparison of both structures: (A,B). The red circles in (A,B) denote areas where the difference in probe points is most visible. The arrows in (C) point to semi-transparent probes, representing the probes which are not shared between the two structures. Legend for the six (6) probe types legend: hydrophobic in light blue, aromatic in orange, H-bond donor in dark blue and acceptor in red, positive in green and negative charge in purple.

binding site of reference and structural models of HIV1-PR. Such similarities were then quantified by the Tanimoto coefficient (T_c), calculated as in Equation 1:

$$T_c = \frac{N_C}{N_R + N_Q - N_C} \tag{1}$$

where N_c is the number of common probe points to the two MIF maps under comparison; N_r and N_q represent the number of probe points present in the reference and query structure, respectively (**Figure 2C**) (Chartier and Najmanovich, 2015). The measurement of dissimilarity (Equation 2) between binding sites is justified by the fact that the focus of this work is set on the discrimination of resistant structures, when compared with a susceptible reference. Therefore, the chosen metric was dissimilarity rather than similarity:

$$dissimilarity\ coefficient = 1.0 - T_c \tag{2}$$

Analysis of Mutation Patterns Across Thousands of HIV1-PR Sequences

Analyses of the number and position of mutations were performed on HIV1-PR sequences in order to obtain information supporting and justifying the development of a *sequence-to-structure*-, MIF-based approach to antiviral resistance classification and prediction.

R version 3.4.3 (R Core Team, 2018) was used to conduct the analysis and generating the associated graphical representations. The R packages used in this work were ggplot2 (Wickham, 2009), *gplots* (Warnes et al., 2019), and ROCit (Khan and Brandenburger, 2019).

"Outlier" Detection on Binding-Site MIF Dissimilarities

Tukey's method (Tukey, 1949; Hoaglin, 2003), also referred to as Tukey's fences method, was used to detect outliers in the binding-site MIF dissimilarities results. Tukey's method is a statistical approach used to determine whether a value should be considered an outlier or not: the method relies on the interquartile range (IQR) measurement, which is calculated by the difference between the first quartile (Q1) and the third quartile (Q3) (see Equation 3). Q1 stands for the value in the dataset that holds 25% of the values below it and Q3 is the value in the dataset that holds 25% of the values above it.

$$IQR = Q3 - Q1 \tag{3}$$

According to Tukey's method, a value is considered an outlier if it is observed in the range described in Equation 4:

$$\begin{aligned} outlier &< Q1 - k \times IQR \vee outlier > Q3 + k \times IQR \\ outlier &< LowerBound \vee outlier > UpperBound \end{aligned} \tag{4}$$

where $k = 1.5$ indicates an outlier and $k = 3$ indicates an extreme outlier. For the purpose of the present work, only extreme outliers were discarded.

Evaluation of the Algorithm's Predictive Performance

The performance of our method at discriminating resistant from susceptible models was assessed by calculation of several metrics typically employed in the fields of predictive modeling and machine learning, particularly in cases where binary classification occurs. These included the Receiver Operating Characteristic (ROC) and the respective Area Under the Curve (ROC AUC). The ROC curve is a graphical representation of the True Positive Rate (TPR) as a function of the True Negative Rate (TNR), i.e., at various cut-off settings. The TPR is also known as Sensitivity (Equation 5), which measures the proportion of positive cases. On the other hand, the TNR is also calculated as 1—Specificity (Equation 6) and measures the proportion of true negative cases.

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

where *TP* represents the number of correctly identified resistant structures (true positives), *TN*, the number of correctly identified susceptible structures (true negatives), *FP*, the number of susceptible incorrectly predicted as resistant (false positives), and

FN the number of resistant incorrectly predicted as susceptible (false negatives).

Additional performance metrics included Accuracy (Equation 7) and Matthews Correlation Coefficient (MCC; see Equation 8) (Matthews, 1975; Florkowski, 2008; Powers, 2011).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\,(TP+FN)\,(TN+FP)\,(TN+FN)}} \quad (8)$$

The dissimilarity threshold used for classification in resistant or susceptible *sequence-structure* pairs was derived from ROC curves, corresponding to the highest Youden's index (Youden, 1950), J, calculated as in Equation 9:

$$J = Sensitivity + Specificity - 1 \quad (9)$$

This index defines the maximum potential effectiveness of a classifier. It can be determined for all points of an ROC curve, although its maximum value represents the classifier optimal differentiating ability cut-point when equal weight is given to Sensitivity and Specificity (Ruopp et al., 2008).

RESULTS AND DISCUSSION

In this work, we describe a *sequence-to-structure*-, MIF-based method to assess binding-site dissimilarities across *sequence-structure* pairs, with the aim of predicting antiviral resistance—and using HIV1-PR as a case-study. It is generally accepted that the majority of resistance-conferring mutations occur in the binding site regions of viral enzymes (Weber and Agniswamy, 2009; Weber and Harrison, 2016). In order to further support the *rationale* and underlying assumptions of the proposed approach, we performed analysis of *major* and *minor* mutations of HIV1-PR binding site residues focusing on sequences known to be fully resistant and fully susceptible. For the sake of comparison, the quantification of mutations was also extended to *major* and *minor* mutations occurring in the remainder residues, i.e., residues not comprising the binding site region of HIV1-PR.

Counting of PI-Resistant Mutations in HIV1-PR Sequences

Resistance to PIs develops upon accumulation of mutations that increasingly impact the structure of HIV1-PR, resulting in highly-resistant variants of HIV-1. As mentioned by Weber and Agniswamy (2009), PI resistance is linked to the occurrence of primary (*major*) mutations, commonly associated with the active site where HIV PIs typically bind, resulting from structural changes that disrupt the van der Waals contacts and/or hydrogen bonding patterns in the inhibitor-protein interaction and promote direct steric hindrance, by altering the pocket volume or its physicochemical environment. Secondary (*minor*) mutations occur in addition to *major* mutations, acting like accessory mutations that compensate the flaws produced by *major* mutations and enhancing the resistance level (synergistic effect). Being less obvious, they seem to affect HIV1-PR catalysis, dimer stability, inhibitor binding kinetics, and/or active site re-shaping through long-range structural perturbations (Weber and Agniswamy, 2009; Weber and Harrison, 2016).

Our workflow follows a *sequence-to-structure* approach in attempt to capture changes to the structural and physicochemical determinants of HIV1-PR's binding site upon mutation, based on the assumption that these changes represent the main driver of antiviral resistance. To support this assumption, quantification of mutations known to contribute to PI resistance was carried out across the retrieved datasets. The version 8.7 HIVDB (Stanford University, 1998b,c,d) listed the following PI-resistant mutations for HIV1-PR:

- *Major* mutations: D30N, V32I, L33F, M46IL, I47VA, G48VM, I50VL, I54VTALM, L76V, V82AFTSL, I84V, N88SD, and L90M;
- *Minor* mutations: L10FIVRY, V11IL, K20RIMTV, L23I, L24IFM, M36I, K43T, M46V, G48ASTQL, F53LY, I54ST, Q58E, A71VTIL, G73STCADV, T74PS, V82MC, N83DS, I84AC, I85V, N88TG, and L89VT.

Even though not all sequences exhibit the same degree of resistance to each PI, we selected these two groups of *major* and *minor* PI-resistant mutations and quantitatively characterized their presence in our subsets. Since all HIV1-PR sequences in our dataset were retrieved from the same unique source, HIVDB (Stanford University, 1998c; Rhee et al., 2003), the percentage of sequences holding PI-resistant mutations distributed across the entire HIV1-PR sequence, as well as the percentage of PI-resistant mutations manifesting in residues comprising the binding site of HIV1-PR, were determined and compared among all four subsets: [Susceptible], [Res_{10}], [Res_{20}], and [Resistant*]–as represented in **Figure 3**.

Figures 3A,B shows that, as expected, all HIV1-PR sequences belonging to the *Susceptible* subset hold much less PIs-resistant mutations than those belonging to the *Resistant* subsets. The majority (98.24%) of susceptible HIV1-PR sequences does not hold any *major* mutations, while 1.74% contain one *major* mutation, and only one sequence (0.01%) comprises three *major* mutations. The presence of *major* mutations across drug-resistant sequences is higher, ranging from three to seven *major* mutations, implying that among these subsets the *major* mutations appear in the shape of mutation patterns rather than individual mutations. The presence of *minor* mutations (**Figure 3B**) follows a similar trend to that witnessed for *major* mutations, with susceptible sequences denoting a lower number when compared to their resistant counterparts. Approximately 98.25% of the susceptible sequences present two or less *minor* mutations, with about half of susceptible HIV1-PR sequences (53.3 %) displaying no *minor* mutations.

When comparing susceptible vs. drug-resistant sequences, it can be observed that resistance against PIs is linked to the presence of *major* mutations, as implied above (Weber and Harrison, 2016). However, within the subsets of drug-resistant sequences, a direct relation between the number of *major* mutations and the increase of resistance is not observed. Drug-resistant sequences show a higher frequency of *minor*

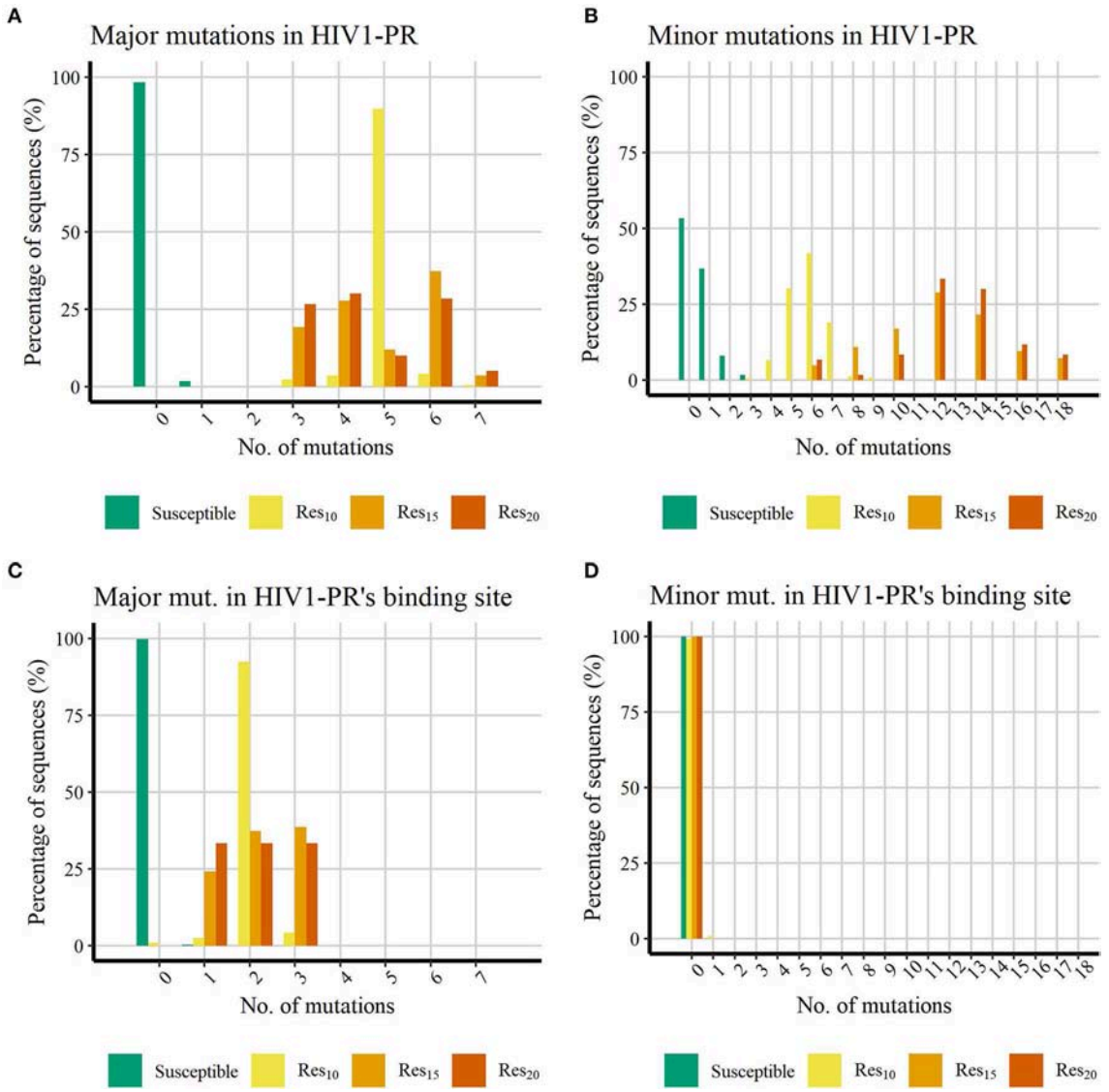


FIGURE 3 | Histograms representing the percentage of PI-resistant mutations in the datasets showing increasing number of PI-resistant mutations in the datasets retrieved from HIVDB. (A) Percentage of major mutations in the whole HIV1-PR. (B) Percentage of minor mutations in the whole HIV1-PR. (C) Percentage of major mutations in the HIV1-PR binding site. (D) Percentage of minor mutations in the HIV1-PR binding site. Represented datasets: [Susceptible] (green); [Res_{10}] (yellow); [Res_{15}] (orange); [Res_{20}] (dark orange).

mutations, ranging from three to 18, with a visual apparent difference between sequences with lower resistance ([Res_{10}]) and the more resistant sequences ([Res_{15}] and [Res_{20}]). In [Res_{10}], 98.2% of the sequences have up to seven *minor* mutations, while 78.3% in [Res_{15}] and 93.3% [Res_{20}] have more than eight *minor* mutations. This trend in the profile of mutation distribution among the resistant sequences is in line with *minor* mutations acting as accessory mutations, appearing as patterns and not as individual mutations, and showing a similar trait as the one observed for the distribution of *major* mutations.

Analysis of *major* mutations located in HIV1-PR's binding site residues (**Figure 3C**), corresponding to sequence positions 30, 32, 47, 48, 50, 82, and 84, shows that 99.78% of the susceptible sequences do not display *major* mutations, while the remainder show only one *major* mutation. In contrast, less than 1% of resistant sequences lack *major* mutations in the drug binding site. Interestingly, the eight sequences representing this small fraction (0.91%) belong to the lower (10-fold) resistance subset ([Res_{10}]). All remaining drug-resistant sequences hold from one to three *major* mutations in the enzyme's binding site.

Counting of mutations in binding site residues of HIV1-PR exposes a systematic presence of *major* mutations in resistant HIV1-PR sequences, while also highlighting the absence of such mutations on 99.78% of their susceptible counterparts. This contrasting trait observed between the binding site region of susceptible and resistant HIV1-PR supports the development

of a structure-based drug-resistance classifier focusing on the detection and quantification of binding site dissimilarities.

Regarding the distribution of *minor* mutations across binding site residues, as represented in **Figure 3D**, mutations localized in sequence positions 23, 48, 82, and 84 were quantified among both HIV1-PR susceptible and drug-resistant sequences, revealing that the great majority does not present *minor* mutations in their respective binding sites. Only a small percentage of susceptible (0.01%) and resistant sequences (0.91%) show *minor* mutations in this region. It should be noted that the small subset of resistant sequences holding a *minor* mutation in their binding site region correspond to sequences that do not display *major* mutations in the active site.

These results show that the binding site *minor* mutations are uncommon on the datasets of HIV1-PR sequences—be they resistant or susceptible. Although such mutations appear to be important to increase the enzyme resistance's by stabilizing the mutated protein structure, they seem to produce limited direct effect on the enzyme's binding site, where they are mostly absent. Thus, these results seem to be in agreement with our motivation to explore a quantitative detection of binding-site dissimilarities to predict HIV1-PR resistance to PIs, as the *major* mutations play the main role on altering the binding site conformation, volume and/or physicochemical environment.

The quantification of mutations in the datasets retrieved from HIVDB yielded distinct results between the susceptible and drug-resistant sequences. Most of the resistant sequences show a higher frequency of *major* mutations when compared to the susceptible set. All resistant sequences present at least one mutation in the binding site region, contrasting with 98% of susceptible sequences that do not present any *major* mutations in that site. It is worth noticing that half of the *major* mutations are found in the binding site of resistant sequences. However, when considering the total number of mutations, the increase in the number of mutations per sequence seems to hold a reflection on the increase in the resistance of the observed sequence. Furthermore, binding site *major* mutations are more likely to cause changes on the HIV1-PR binding cleft physicochemical environment when compared with susceptible enzymes which do not have such type of mutations.

A Fast, *Sequence-to-Structure*-, MIF-Based Antiviral Drug Resistance Classifier

The quantification of resistance-conferring mutations in HIV1-PR sequences, using the datasets retrieved from HIVDB, prompted us to further develop a discriminative resistance-classifier approach focused on analysis and comparison of binding-site MIFs. In practice, the proposed workflow involves performing structural modeling of input HIV1-PR sequences using the same template (i.e., 1NH0) and a script (Alves et al., 2018d) that calls *mutate_model.py* (Šali, 2019b) to conduct local energy minimization around the mutated residues of the HIV1-PR structure. Once the generation of structure models is concluded, the modules belonging to the IsoMIF package are deployed for cavity detection (GetCleft module), calculation of MIFs within the selected cavity volume (MIF module), field alignment and quantification of dissimilarities between MIF points computed for the dataset HIV1-PR structural models and those computed for a high quality [Susceptible] reference HIV1-PR structure (1NH0) and, finally, scoring by means of a Tanimoto coefficient (IsoMIF module). The average running time of the workflow is ≈ 77 s per sequence (**Supplementary Figure S1** and **Supplementary Datasheet S1**), considering that this value varies with the amount of mutations present in the HIV1-PR.

TABLE 1 | Tukey's method results to determine outliers.

	Susceptible	*Res_{10}*	*Res_{15}*	*Res_{20}*
Q1	0.0057	0.1075	0.1173	0.0649
Q3	0.0225	0.2041	0.2041	0.2171
IQR	0.0168	0.0966	0.0868	0.1522
Lower Bound	−0.0447*	−0.1823*	−0.1431*	−0.3917*
Upper Bound	0.0729	0.4939	0.4645	0.6737

*Quartile 1 (Q1), Quartile 3 (Q3), Inter Quartile Range (IQR), Upper Bound and Lower Bound values for susceptible sequences dissimilarity coefficient distribution. Upper and Lower Bound were calculated as described in Equation 4, with k = 3. Values above the upper bound and below the lower bound were considered outliers. *Negative values are not realistic lower bounds; the minimum value must be 0.*

Analysis of MIF Dissimilarities in HIV1-PR Binding Site

Figure 4 discloses the frequency of HIV1-PR *sequence-structure* pairs scattered across a spectrum of Tanimoto coefficient (Tc) values (varying from 0.00 to 1.00), in turn reflecting binding-site MIF dissimilarities in the subset of susceptible sequences (containing 7,768 *sequence-structure* pairs) against the selected naïve, template structure. Analyzing this profile of binding site dissimilarities, we observe that there are substantially more susceptible sequences concentrated on lower end of the dissimilarity spectrum. However, a small number of sequences ($N = 81$) present higher values, more visibly around the Tc value of 0.35. Since susceptible HIV1-PR *sequence-structure* pairs display a lower frequency of mutations in the binding site residues, we assume that Tc values deviating from the normal trend may highlight inconsistent data, errors and/or any form of outliers worthy of further investigation.

In order to verify if the higher Tc values could reflect true outliers, Tukey's outlier detection method was used (Tukey, 1949; Hoaglin, 2003). **Table 1** shows the result of applying the statistical Tukey method to the MIF dissimilarity Tc values obtained for the dataset of susceptible *sequence-structure* pairs, and to the [Res_{10}], [Res_{15}], and [Res_{20}] subsets. For each of the four groups, **Figure 5** shows boxplots summarizing the distribution of the MIF dissimilarity Tc values. On the susceptible subset, the higher Tc values were identified as significantly different from the central tendency (values were below the determined lower bound; see Equation 4 in Methods). Looking at the dataset of resistant *sequence-structure* pairs, *extreme* outliers (as described in the Methods section) were only found in the [Res_{10}] subset. These outliers were found to be associated with a software limitation wherein the same reference grid

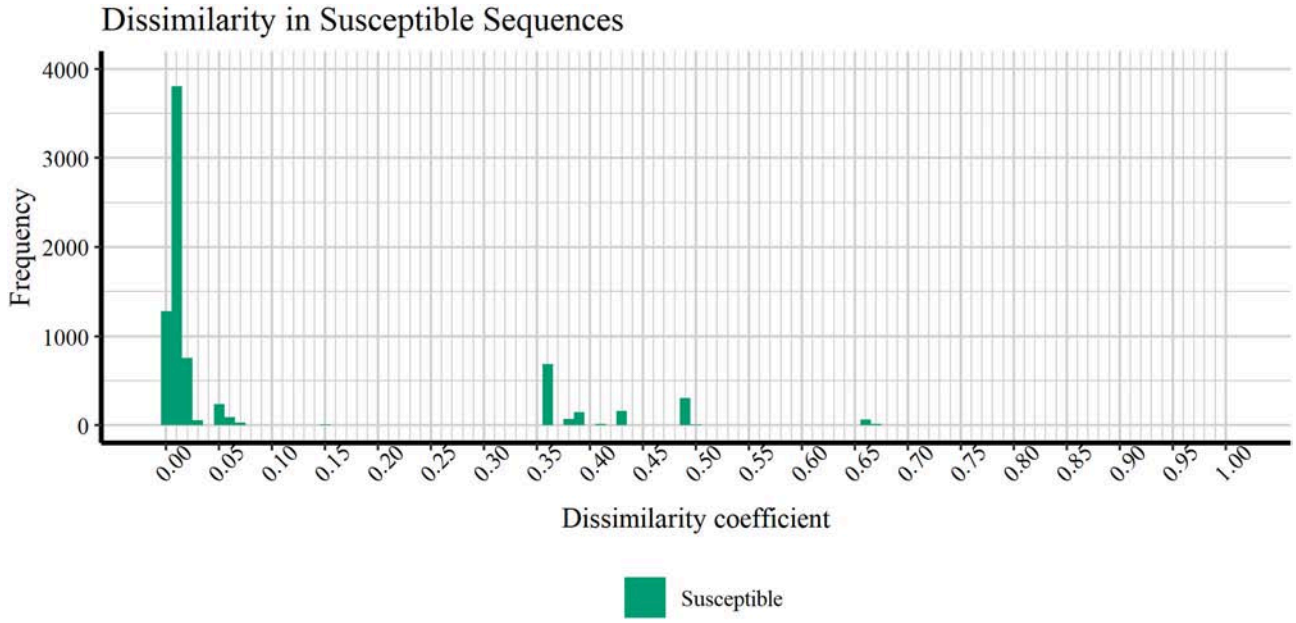


FIGURE 4 | Frequency of susceptible sequences distributed across a range of HIV1-PR binding site 3D dissimilarity coefficients. The maximum value reported is ~0.667.

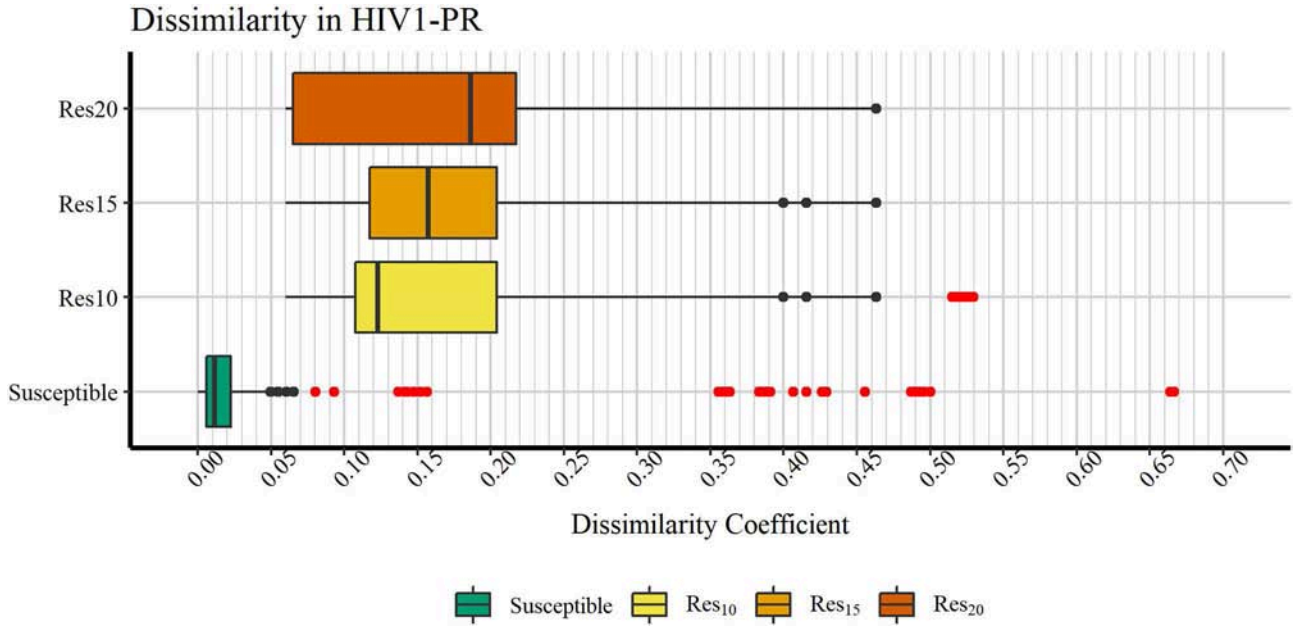


FIGURE 5 | Boxplot representation of dissimilarity coefficients obtained for the HIVDB datasets. The maximum value reported is ~0.667. Outliers are marked in red. Represented datasets: [Susceptible] (green); [Res_{10}] (yellow); [Res_{15}] (orange); [Res_{20}] (dark orange).

(generated by GetCleft), covering the entire binding site volume, was not homogeneous across all HIV1-PR structure models. In fact, a wider grid was calculated for some structures when compared to the reference HIV1-PR structure, which resulted on a different number of grid points, consequently leading to an increase of dissimilarities. Thus, these *sequence-structure* pairs were not considered relevant for performance evaluations, as they could introduce performance bias. The Tukey's boxplot analysis thus allowed the identification and removal of *extreme* outliers in the [Susceptible] and [Res_{10}] subsets, resulting in 6269 and 680 HIV1-PR structural models, respectively. The [Res_{15}] and [Res_{20}] subsets remained unchanged with 83 and 60

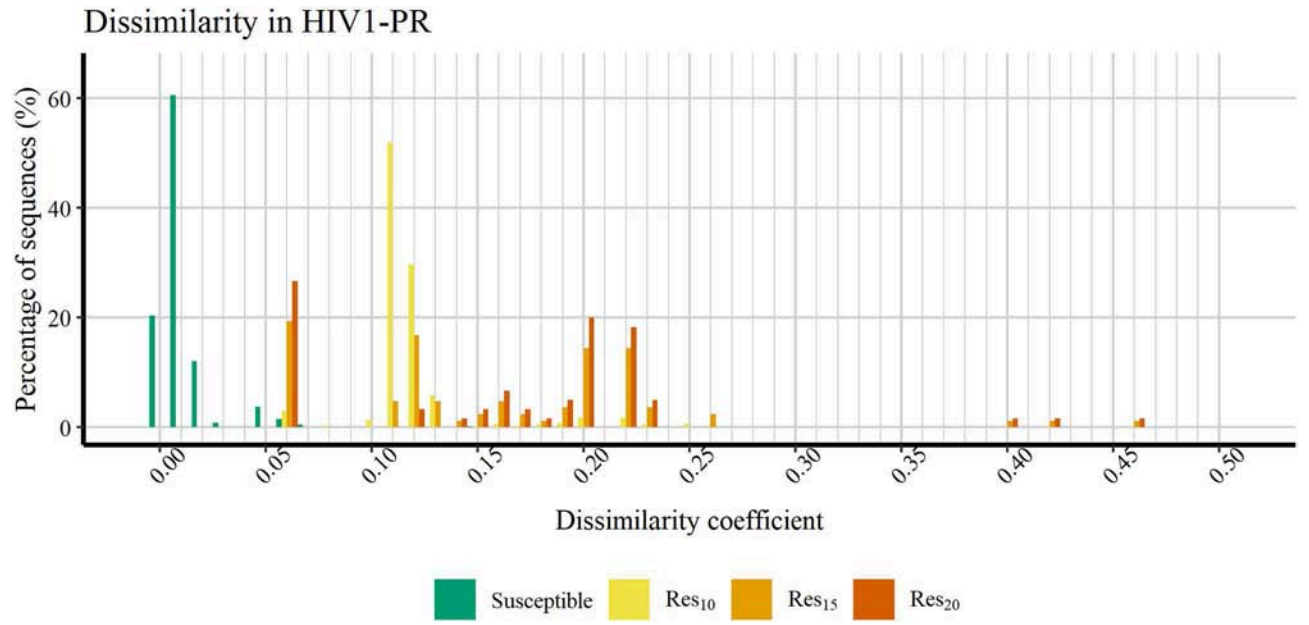


FIGURE 6 | Percentage of sequences distributed across a range of HIV1-PR binding site 3D dissimilarity coefficients. Maximum value reported is ~0.463. Represented datasets: [Susceptible] (green); [Res_{10}] (yellow); [Res_{15}] (orange); [Res_{20}] (dark orange).

HIV1-PR structural models, respectively. The resulting dataset has been used for further statistical analysis and as *test set* for performance calculations.

Figure 6 shows a profile of the HIV1-PR binding-site MIF dissimilarities across the susceptible dataset withdrawn of extreme outliers ([Susceptible*]) and the stratified resistant data set (encompassing [Res_{10}], [Res_{15}], and [Res_{20}]) also withdrawn of extreme outliers ([Susceptible*]). As seen, susceptible HIV1-PR structures tend to present very low to null binding-site MIF dissimilarities compared to the ([Susceptible]) structure modeled from the *consensus* sequence. In fact, 93.91% of the *sequence-structure* pairs in the susceptible group show dissimilarities lower than 0.02, indicating a considerable degree of conservation within the binding site. Overall, these results show a segregation between susceptible and resistant *sequence-structure* pairs, when analyzing their binding-site MIF dissimilarities against a susceptible reference sequence-structure pair, suggesting that our method is able to quantitatively capture differences among susceptible and resistant HIV1-PR structures.

Evaluation of the Classification Performance of Our Drug Resistance Classifier

At the current stage of development, the proposed workflow only performs binary classification, meaning that each input sequence gets classified as either susceptible or resistant. Sequence data are used exclusively for the generation of the structural models on which dissimilarities are analyzed, but not to aid the classification itself. It is worth highlighting that our workflow relies on the detection of structural and chemical changes in viral enzymes that dictate susceptibility or resistance to drugs—rather than on the training of predictive models using sequences with known phenotypic response to drugs. Therefore, instead of using performance evaluation methods, such as cross-validation, that assess the impact of hiding a portion of training data (observations) on the accuracy of the resulting predictions, we resorted to the calculation of metrics of overall performance of our binary classifier.

The Receiver Operating Characteristic (ROC) curve was used to assess the overall discriminatory performance of our method. The score assigned to each dataset entry (here used for testing), corresponding to binding-site dissimilarities between each input *sequence-structure* pair and the template *consensus sequence-*structure, were thus plotted as a ROC curve. ROC curves are conceptually simple plots that depicts a binary classifier's discriminative capability as its discrimination threshold is varied. Such graphical plots are created by plotting the method's true positive rate (sensitivity) against its false positive rate (1-specificity), at varying thresholds. The area under the ROC curve (ROC AUC) value is a single scalar value varying between 0 and 1, providing a measure of the overall discriminatory power of the method. A ROC AUC value of 1 (or 100%) entails a *perfect* discrimination, a value of 0.5 represents random classification, while values above 0.8 are commonly accepted as indicators of an acceptable discriminatory performance (Fawcett, 2006; Pines and Everett, 2008; Powers, 2011; Tape). Furthermore, several performance measures, such as the Sensitivity (Equation 5), Specificity (Equation 6), Accuracy (Equation 7), and MCC (Equation 8) were also determined.

Figure 7 represents the obtained ROC curves and their respective ROC AUC values for the susceptible and resistant HIV1-PR binding-site MIF dissimilarities. ROC AUC values for [Res_{10}], [Res_{15}], and [Res_{20}] subsets were

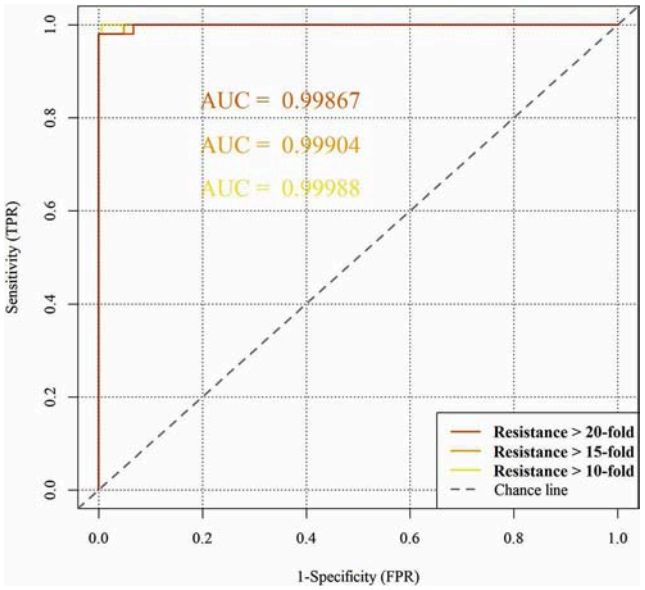


FIGURE 7 | Predictive performance of the binary classification (resistant vs. susceptible) produced by the algorithm/workflow presented herein, quantified by means of Receiver Operating Characteristic (ROC) curves and their respective Area Under the Curve (ROC AUC). The colors represent the ROC curves as follows: yellow for HIV1-PR sequences associated with 10-fold resistance; orange for HIV1-PR sequences associated with 15-fold resistance; and dark orange for HIV1-PR sequences associated with 20-fold resistance.

found to be similarly very high-—0.9999, 0.9990, and 0.9987, respectively – suggesting that the method holds significant discriminatory power to distinguishing susceptible from fully resistant HIV1-PR *sequence-structure* pairs—based on their binding-site MIF dissimilarities to the [Susceptible] reference *sequence-structure* pair.

We have also used ROC curve analysis to guide the definition of an *optimal* discrimination threshold based on Youden's index (Equation 9) (Youden, 1950). The optimal threshold observed corresponded to a 0.06 dissimilarity Tc for all [Res_{10}], [Res_{15}], and [Res_{20}] subsets. **Table 2** presents the values of each performance metric obtained for each subset, when applying a classification threshold of 0.06. At this classification cut-off, the specificities and sensitivities were found to be 0.997 and 0.994 for the [Res_{10}] subset, 0.997 and 0.952 for the [Res_{15}] subset and 0.997 and 0.933 for the [Res_{20}] subset, respectively. In all cases, there is strong discriminative performance toward susceptibility or resistance—as it can be appreciated by the high accuracy values highlighted in **Table 2**. Nevertheless, the best results are found for the [Res_{10}] subset, with an accuracy of about 0.997. On the other hand, the subsets with increasing degree of resistance, [Res_{15}] and [Res_{20}], show only slightly worst results concerning Sensitivity determined at a threshold of 0.06.

The overall predictive performance of our method was also evaluated by the Matthews correlation coefficient (MCC) on the three resistant subsets, which summarizes the sensitivity and the specificity of a classification method within a unique value, also varying between 0 and 1. A higher value of MCC indicates that the method has a better discriminatory performance. For the [Res_{10}], [Res_{15}], and [Res_{20}] groups, MCC values of 0.982, 0.874, and 0.833 were, respectively, obtained. Still, such performance metrics seems to highlight the clear potential of our MIF-based method to predict drug resistance, especially within the most populated [Res_{10}] group (MCC value close to 1).

TABLE 2 | Performance metrics obtained using a dissimilarity threshold of 0.0603.

Dissimilarity Threshold = 0.0603	Res_{10}	Res_{15}	Res_{20}
ROC AUC	0.99988	0.99904	0.99867
Sensitivity	0.994118	0. 951807	0. 933333
Specificity	0.992184	0. 992184	0. 992184
Accuracy	0.99669	0.996379	0.996366
MCC	0.98151	0.874199	0.833085

Positioning and Differentiation vs. Sequence-Based, PI-Resistance Prediction Tools

More than a decade ago, Lengauer and Sing pointed out the lack of commonly agreed benchmark (or test) datasets to assess and compare the performance of different prediction methods (Lengauer and Sing, 2006). The amount of available information on matched HIV genotype–resistance phenotype has increased significantly over recent years, with HIVDB embodying an important role as a centralized data repository (Rhee et al., 2003). As expected, sequence-based methods can make use of as much information as available to train their predictions, resulting in that they become proficient at "predicting" the phenotypic response for the sequences they have been trained on. Only in a few cases do we witness a concern in drawing prospective validation on unseen sequence sets and in making those test sets available to the community (Tarasova et al., 2018). This hinders the design of fair comparisons with methods that do not make direct use of sequence data for training, such as the one we propose here. On the other hand, over the past years genotypic-based methods have reached a level of sophistication that allows them to perform resistance predictions to specific drugs, exclusively based on sequence data matched to phenotypic response, while, at its current stage of development, our MIF-based method can only perform binary classification (susceptible or resistant) of input sequences.

Taken together, these aspects render the comparison of our algorithm with existing, sequence-trained, multi-classification predictors *non-trivial* to say the least. Further developments of our methodology, aiming at a more exhaustive exploration of specific MIF areas around the mutated binding sites, may enable stratification of classification into multiple drug classes by detecting the determinants of resistance to specific PIs. For the time being, we center the analysis of differentiation of our method on the answer to a recurrent question in the mind virologists or physicians who prescribe HIV-1 medications: *would it be possible to accurately predict whether a new, unknown HIV-1 strain will be susceptible to known PIs?*

TABLE 3 | Performance metrics for exemplary sequence-based prediction tools tested against the datasets compiled in this work.

PI-resistance predictor	Sensitivity$_{(A)}$	Sensitivity$_{(B)}$	Specificity	FN$_{(A)}$‡	FN$_{(B)}$‡	FP‡
HIV-GRADE 07/2019	1.0000	0.1471	0.9809	0	580 *12*	120 *8*
ANRS 29_11/2018	1.0000	0.1868	0.8493	0	553 *14*	945 *61*
HIVdb 8.9.1	1.0000	1.0000	0.8818	0	0	741 *21*
Rega 10.0.0	1.0000	0.0838	0.9804	0	623 *7*	123 *10*
MIF-based Drug Resistance Classifier†	0.9941	0.9941	0.9922	4 *1*	4 *1*	49 *11*

†*The proposed MIF-based drug resistance classifier is shown in the last row for comparison purposes.*

‡*False negatives (FN) corresponds to the number of sequences belonging to the Resistant* dataset (withdrawn of extreme outliers) that were predicted susceptible to all PIs. False positives (FP) corresponds to the number of sequences belonging to the Susceptible* dataset (withdrawn of extreme outliers) that were predicted resistant to at least one PI. In italics are indicated the number of viral isolates to which the sequences misclassified as FP belong. Rules for sensitivity analysis in (1) benchmark A [Sensitivity$_{(A)}$]: resistance to one or more PIs is considered a correct prediction; and (2) benchmark B [Sensitivity$_{(B)}$]: resistance to all PIs is considered a correct prediction.*

In order to answer to this question, we first converted our *test set* containing susceptible and resistant HIV1-PR sequences withdrawn of *extreme* outliers ($N = 6{,}269$ [[Susceptible*]] and $N = 680$ [[Resistant*]], respectively) into codon code, using the EMBOSS Backtranseq online tool (Madeira et al., 2019a,b), and then submitted it to the HIV-GRADE web server (Obermeier et al., 2012a,b) for comparison with the sequence-based algorithms ANRS-rules (Brun-Vézinet et al., 2003), HIVdb (Rhee et al., 2003; Tang et al., 2012) and Rega (Van Laethem et al., 2002; Camacho et al., 2017). Unexpectedly, we were not able to obtain predictions from *geno2pheno* via HIV Grade due to a technical issue of the web platform. To eschew this problem, we tried to submit the *test set* directly through *geno2pheno*'s web server, but the interface is limited to an unpractical maximum of 20 sequences per run.

Because the existing sequence-based interpretation systems try to predict phenotypical susceptibility or resistance to the individual drugs for a given genotype, whereas our approach only performs binary classification (susceptibility or resistance to all PIs), in order to draw comparison between the methods we tried to "level the playing field" by converting the predictions made by *sequence-based* algorithms into simpler binary classifications. In a first benchmark (benchmark A), the prediction outputs were converted into (i) susceptibility to all PIs ([Susceptible]) or (ii) resistance to any PI (*Resistant*). In a second, more challenging benchmark (benchmark B), the outputs were encoded as either (i) susceptible to all PIs (*Susceptible*) or (ii) resistant to all PIs (*Resistant*). The full list of criteria applied to the conversion of multiple classifiers into binary classification is given in **Supplementary Table S2**. The full raw output of HIV-GRADE is available in **Supplementary Datasheet S2**.

The ability to accurately predict the susceptibility of the input sequences to all PIs was assessed by determining the rate of correct predictions, with reflection into the calculated methods' Sensitivity (Equation 5) and Specificity (Equation 6). **Table 3** lists calculated performance metrics for the *sequence-based* algorithms on both benchmarks A and B, contrasted with the performance of our *sequence-to-structure*-, MIF-based algorithm. Sensitivity$_{(A)}$ and the number of detected false negatives FN$_{(A)}$ translate the methods' ability to classifying a HIV1-PR sequence known to be resistant to all PIs as *Resistant to at least one PI*. In contrast, Sensitivity$_{(B)}$ and FN$_{(B)}$ translate the methods' ability to correctly predict the same sequences (known to be resistant to all PIs) as *resistant to all PIs*. From the methods' sensitivity viewpoint, the assessment of the results of both benchmarks A and B has been important to counterbalance the crudeness of the conversion of a multiple classifier of resistance toward specific PIs into a binary classification. Benchmark A clearly biases sensitivity in favor of a multi-classifier by considering any resistance prediction (in number or kind of PI) for sequences known to be resistant to all PIs as *correct*, whereas benchmark B offers a more stringent evaluation of sensitivity wherein only *resistant-to-all-PIs* predictions for the same set of fully resistant sequences are considered as *correct*.

As expected the discriminatory power of the methods in benchmark A is in stark contrast with that calculated for benchmark B. Sensitivity$_{(A)}$ suggests that *sequence-based* methods slightly outperform our *sequence-to-structure*-, MIF-based classifier, with 100% correct predictions of *Resistant* sequences vs. a Sensitivity$_{(A)}$ value of 0.994 obtained by our method. By contrast, benchmark B shows a considerable drop in performance by *sequence-based* methods at correctly predicting HIV1-PR sequences resistant to all PIs—aside HIVdb, which retains a Sensitivity of 1.000.

The results in **Table 3** indicate that our workflow outperforms all other algorithms at identifying sequences susceptible to all PIs, with a Specificity of approximately 0.992, while its sequence-based counterparts display Specificities ranging from approximately 0.849 to 0.981. Still, it is worth noting that the large number of FP from the other *sequence-based* methods mostly come from the same isolates, similarly as mentioned above for FN$_{(B)}$. This fact highlights the advantage of accounting for structural information besides genotypic data. While MIFs allow searching for differences in the structural and physicochemical

environment of proteins, which might not be significantly affected by mutations for similar amino acids, *sequence-based* approaches will consistently search for mutations at positions of interest and consistently assign them the same classification. At an early-stage of development, our workflow's performance is quite satisfactory, considering that the ability of correctly classifying a sequence as susceptible to all PIs is a highly relevant step at the beginning of antiretroviral therapy—where a false positive weights more on the flexibility of first-line therapy regimens and, consequently, quality of life of the patient.

CONCLUSION AND FUTURE PERSPECTIVES

In recent years, the availability of data in the form of matched HIV genotype–resistance phenotype has expanded greatly, enabling further training of statistical learning methods relating genotype to different levels of phenotypic resistance and against specific drugs. However, in spite of the increased access to and routine sequencing of HIV's genome in many countries, as well as the constant evolution of machine learning (ML)-based techniques, HIV's high mutation rate (estimated in 3×10^{-5} per nucleotide per replication) will continue to pose significant challenges: not only in terms of the constant demand for curation of genotypic and phenotypic data to be fed into ML algorithms, but also from the viewpoint of the interpretability and translation of said data into knowledge to assist the design of novel anti-microbial agents. Therefore, the exploration of innovative structure-based *in silico* approaches to the prediction of drug resistance, focusing at the molecular interface that bridges to drug design, holds clear interest and appeal as alternative or complement to some of the most developed sequence-based statistical methods.

In this contribution, we propose a novel approach to drug resistance prediction, which captures structural and physicochemical modifications induced by mutations in the binding site of an extensively studied viral target, HIV1-PR. We demonstrate that, even at an early, proof-of-principle stage of development, our methodology can identify HIV1-PR *sequence-structure* pairs belonging to three levels of increasing resistance—with impressively high accuracy—thus anticipating, on a purely structural basis, whether a given HIV1-PR sequence will translate into phenotypic resistance or susceptibility to PIs. Since our *sequence-to-structure*-based classifier does not rely on *training* from genotypic data and only uses an individual input sequence to derive the corresponding viral enzyme structure and yield a prediction, its potential real-world value in supporting clinical decision is clearly relevant. Due to the fact that the proposed workflow produced predictions of complete drug susceptibility to the HIV1-PR datasets with high predictive accuracy, said results highlight this methodology as a potential valuable resource on clinical practice. Being able to use the clinical isolate sequence data to accurately predict susceptibility to known PIs, before starting a therapeutic regimen, is of paramount importance to allow the initiation of PI-based therapy with the less expensive 1st generation PIs, resulting in an economic benefit to the healthcare systems. Importantly, even though the method performs analysis on thousands of structural data points (atomic coordinates and MIF points), classification into *susceptible* or *resistant* takes place in a *couple-of-minutes* time scale.

It is worth emphasizing, nevertheless, that there is obvious room for methodological improvement and expansion. The upgrade to multi-classification functionality, where target structures known to be susceptible to specific inhibitors and drugs are used as template for structural modeling, is a critical milestone that will pave the way to predicting resistance to those specific anti-microbial agents. The growing amount of three-dimensional structural data on microbial target-inhibitor complexes, coupled with more elaborate use of sequence data, fuels our belief in that an improved *sequence-to-structure* -, MIF-based drug resistance classifier, will be able to combine the strengths and overcome the shortcomings of current approaches.

Claims of greatness must be backed by adequate validation designs. While the current version of our workflow does not allow drawing comprehensive and direct comparisons with more advanced sequence-based predictors of resistance to specific HIV1-PR inhibitors, further developments to our method will also be accompanied by the assembly and sharing of stratified benchmark sets of susceptible and resistant microbial target sequences—enabling fairer comparisons to be made both by ourselves and the scientific community.

As implied in our concluding words, a clear expectation around this work involves extending the application of our method to other targets, other than HIV1-PR, with inherent and multiple patterns of genetic variation. We realize, however, that this expectation may only be fulfilled if workable amounts of data are shared among the scientific community. Undoubtedly, one of the most critical aspects facing drug resistance prediction is the development of community-wide efforts to prepare and share useful datasets and tools to facilitate improvement and performance evaluation of existing and novel methodologies—which should be a clear priority for researchers working in the field. By basing its development on the use of freeware, our method is freely-available for non-commercial use.

To conclude, we see the results presented here as a promising example of the potential application of combined sequence- and structure-based *in silico* methods to achieve a more detailed interpretation and prediction of the impact of mutations in drug resistance. The ever-increasing emergence and widespread of drug-resistance calls in for the development of more efficient strategies to combat microbial threats in several fronts—be that in the drug discovery research setting or the clinical and medical therapeutic decision realm.

DATA AVAILABILITY STATEMENT

The datasets analyzed and scripts for this study can be found in the PI-resistance_Prediction GitHub [https://github.com/subject-am/PI-resistance_Prediction]. Raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

CS and RB developed the idea for the present work and provided critical revisions. NA, AM, and JL contributed equally to its conception, literature search, and manuscript writing. All authors contributed to manuscript revision, read, and approved the submitted version.

ACKNOWLEDGMENTS

NA, AM, JL, CS, and RB thank Daniela Vaz, João Vaz, and Vítor Duque for the informative discussions on the HIV1 drug resistance topic, which lead to devising and developing of the present work. NA, AM, JL, CS, and RB also thank the Coimbra Chemistry Centre (CQC) supported by the Portuguese Agency for Scientific Research, Foundation for Science and Technology (FCT), through Project UID/QUI/00313/2019. NA, AM, and JL acknowledge the MedChemTrain Ph.D. programme (PD/00147/2013) in Medicinal Chemistry—Ministry of Science, Technology, and Higher Education (MCTES), Portugal—for Ph.D. fellowships PD/BD/135287/2017, PD/BD/135289/2017 and PD/BD/135292/2017, respectively.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2020.00243/full#supplementary-material

REFERENCES

Agniswamy, J., Louis, J. M., Roche, J., Harrison, R. W., and Weber, I. T. (2016). Structural studies of a rationally selected multi-drug resistant HIV-1 protease reveal synergistic effect of distal mutations on flap dynamics. *PLoS ONE* 11:e0168616. doi: 10.1371/journal.pone.0168616

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Alves, N. G., Mata, A. I., and Luís, J. P. (2018a). *GitHub: FilterMajor.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/fasta-processing/filterMajor.sh (accessed October 27, 2019).

Alves, N. G., Mata, A. I., and Luís, J. P. (2018b). *GitHub: FilterMinor.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/fasta-processing/filterMinor.sh (accessed October 27, 2019).

Alves, N. G., Mata, A. I., and Luís, J. P. (2018c). *GitHub: MutModels.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/model-structures/MutModels.sh (accessed October 27, 2019).

Alves, N. G., Mata, A. I., and Luís, J. P. (2018d). *GitHub: Pattern_HIVp.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/model-structures/pattern-HIVp.sh (accessed October 27, 2019).

Alves, N. G., Mata, A. I., and Luís, J. P. (2018e). *GitHub: Ref_process.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/isomif_run/ref_process.sh (accessed October 27, 2019).

Alves, N. G., Mata, A. I., and Luís, J. P. (2018f). *GitHub: Separate_sets.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/Dataset-processing/Separate_sets.sh (accessed October 27, 2019).

Alves, N. G., Mata, A. I., and Luís, J. P. (2019a). *GitHub: Count_mut.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/fasta-processing/count_mut.sh (accessed October 27, 2019).

Alves, N. G., Mata, A. I., and Luís, J. P. (2019b). *GitHub: HIV1predict.sh*. Available online at: https://github.com/subject-am/PI-resistance_Prediction/blob/master/bin/HIV1predict.sh (accessed November 30, 2019).

Artese, A., Cross, S., Costa, G., Distinto, S., Parrotta, L., Alcaro, S., et al. (2013). Molecular interaction fields in drug discovery: recent advances and future perspectives. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 3, 594–613. doi: 10.1002/wcms.1150

Beerenwinkel, N., Däumer, M., Oette, M., Korn, K., Hoffmann, D., Kaiser, R., et al. (2003). Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res.* 31, 3850–3855. doi: 10.1093/nar/gkg575

Bonet, I. (2015). Machine learning for prediction of HIV drug resistance: a review. *Curr. Bioinform.* 10, 579–585. doi: 10.2174/1574893610666151008011731

Bron, C., and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM* 16, 575–577. doi: 10.1145/362342.362367

Brun-Vézinet, F., Descamps, D., Ruffault, A., Masquelier, B., Calvez, V., Peytavin, G., et al. (2003). Clinically relevant interpretation of genotype for resistance to abacavir. *AIDS* 17, 1795–1802. doi: 10.1097/00002030-200308150-00008

Camacho, R., Laethem, K., Van Geretti, A. M., Verheyen, J., Paredes, R., and Vandamme, A.-M. (2017). *Algorithm for the Use of Genotypic HIV-1 Resistance Data (Version Rega v10.0.0)*. Available online at: https://rega.kuleuven.be/cev/avd/software/rega-hiv1-rules-v10.pdf

Cao, Z. W., Han, L. Y., Zheng, C. J., Ji, Z. L., Chen, X., Lin, H. H., et al. (2005). Computer prediction of drug resistance mutations in proteins. *Drug Discov. Today* 10, 521–529. doi: 10.1016/S1359-6446(05)03377-5

Chartier, M., and Najmanovich, R. (2015). Detection of binding site molecular interaction field similarities. *J. Chem. Inf. Model* 55, 1600–1615. doi: 10.1021/acs.jcim.5b00333

Cruciani, G. (2005). *Molecular Interaction Fields*. Weinheim: FRG, Wiley-VCH Verlag GmbH and Co. KGaA.

draw.io. (2005). *Flowchart Maker & Online Diagram Software*. Available online at: https://www.draw.io/ (accessed October 25, 2019).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognit. Lett.* 27, 861–874. doi: 10.1016/j.patrec.2005.10.010

Florkowski, C. M. (2008). Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *Clin. Biochem. Rev.* 29(Suppl 1), S83–S87.

Gaudreault, F., Morency, L.-P., and Najmanovich, R. J. (2015). NRGsuite: a PyMOL plugin to perform docking simulations in real time using FlexAID. *Bioinformatics* 31, 3856–3858. doi: 10.1093/bioinformatics/btv458

Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* 28, 849–857. doi: 10.1021/jm00145a002

Hao, G., Yang, G., and Zhan, C. (2010). Computational mutation scanning and drug resistance mechanisms of HIV-1 protease inhibitors. *J. Phys. Chem. B* 114, 9663–9676. doi: 10.1021/jp102546s

Hao, G.-F., Guang-Fu, Y., and Zhan, C.-G. (2012). Structure-based methods for predicting target mutation-induced drug resistance and rational drug design to overcome the problem. *Drug Discov. Today* 17, 1121–1126. doi: 10.1016/j.drudis.2012.06.018

Hoaglin, D. C. (2003). John W. Tukey and data analysis. *Stat. Sci.* 18, 311–18. doi: 10.1214/ss/1076102418

Hou, T., and Yu, R. (2007). Molecular dynamics and free energy studies on the wild-type and double mutant HIV-1 protease complexed with amprenavir and two amprenavir-related inhibitors: mechanism for binding and drug resistance. *J. Med. Chem.* 50, 1177–1188. doi: 10.1021/jm0609162

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graph.* 14, 33–38. doi: 10.1016/0263-7855(96)00018-5

Irwin, K. K., Renzette, N., Kowalik, T. F., and Jensen, J. D. (2016). Antiviral drug resistance as an adaptive process. *Virus Evol.* 2:vew014. doi: 10.1093/ve/vew014

Jenwitheesuk, E., and Samudrala, R. (2005). Prediction of HIV-1 protease inhibitor resistance using a protein-inhibitor flexible docking approach. *Antivir. Ther.* 10, 157–166.

Khalid, Z., and Sezerman, O. U. (2018). Prediction of HIV drug resistance by combining sequence and structural properties. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 15, 966–973. doi: 10.1109/TCBB.2016.2638821

Khan, R. A., and Brandenburger, T. (2019). *ROCit: Performance Assessment of Binary Classifier with Visualization*. Available online at: https://cran.r-project.org/package=ROCit

King, N. M., Prabu-Jeyabalan, M., Nalivaika, E. A., Wigerinck, P., de Bethune, M.-P., and Schiffer, C. A. (2004). Structural and thermodynamic basis for the binding of TMC114, a next-generation human immunodeficiency virus type 1 protease inhibitor. *J. Virol.* 78, 12012–12021. doi: 10.1128/JVI.78.21.12012-12021.2004

Lengauer, T., and Sing, T. (2006). Bioinformatics-assisted anti-HIV therapy. *Nat. Rev. Microbiol.* 4, 790–797. doi: 10.1038/nrmicro1477

Madeira, F., Mi Park, Y., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al. (2019a). EMBOSS Backtranseq. *Nucleic Acids Res.* 47, W636–W641. doi: 10.1093/nar/gkz26

Madeira, F., Mi Park, Y., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al. (2019b). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 47, W636–W641. doi: 10.1093/nar/gkz268

Martínez, L., Andreani, R., and Martínez, J. M. (2007). Convergent algorithms for protein structural alignment. *BMC Bioinformatics* 8:306. doi: 10.1186/1471-2105-8-306

Mason, S., Devincenzo, J. P., Toovey, S., Wu, J. Z., and Whitley, R. J. (2018). Comparison of antiviral resistance across acute and chronic viral infections. *Antiviral Res.* 158, 103–112. doi: 10.1016/j.antiviral.2018.07.020

Masso, M., and Vaisman, I. I. (2013). Sequence and structure based models of HIV-1 protease and reverse transcriptase drug resistance. *BMC Genomics* 14:S3. doi: 10.1186/1471-2164-14-S4-S3

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta Protein Struct.* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

McKeegan, K. S., Borges-Walmsley, M. I., and Walmsley, A. R. (2002). Microbial and viral drug resistance mechanisms. *Trends Microbiol.* 10, 8–14. doi: 10.1016/S0966-842X(02)02429-0

Nayak, C., Chandra, I., and Singh, S. K. (2019). An *in silico* pharmacological approach toward the discovery of potent inhibitors to combat drug resistance HIV-1 protease variants. *J. Cell. Biochem.* 120, 9063–9081. doi: 10.1002/jcb.28181

Obermeier, M., Pironti, A., Berg, T., Braun, P., Däumer, M., Eberle, J., et al. (2012a). HIV-GRADE: a publicly available, rules-based drug resistance interpretation algorithm integrating bioinformatic knowledge. *Intervirology* 55, 102–107. doi: 10.1159/000331999

Obermeier, M., Pironti, A., Berg, T., Braun, P., Däumer, M., Eberle, J., et al. (2012b). HIV-GRADE: HIV-1 Tool. *Intervirology* 55, 102–107.

Pawar, S., Wang, Y.-F., Wong-Sam, A., Agniswamy, J., Ghosh, A. K., Harrison, R. W., et al. (2019). Structural studies of antiviral inhibitor with HIV-1 protease bearing drug resistant substitutions of V32I, I47V and V82I. *Biochem. Biophys. Res. Commun.* 514, 974–978. doi: 10.1016/j.bbrc.2019.05.064

Pines, J. M., and Everett, W. W. (2008). *Evidence-Based Emergency Care*, 2nd Edn, eds J. M. Pines, C. R. Carpentee, A. S. Raja, and J. D. Schuur. Oxford: Blackwell Publishing Ltd.

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-factor to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol.* 2, 37–63. Available online at: http://bioinfopublication.org/viewhtml.php?artid=BIA0001114

Qiu, X., and Liu, Z.-P. (2011). Recent developments of peptidomimetic HIV-1 protease inhibitors. *Curr. Med. Chem.* 18, 4513–4537. doi: 10.2174/092986711797287566

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Found. Stat. Comput. Available online at: https://www.r-project.org/

RCSB PDB (2000). *RCSB PDB*. Available at: https://www.rcsb.org/ (accessed September 7, 2019).

[Res_{10}] Alves, Nuno G., Ana Isabel Mata, and João Pedro Luís. (2019) HIV-1 protease isolates from HIVdb Stanford University with more than 10-fold resistance to all inhibitors. GitHub. 20181102. https://github.com/subject-am/PI-resistance_Prediction/blob/Dataset-processing/PI_dataset-res10

[Res_{15}] Alves, Nuno G., Ana Isabel Mata, and João Pedro Luís. (2019) HIV-1 protease isolates from HIVdb Stanford University with more than 15-fold resistance to all inhibitors. GitHub. 20181102. https://github.com/subject-am/PI-resistance_Prediction/blob/Dataset-processing/PI_dataset-res15

[Res_{20}] Alves, Nuno G., Ana Isabel Mata, and João Pedro Luís. (2019) HIV-1 protease isolates from HIVdb Stanford University with more than 20-fold resistance to all inhibitors. GitHub. 20181102. https://github.com/subject-am/PI-resistance_Prediction/blob/Dataset-processing/PI_dataset-res20

[Resistant*] Alves, Nuno G., Ana Isabel Mata, and João Pedro Luís. (2019) HIV-1 protease isolates from HIVdb Stanford University with more than 10-fold resistance to all inhibitors, without extreme outliers. GitHub. 20181102. https://github.com/subject-am/PI-resistance_Prediction/blob/Dataset-processing/Resistant_noOut

Rhee, S.-Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* 31, 298–303. doi: 10.1093/nar/gkg100

Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L., and Shafer, R. W. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc. Natl. Acad. Sci. U. S. A.* 103, 17355–17360. doi: 10.1073/pnas.0607274103

Riemenschneider, M., Hummel, T., and Heider, D. (2016). SHIVA - a web application for drug resistance and tropism testing in HIV. *BMC Bioinformatics* 17:314. doi: 10.1186/s12859-016-1179-2

Ruopp, M. D., Perkins, N. J., Whitcomb, B. W., and Schisterman, E. F. (2008). Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical J.* 50, 419–430. doi: 10.1002/bimj.200710415

Šali, A. (2019a). *Modeller*. Available at: https://salilab.org/modeller/ (accessed August 20, 2019).

Šali, A. (2019b). *Modeller Wiki*. Available online at: https://salilab.org/modeller/wiki/Mutate model (accessed September 30, 2019).

Šali, A., and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779–815. doi: 10.1006/jmbi.1993.1626

Sheik Amamuddy, O., Bishop, N. T., and Tastan Bishop, Ö. (2018). Characterizing early drug resistance-related events using geometric ensembles from HIV protease dynamics. *Sci. Rep.* 8:17938. doi: 10.1038/s41598-018-36041-8

Stanford University (1998a). *HIVDB: Consensus B Amino Acid Sequences*. Available online at: https://hivdb.stanford.edu/pages/documentPage/consensus_amino_acid_sequences.html (accessed November 2, 2019).

Stanford University (1998b). *HIVDB: Genotype-Phenotype Datasets*. Available online at: https://hivdb.stanford.edu/pages/genopheno.dataset.html (accessed November 2, 2018).

Stanford University (1998c). *HIVDB: HIV Drug Resistance Database*. Available online at: https://hivdb.stanford.edu/ (accessed October 23, 2019).

Stanford University (1998d). *HIVDB: PI Resistance Notes*. Available online at: https://hivdb.stanford.edu/dr-summary/resistance-notes/PI/ (accessed October 23, 2019).

Strasfeld, L., and Chou, S. (2010). Antiviral drug resistance: mechanisms and clinical implications. *Infect. Dis. Clin. North Am.* 24, 413–437. doi: 10.1016/j.idc.2010.01.001

[Susceptible] Alves, Nuno G., Ana Isabel Mata and João Pedro Luís. (2019) HIV-1 protease isolates from HIVdb Stanford University susceptible to all inhibitors. GitHub. 20181102. https://github.com/subject-am/PI-resistance_Prediction/blob/Dataset-processing/PI_dataset-susceptible

[Susceptible*] Alves, Nuno G., Ana Isabel Mata, and João Pedro Luís. (2019) HIV-1 protease isolates from HIVdb Stanford University susceptible to all inhibitors, without *extreme* outliers. GitHub. 20181102. https://github.com/subject-am/PI-resistance_Prediction/blob/Dataset-processing/Susceptible_noOut

Tang, M. W., Liu, T. F., and Shafer, R. W. (2012). The HIVdb system for HIV-1 genotypic resistance interpretation. *Intervirology* 55, 98–101. doi: 10.1159/000331998

Tape, T. G. (1990). *Interpreting Diagnostic Tests*. University of Nebraska Medical Center. Available online at: http://gim.unmc.edu/dxtests/ (accessed October 24, 2019).

Tarasova, O., Biziukova, N., Filimonov, D., and Poroikov, V. (2018). A computational approach for the prediction of HIV resistance based on amino acid and nucleotide descriptors. *Molecules* 23:2751. doi: 10.3390/molecules23112751

Toor, J. S., Sharma, A., Kumar, R., Gupta, P., Garg, P., and Arora, S. K. (2011). Prediction of drug-resistance in HIV-1 subtype C based on protease sequences

from ART naive and first-line treatment failures in North India using genotypic and docking analysis. *Antiviral Res.* 92, 213–218. doi: 10.1016/j.antiviral.2011.08.005

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics* 5:99. doi: 10.2307/3001913

Van Laethem, K., De Luca, A., Antinori, A., Cingolani, A., Perna, C. F., and Vandamme, A.-M. (2002). A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. *Antivir. Ther.* 7, 123–9.

Vercauteren, J., and Vandamme, A.-M. (2006). Algorithms for the interpretation of HIV-1 genotypic drug resistance information. *Antiviral Res.* 71, 335–342. doi: 10.1016/j.antiviral.2006.05.003

Vere Hodge, A., and Field, H. J. (2011). "General mechanisms of antiviral resistance," in *Genetics and Evolution of Infectious Disease*, ed M. Tibayrenc (Amsterdam: Elsevier), 339–362.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., et al. (2019). *gplots: Various R Programming Tools for Plotting Data*. Available online at: https://cran.r-project.org/package=gplots

Weber, I. T., and Agniswamy, J. (2009). HIV-1 protease: structural perspectives on drug resistance. *Viruses* 1, 1110–1136. doi: 10.3390/v1031110

Weber, I. T., and Harrison, R. W. (2016). Tackling the problem of HIV drug resistance. *Postepy Biochem.* 62, 273–279.

Wensing, A. M., Calvez, V., Ceccherini-Silberstein, F., Charpentier, C., Günthard, H. F., Paredes, R., et al. (2019). 2019 update of the drug resistance mutations in HIV-1. *Top. Antivir. Med.* 27, 111–121. Available at: http://www.ncbi.nlm.nih.gov/pubmed/31634862

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag. Available online at: https://ggplot2-book.org/. doi: 10.1007/978-0-387-98141-3

Wlodawer, A., and Erickson, J. W. (1993). Structure-based inhibitors of HIV-1 protease. *Annu. Rev. Biochem.* 62, 543–585. doi: 10.1146/annurev.bi.62.070193.002551

Word, J. M., Lovell, S. C., Richardson, J. S., and Richardson, D. C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* 285, 1735–1747. doi: 10.1006/jmbi.1998.2401

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer* 3, 32–35.

Yu, X., Weber, I. T., and Harrison, R. W. (2014). Prediction of HIV drug resistance from genotype with encoded three-dimensional protein structure. *BMC Genomics* 15:S1. doi: 10.1186/1471-2164-15-S5-S1

Zhang, J., Rhee, S.-Y., Taylor, J., and Shafer, R. W. (2005). Comparison of the precision and sensitivity of the antivirogram and phenosense HIV drug susceptibility assays. *J. Acquir. Immune Defic. Syndr.* 38, 439–444. doi: 10.1097/01.qai.0000147526.64863.53

Conflict of Interest: CS and RB are cofounders of the company BSIM Therapeutics, however all work reported in this article was carried out at the University of Coimbra.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Alves, Mata, Luís, Brito and Simões. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 29 May 2020
doi: 10.3389/fchem.2020.00443

In silico Druggability Assessment of the NUDIX Hydrolase Protein Family as a Workflow for Target Prioritization

Maurice Michel[1]*, *Evert J. Homan*[1], *Elisée Wiita*[1], *Kia Pedersen*[1], *Ingrid Almlöf*[1], *Anna-Lena Gustavsson*[2], *Thomas Lundbäck*[2,3], *Thomas Helleday*[1,4] and *Ulrika Warpman Berglund*[1]*

[1] Science for Life Laboratory, Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden, [2] Chemical Biology Consortium Sweden (CBCS), Science for Life Laboratory, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden, [3] Mechanistic Biology and Profiling, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden, [4] Department of Oncology and Metabolism, Sheffield Cancer Centre, University of Sheffield, Sheffield, United Kingdom

OPEN ACCESS

Edited by:
Simone Brogi,
University of Pisa, Italy

Reviewed by:
Ariel Fernandez,
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina
Mohane S. Coumar,
Pondicherry University, India
Claudio Norberto Cavasotto,
Austral University, Argentina

***Correspondence:**
Maurice Michel
maurice.grube@scilifelab.se
Ulrika Warpman Berglund
ulrika.warpmanberglund@scilifelab.se

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: 23 October 2019
Accepted: 28 April 2020
Published: 29 May 2020

Citation:
Michel M, Homan EJ, Wiita E, Pedersen K, Almlöf I, Gustavsson A-L, Lundbäck T, Helleday T and Warpman Berglund U (2020) In silico Druggability Assessment of the NUDIX Hydrolase Protein Family as a Workflow for Target Prioritization. Front. Chem. 8:443. doi: 10.3389/fchem.2020.00443

Computational chemistry has now been widely accepted as a useful tool for shortening lead times in early drug discovery. When selecting new potential drug targets, it is important to assess the likelihood of finding suitable starting points for lead generation before pursuing costly high-throughput screening campaigns. By exploiting available high-resolution crystal structures, an *in silico* druggability assessment can facilitate the decision of whether, and in cases where several protein family members exist, which of these to pursue experimentally. Many of the algorithms and software suites commonly applied for *in silico* druggability assessment are complex, technically challenging and not always user-friendly. Here we applied the intuitive open access servers of DoGSite, FTMap and CryptoSite to comprehensively predict ligand binding pockets, druggability scores and conformationally active regions of the NUDIX protein family. In parallel we analyzed potential ligand binding sites, their druggability and pocket parameter using Schrödinger's SiteMap. Then an *in silico* docking cascade of a subset of the ZINC FragNow library using the Glide docking program was performed to assess identified pockets for large-scale small-molecule binding. Subsequently, this initial dual ranking of druggable sites within the NUDIX protein family was benchmarked against experimental hit rates obtained both in-house and by others from traditional biochemical and fragment screening campaigns. The observed correlation suggests that the presented user-friendly workflow of a dual parallel *in silico* druggability assessment is applicable as a standalone method for decision on target prioritization and exclusion in future screening campaigns.

Keywords: druggability, nudix, drug discovery, workflow, malachite green

INTRODUCTION

The nucleoside diphosphates attached to sequence-x (NUDIX) hydrolase protein family was recently comprehensively and exhaustively reviewed by Carreras-Puigvert et al. (2017) NUDIX proteins possess a conserved sequence, called the NUDIX box, i.e., $Gx_5Ex_5[UA]xREx_2EExGU$), which differs little between individual members which are otherwise of low sequence similarity.

Structural and domain analysis revealed three major groups and one outlier, NUDT22, mostly based on their already reported activity against substrate classes such as diphosphoinositol polyphosphates (Caffrey et al., 1999, 2000) and NADH diphosphates (Abdelraheim et al., 2003). Subsequently, a systematical screening against a large set of substrates was performed and painted a rather promiscuous picture of the NUDIX hydrolases, indicating backup functionality or redundancy. Consequently, a global expression analysis was performed and showed a clear dependency on tissue of origin and the corresponding cancer tissue. Interestingly, NUDT1, NUDT5, and NUDT14 amongst others were present in a cluster of highly expressed proteins, confirming a potential role in cancer as reported earlier (Choi et al., 2011; Gad et al., 2014; Huber et al., 2014; Wright et al., 2016). Importantly, when evaluated for epistasis, it became apparent that several NUDIX members sustain relations as measured in cell viability and cell cycle perturbations and that these interactions are more important for cancerous cells. With this overview in structure, expression, substrate specificity and relation, the NUDIX protein family members gained considerable attention as potential drug targets. The original interest in pharmacological modulation of NUDIX members was sparked by the notion that NUDT1 is overexpressed in several cancer cell types, while its role in healthy cells can largely be compensated for as evidenced by the normal life-span of knock-out mice (Tsuzuki et al., 2001). Besides GTP and dGTP, NUDT1 hydrolyzes several oxidatively damaged DNA nucleotides including 8-oxo-dGTP and 2-OH-dATP, thus preventing their incorporation into DNA, which otherwise would lead to DNA damage and ultimately cell death. This led to the hypothesis that increased expression of NUDT1, and hence improved sanitization capacity of oxidatively damaged DNA bases from the nucleotide pool, would enable cancer cells to cope with the increased oxidative stress they are exposed to compared with healthy cells. Gad and coworkers published TH588 (**Figure 1**) as the first small-molecule NUDT1 inhibitor with efficacy in mouse xenograft models (Gad et al., 2014), although subsequent potent and selective NUDT1 inhibitors disclosed by AstraZeneca, MD Anderson, Gilead and Sprint Bioscience/Bayer failed to reproduce these findings with regards to cytotoxicity (**Figure 1**) (Kettle et al., 2016; Petrocchi et al., 2016; Ellermann et al., 2017; Farand et al., 2020). The validity of NUDT1 as an anticancer target has thus been questioned and is still under debate (Warpman Berglund et al., 2016; Samaranayake et al., 2017). Regardless, these studies served to demonstrate significant amenability to small-molecule inhibition of NUDT1, justifying the question as to how this translates to other members of the NUDIX family.

Besides NUDT1, a series of potent, drug-like NUDT5 inhibitors have been described by Page and coworkers (Page et al., 2018). The lead compound TH5427 (**Figure 1**) was shown to block progestin-dependent, PAR-derived nuclear ATP synthesis and subsequent chromatin remodeling, gene regulation and proliferation in breast cancer cells, suggesting that targeting NUDT5 may represent a novel therapeutic approach for breast cancer treatment. Most recently, the covalent NUDT7 inhibitor NUDT7-COV-1 was developed by employing electrophilic fragment screening and a fragment combination approach

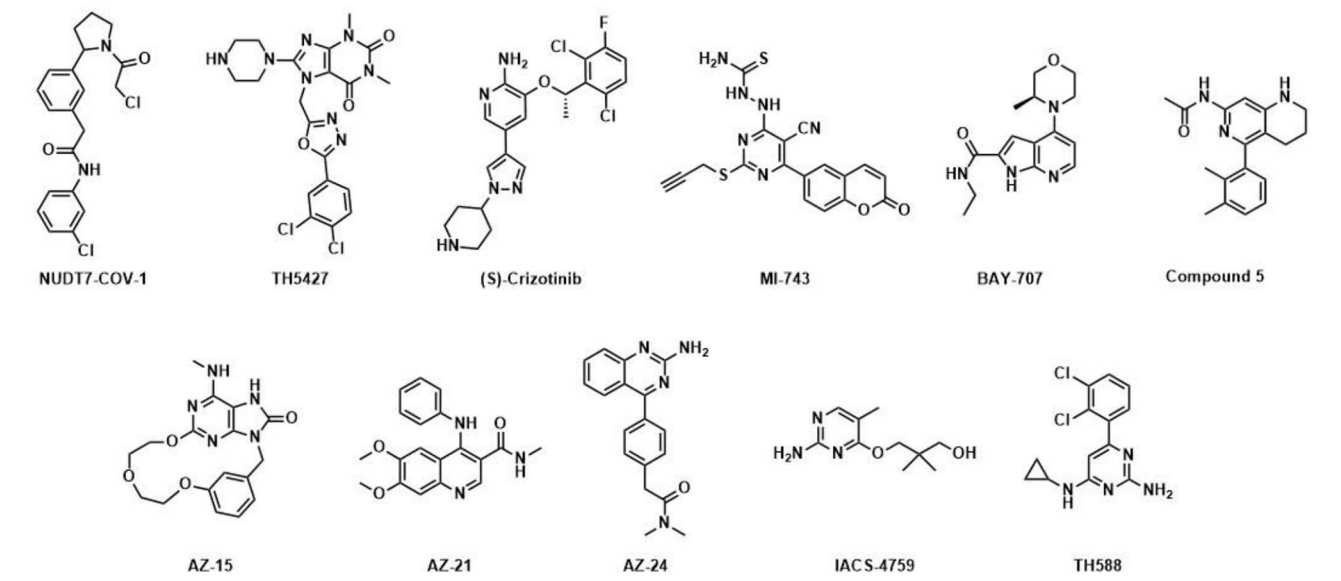


FIGURE 1 | Published NUDIX inhibitors: TH588 was developed as a first in class NUDT1 inhibitor at Science for Life Laboratory and Karolinska Institutet (Gad et al., 2014); (S)-Crizotinib is a potent NUDT1 inhibitor and the enantiomer of (R)-Crizotinib (Huber et al., 2014), a clinically applied tyrosine kinase inhibitor; optimized by Astra-Zeneca; AZ-15, AZ-21 and AZ-24 are distinct chemotype inhibitors targeting NUDT1 (Kettle et al., 2016); BAY-707 (Ellermann et al., 2017) was discovered as a NUTD1 inhibitor by Sprint Bioscience; IACS-4759 (Petrocchi et al., 2016) is a NUDT1 inhibitor developed by MD Anderson; MI-743 is a selective inhibitor of NUDT1 in gastric cancers (Zhou et al., 2019); Compound 5 was reported by Gilead and inhibits NUDT1 (Farand et al., 2020); TH5427 was synthesized as a lead compound against NUDT5 (Page et al., 2018); NUDT7-COV-1 is a covalent inhibitor generated by electrophile screening and fragment combination (Resnick et al., 2019).

(**Figure 1**) (Resnick et al., 2019). To the best of our knowledge no potent inhibitors for any of the other NUDIX hydrolase members have been disclosed to date, although there are public data on hit rates for selected family members on the Structural Genomics Consortium homepage[1, 2].

One aspect not addressed in the recent comprehensive review (Carreras-Puigvert et al., 2017) is an assessment of the potential druggability of the different NUDIX family members, i.e., their amenability to be modulated by drug-like small molecules. With the recent dawn of PROTACs, synthetic neoantigens and biologics, but also established targeting strategies like allosteric modulation or active site inhibition, several scenarios of how to target a protein may be exploited. With that in mind, druggability as such is no longer restricted to active site inhibition of a protein by a small molecule with an optimized small-molecule drug-like profile. Both orthosteric or catalytic sites and secondary, allosteric sites, may be equally interesting to be targeted for the development of small-molecule chemical probes and potential drug candidates. As high-resolution crystal structures of 18 out of the 22 human NUDIX hydrolases are now available, a family-wide *in silico* druggability assessment for available sites is feasible.

Here we use several open-access binding site analysis methods, i.e., DoGSite (Volkamer et al., 2012)[3], CryptoSite (Cimermancic et al., 2016)[4], and FTMap (Kozakov et al., 2015)[5], as well as the commercial SiteMap and *in silico* fragment screening of a fragment library using Glide to probe the NUDIX hydrolase protein structures for potential small-molecule binding sites and assess their druggability and suitability for a prospective drug discovery campaign. This established *in silico* prioritization workflow within the NUDIX family is further supported by results obtained from biochemical screens employing the malachite green assay (Baykov et al., 1988) as well as differential scanning fluorimetry (DSF) (Niesen et al., 2007) fragment screens for some of the family members. This correlation with own experimental results and those published previously highlights the benefit of this comparably low-cost computational assessment workflow prior to applying experimental screening methods for the rapid evaluation of target druggability.

MATERIALS AND METHODS

Protein Preparation and Validation

Available crystal structures of human NUDIX hydrolases with the highest resolution were imported into Maestro (Schrödinger Suite 2019-1, Schrödinger, LLC, New York, NY, 2019.) The structures were then prepared using the Protein Preparation Wizard as implemented in the Schrödinger Suite. Briefly, raw PDB structures were processed by automatically assigning bond orders, adding hydrogens, creating zero-order bonds to metals, converting selenomethionine to methionine, adding missing side-chains, creating possible disulfide bridges, deleting waters beyond 5.0 Å of hetero groups (if present), and generating hetero protonation states at pH 7.0. Residues with alternate positions were locked in the conformations with the highest average occupancy. Small ligands and metal ions originating from crystallization buffer were removed. The hydrogen bonding networks were optimized automatically, by sampling water orientations and optimization of hydroxyls, Asn, Gln, and His residue states using ProtAssign. Any remaining water molecules were subsequently removed. A restrained minimization was then performed using the OPLS3e force field, until an RMSD convergence of 0.30 Å was reached for the heavy atoms. Finally, the minimized NUDIX structures were aligned to the structure of NUDT1 (3Q93) with respect to the backbone atoms of the A chain.

DoGSite

The protein structures as prepared above were exported as PDB files, uploaded to the DoGSite server and assessed for binding sites and their corresponding DrugScores according to the published protocol (Volkamer et al., 2012). Pocket Size and DrugScores were extracted for all identified sites and annotated to pocket numbers.

FTMap

All prepared PDB files were uploaded to the FTMap server and interrogated for number of probes per cluster found according to the published protocols (Kozakov et al., 2015; Vajda et al., 2018).

CryptoSite

All prepared PDB files were uploaded to CryptoSite server and assessed for amino acid flexibility according to the published protocol (Cimermancic et al., 2016). Amino acid residues exceeding a Cryptic Site Score of 0.10 were extracted.

SiteMap

Prepared protein structures were submitted to SiteMap analyses as implemented in Schrödinger Suite 2019-1. The 5 top-ranked potential binding sites were identified. At least 15 site points per reported sites were required. The more restricted definition of hydrophobicity together with a standard grid (0.7 Å) were used. Site maps at 4 Å or more from the nearest site points were cropped. Clustering of the SiteMap parameters was performed using the heatmaply library in R[6]. The SiteMap parameters were transformed using "percentize," and average linking was used for clustering.

Virtual Fragment Screening

1) Fragment subset selection: a subset of the ZINC Frags Now set (Irwin et al., 2012) was created by applying a number of filters implemented in a Knime workflow (Knime 3.5.2, Berthold et al., 2008). Foremost, only fragments available from a list of 19 preferred suppliers, composed by a team of experienced

[1] *Fragment Screening*. Available online at: https://www.thesgc.org/fragment-screening (accessed August 8, 2019).
[2] NUDT15. *Manuscript Submitted*.
[3] *Zentrum für Bioinformatik: Universität Hamburg - Proteins Plus Server*. Available online at: https://proteins.plus/ (accessed June 3, 2019).
[4] *Cryptic binding site*. Available online at: https://modbase.compbio.ucsf.edu/cryptosite/ (accessed June 3, 2019).
[5] *FTMap: A Small Molecule Mapping Server*. Available online at: http://ftmap.bu.edu/login.php (accessed June 3, 2019).
[6] *Introduction to Heatmaply*. Available online at: https://cran.r-project.org/web/packages/heatmaply/vignettes/heatmaply.html (accessed April 16, 2020).

medicinal chemists were considered. These were then filtered using a cascade of structural filters, including REOS (Walters and Murcko, 2002), PAINS (Baell and Holloway, 2010) and a set of in-house filters (ScrapFilter) compiled over the years. Lipinski-type descriptors (SlogP, TPSA, AMW, NumLipinskiHBA, NumLipinskiHBD, NumRotatableBonds, NumHeavyAtoms, NumRings, NumAromaticRings) were then calculated using the RDKit Descriptor Calculation node. An additional descriptor HetRatio was then calculated as the ratio of NumLipinskiHBA and NumHeavyAtoms, and fragments with HetRatio <0.2 or >0.5 were filtered out. Finally, remaining outliers were removed by applying Gaussian Z-score normalization on the descriptor space, and then filtering out fragments with descriptor values deviating more than 3 units from the mean. The entire filtering cascade reduced the original input file of 704,041 structures as downloaded from ZINC to 205,891 fragments (**Supplementary Data Sheet 1**).

2) Ligand preparation: the selected fragment subset was then prepared for docking using LigPrep (Schrödinger): the OPLS3e force field was used for minimizations; possible ionization states at pH 7.0 ± 2.0 were generated using Epik (Shelley et al., 2007; Greenwood et al., 2010), metal binding states were added, and tautomers were generated; specified chiralities were retained and at most 4 stereoisomers were generated per structure. This yielded 345,044 structures for docking.
3) Receptor grid generation: Glide docking grids (Friesner et al., 2004, 2006; Halgren et al., 2004) were generated for each target protein by focusing the grid box on the center of the site with the highest Dscore as determined by SiteMap (Halgren, 2007, 2009). The size of the box enclosing the grid was set to 16 Å. No constraints, rotatable groups or excluded volumes were defined.
4) Virtual screening: The Virtual Screening Workflow as implemented in Schrödinger Suite was used for docking, scoring, and ranking of the top-1,000 fragments against the sites with the highest Dscore as determined by SiteMap. The workflow comprised a cascade of docking steps with increased accuracy (Glide HTVS → SP → XP), where the top-10% ranked ligands are passed on to the next step. After Glide XP docking the top-1,000 ranked fragments were retained for druggability assessment based on their combined docking scores.

Biochemical Screening

Small-molecule screening of NUDT2, NUDT15, and NUDT16 at a compound concentration of 10 μM was conducted using coupled enzymatic assays as already described for NUDT1 (Gad et al., 2014) and NUDT5 (Page et al., 2018). In brief this involved the purification of recombinant proteins following overexpression in *E. coli* and subsequent validation of coupled enzymatic assays based on cognate substrates for each of these [Ap4A for NUDT2, dGTP for NUDT15 and ADP for NUDT16 (Trésaugues et al., 2015)][2]. The assays for NUDT2 and NUDT15 were based on enzymatic release of inorganic pyrophosphate and subsequent degradation to two molecules of inorganic phosphate in the presence of excess inorganic pyrophosphatase. Levels of inorganic phosphate are measured using an established procedure for such measurements in 384-well format in our lab (see e.g., Gad et al., 2014; Page et al., 2018). The screening of NUDT16 was based on enzymatic processing of ADP to release one molecule of inorganic phosphate, such that the coupled enzyme was not needed in this assay. All assays were optimized to allow their application at close to the K_m of each substrate and with an incubation time chosen to ensure consumption of <30% of substrate and near linearity of assay signal increase with time.

Slightly different screening sets have been applied for the family members, with only a smaller subset of 5,500 compounds in common. All screens conducted at Chemical Biology Consortium Sweden have 16 each of negative (DMSO only – 0% inhibition) and positive controls (no enzyme or inhibitor at concentration that gives 100% inhibition). These are located in columns 23 and 24 of the 384-well plates and they are used to normalize the response in each well-containing library compounds to a % inhibition value. Hit limits are defined based on the average plus three standard deviations of the response for all library compounds and hit rates are provided as the percentage of library compounds above this limit. The malachite green assay has been extensively used for screening purposes in our lab as it is associated with low interference rates, as evidenced by the lack of common hits appearing in screens of NUDT1 (Gad et al., 2014), NUDT5 (Page et al., 2018), dCTPase (Llona-Minguez et al., 2016), dUTPase and ITPase besides the herein reported NUDIX proteins (**Supplementary Material** – Screens using malachite green).

DSF Fragment Screening

NUDT1, NUDT2, NUDT5, and NUDT15 druggability was further experimentally assessed through fragment screening by DSF. Different fragments sets were screened over time, reflecting history and development of the available fragment sets. The initial fragment library comprised 450 fragments selected from the Chemical Biology Consortium Sweden reagent store at the Karolinska Institutet, and this set was screened against NUDT1 and NUDT5. Over time this library was complemented with sets of nucleobase analogs acquired from the NCI Developmental Therapeutics Program, which was grown to a subset of 200 compounds. This set, together with the 450-member library, thus totaling 650 fragments, was screened against NUDT2. Subsequently the 450-member library was complemented with 550 additional fragments from the Chemical Biology Consortium Sweden reagent store in order to generate a more diverse generic fragment library of 1,000 compounds. This second version together with the 200 nucleobase analogs acquired from NCI was screened against NUDT15. The proteins were expressed and purified as previously reported (Carreras-Puigvert et al., 2017). Fragment screening by DSF was essentially performed as described in detail by Niesen et al. (2007) All fragments were screened at a final concentration of 500 μM. Positive controls for each target were used at 100 μM. Assay buffer was composed of 100 mM Tris Acetate, 40 mM NaCl, and 10 mM Mg Acetate. Sypro Orange (S6650, Molecular Probes, 5000x) was used as the fluorescent dye. Native melting points of the proteins

under the assay conditions were 50.0, 50.0, 76.0, and 57.0°C for NUDT1, NUDT2, NUDT5, and NUDT15, respectively. Screening was performed in 96-well Q-PCR plates using a BioRad 96CFX real-time PCR detection system with temperature increments of 1.0°C. More details of the assay conditions for each target are provided in the **Supplementary Material** – Fragment screen conditions.

RESULTS AND DISCUSSION

Automated Arm - Step 1: DoGSite and FTMap Predict Druggable Catalytic Sites and Potentially Druggable Secondary Sites

We started by compiling a list of available high-resolution crystal structures of human NUDIX proteins (**Table 1**). Due to the systematic work of the Structural Genomics Consortium, the majority of structures were solved with high sequence coverage (**Supplementary Material** – SiteMap secondary sites) and are often available together with screening data[1]. PDBs were imported to Maestro and prepared as described in the Method part. To enable application in the automated workflow, the prepared proteins were exported as new PDB files (**Figure 2**). In a first step, these files were uploaded to the DoGSite server. DoGSite is a web-based open-access algorithm that interrogates rigid protein structures for binding hotspots, including druggability prediction (Volkamer et al., 2012). Initially, a grid covering the protein identifies grid points that overlap with protein atoms. Application of a difference of Gaussian (DoG) filter then screens for preferred binding spots of sphere-like objects. Combination of several hotspots creates subpockets, which, if neighboring, are merged into a pocket. Several geometric and physico-chemical properties are automatically calculated for the predicted pockets and subpockets. A machine learning model trained on a set of known druggable proteins is then used to predict the druggability of the pockets, expressed as DrugScore. Reported as a factor between 0 and 1.0 a DrugScore over 0.5 and closer to 1.0 corresponds to good druggability.

Application of this algorithm to NUDIX crystal structures identified between two and ten pockets with a wide range of DrugScores (**Figure 3** and **Supplementary Material** - DogSite). Between one and four bindings sites were judged druggable by the DoGSite algorithm. For some of the NUDIX hydrolases the natural substrates and their binding sites are yet to be deciphered. In addition, with the broad targeting possibilities provided by PROTACs (An and Fu, 2018) or allosteric inhibitors (Wenthur et al., 2014; Aretz et al., 2018), it is not necessarily required to target a catalytic pocket to convey a desired phenotype. Thus, the single highest-ranking site of each NUDIX structure, often corresponding to the known substrate binding site, was used to calculate a NUDIX druggability score. With an average druggability score of 0.80, the NUDIX family of proteins qualify as good predicted drug targets. As a positive control and validated target when it comes to chemical amenability, NUDT1 (3Q93) reaches a similar score of 0.81. The protein tyrosine phosphatase 1B (PTP1B) was included into the assessment

TABLE 1 | High-resolution crystal structures used in this study.

Protein name	PDB code	References
NUDT1, MTH1	3Q93	Tresaugues et al., 2011a
NUDT2, APAH1	3U53	Ge et al., 2013
NUDT3, DIPP1	2FVV	Thorsell et al., 2009
NUDT4, DIPP2	5LTU	Srikannathasan et al., 2017a
NUDT5, HSPC115	6GRU	Dubianok et al., 2018
NUDT6, FGF2AS	3H95	Tresaugues et al., 2009a
NUDT7	5T3P	Srikannathasan et al., 2017b
NUDT9	1Q33	Shen et al., 2003
NUDT10, DIPP3A	3MCF	Tresaugues et al., 2010
NUDT12	6SCX	Wu et al., 2019
NUDT14, UGPP	3Q91	Tresaugues et al., 2011b
NUDT15, MTH2	5BON	Carter et al., 2015
NUDT16	3COU	Tresaugues et al., 2008
NUDT17	5LF8	Mathea et al., 2017a
NUDT18, MTH3	3GG6	Tresaugues et al., 2009b
NUDT20, DCP2	5MP0	Mathea et al., 2017b
NUDT21	3BAP	Coseno et al., 2008
NUDT22	5LF9	Tallant et al., 2017
PTP1B*	2HNP	Barford et al., 1994

**Added as reference protein.*

(2HNP) as this is generally known to be a challenging target for classical drug discovery approaches. PTP1B, like other tyrosine phosphatases, contains a relatively polar substrate pocket which can accommodate phosphate isosteres. In the last two decades, small molecules targeting this pocket have been shown to fail eliciting sufficient effects *in vivo* (Zhang and Zhang, 2007; Krishnan et al., 2018). Instead a non-classical approach of allosteric inhibition is currently under evaluation in clinical trials (Mullard, 2018). When interrogated with DoGSite, PTP1B (2HNP) scores 0.72 only by combination of two subpockets through a narrow channel.

An interesting observation is that all NUDIX members, except NUDT4 (0.72, 5LTU) and NUDT18 (0.63, 3GG6), individually score a high DrugScore around 0.80. Furthermore, it can be observed that several members, including NUDT6 (3H95, 0.78), NUDT7 (5T3P, 0.77), NUDT9 (1Q33, 0.82), NUDT17 (5LF8, 0.79), NUDT12 (6SCX, 0.85), and NUDT22 (5LF9, 0.81, **Figure 4**), are predicted to possess a second high-ranking pocket. These sites may increase the potential for pharmacological targeting of the corresponding proteins, for instance by masking a protein-protein interaction or a cofactor binding site. Identification of a second high-ranking pocket remote from the catalytic site, however, may also point toward an artifact in the crystal structure due to the construct used for expression or lack of electron density. For a comparison of resolved and expressed sequences please refer to **Supplementary Material** – SiteMap secondary sites. Thus, when inspected for their location, it became apparent that secondary sites can be distinguished as either neighboring to the top-ranked site or being located more remotely.

The druggability of the identified pockets can be further assessed using FTMap (Kozakov et al., 2015; Yueh et al., 2019). FTMap interrogates the protein surface for contributions

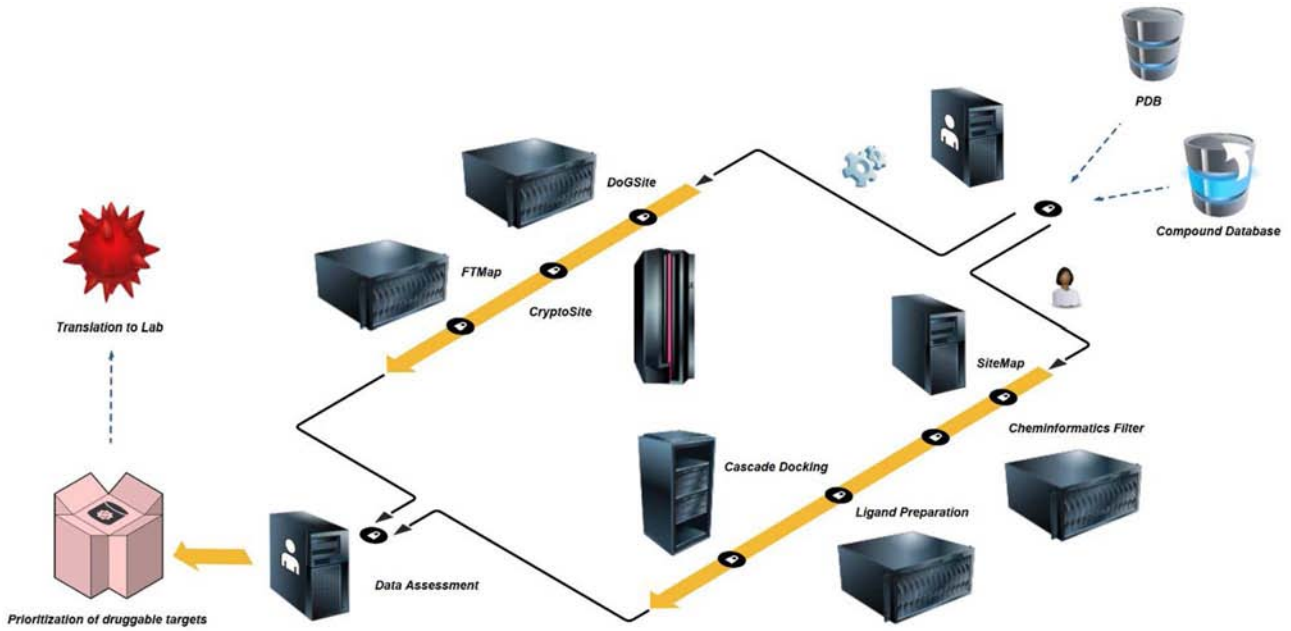


FIGURE 2 | *In silico* druggability prioritization workflow – High-resolution crystal structures and comprehensive commercial space compound databases are freely available. PDBs are prepared by protein preparation wizard (Thorsell et al., 2009), inspected for structural flaws and limitations and exported as new structure files. In a first step, the structures are assessed for binding sites and site druggability by algorithms using rigid proteins or allowing for flexible behavior, i.e. DoGSite (Volkamer et al., 2012) and CryptoSite (Cimermancic et al., 2016). To eliminate artifacts due to protein construct choice, limitations of crystal structure resolution or co-crystallization of additional protein copies, all identified sites are counter-screened by FTMap (Kozakov et al., 2015). In a second and parallel step, the same structures are initially assessed for binding sites, druggability and pocket parameter using SiteMap (Halgren, 2007, 2009). A separate Knime (Berthold et al., 2008) workflow for the elimination of promiscuous functionalities is followed by Ligand Preparation which builds an applicable set of small molecules including a number of tautomers and stereoisomers. In a three-step cascade this set is then docked (Friesner et al., 2004, 2006; Halgren et al., 2004) against the highest-ranking site as identified by SiteMap. The median docking score of the top-1,000 fragments is used to assess druggability based on commercial fragment space. In a final step, prioritization of targets passing both parallel screening schemes may be performed based on published experimental screening data or own future screening efforts during translation to the lab.

to ligand-free energy. Small organic molecules, reflecting the complexity of potential active substances, are scored using a detailed energy function. Some regions bind several clusters of probes and thus identify as a binding hotspot. Earlier, this orthogonal method was applied on pockets identified by CryptoSite (Vajda et al., 2018), where high druggability would correspond to an FTMap cluster populating these sites and containing at least 16 probes. When similarly examined for the number of bound probes, all highest-ranking sites of each NUDIX protein except for NUDT4 (5LTU) and NUDT18 (3GG6), reached more than 16 probes confirming the good druggability of the expected active sites of the enzyme family (**Figure 3** and **Supplementary Material** - FTMap). NUDT4 (5LTU) and NUDT18 (3GG6), which showed a lower DrugScore before, failed to contain more than 16 probes and are the only family members with a lower druggability assessment based on DoGSite and FTMap. Assessment of PTP1B (2HNP) returned all FTMap probe clusters to be located in the smaller of the two sites predicted by DoGSite (DrugScore 0.38). When evaluated with FTMap, secondary sites of NUDT6 (3H95), NUDT7 (5T3P), NUDT9 (1Q33), NUDT17 (5LF8), and NUDT22 (5LF9, **Figure 4**) neighboring the highest-ranking site tend to harbor more probes than those sites found remotely. All remote secondary sites, i.e., NUDT9 (1Q33), NUDT17 (5LF8) and NUDT 12 (6SCX), fail to incorporate the required 16 probes. Of those located much closer to the highest-ranking pocket, only NUDT7 (5T3P) fails to accommodate 16 or more probes underscoring the potential use in pharmacological targeting additionally to the neighboring highest-ranking pocket.

Automated Arm – Step 2: CryptoSite and FTMap Confirm Druggable Active Binding Pockets With High Flexibility

Druggability predictions using DoGSite are based on rigid protein structures, not allowing for flexibility typically induced by larger natural substrates or specifically designed small molecules (Michel et al., 2019). Another aspect is the potential existence of allosteric sites. Typically, a crystal structure of a compound bound to the allosteric site or comprehensive protein dynamics calculations based on several distinct crystal structures are required for their discovery. The CryptoSite algorithm however, can give first insights in whether an already identified active site or a shallow pocket allows for high single amino acid flexibility (Cimermancic et al., 2016). Networks of these flexible cryptic sites could indicate concerted movements of the protein, possibly

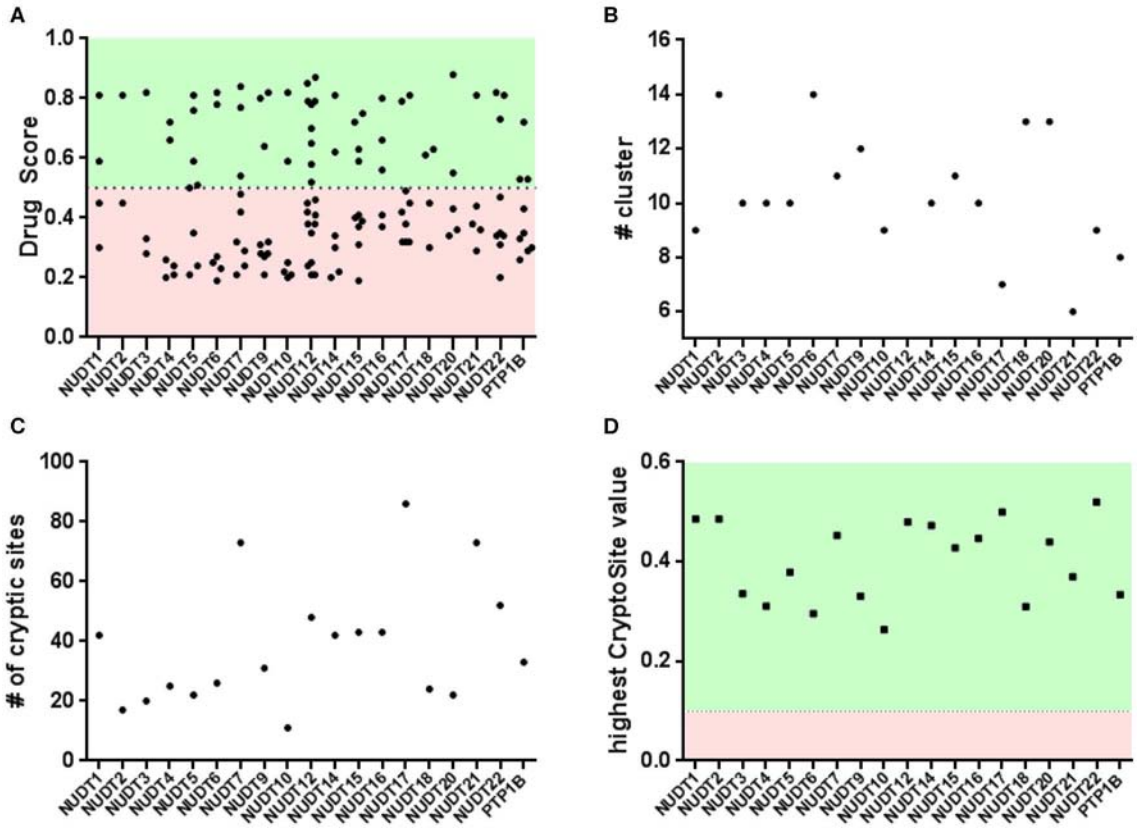


FIGURE 3 | Binding pocket identification and druggability assessment in the automated arm: **(A)** rigid protein assessment with DogSite identifies a number of binding sites within a range of predicted druggability scores; **(B)** FTMap identifies a number of cluster per protein and only NUDT4 and NUDT18 fail to contain enough samples per cluster in the active site; **(C,D)** flexible protein assessment with CryptoSite identifies amino acid residues with different degrees of conformational freedom. Including the top scoring residues these networks largely overlap with pockets identified by DogSite.

forming an allosteric site or conformational changes relevant for substrate binding and protein function. Cryptic scores returned by the algorithm above 0.10 and higher consider a site as cryptic and thus flexible.

When interrogated with CryptoSite, NUDIX hydrolases showed an increased number of cryptic sites around the highest-ranking site as identified before by DoGSite, indicating an extended and flexible three-dimensional network of amino acid residues (**Figure 3** and **Supplementary Material** - CryptoSite). Between 11 and 86 and on average 40 residues scored higher than 0.10 (NUDT1, 42; PTP1B, 33). The highest scoring residues reached values between 0.26 and 0.52 and on average 0.40 (NUDT1, 0.49; PTP1B, 0.33). NUDT10 failed to form a cryptic network while NUDT6, NUDT7 and NUDT22 (**Figure 4**) possessed a second cluster of cryptic sites overlapping with the second highest-ranking sites as identified by DoGSite. Except for NUDT4 (5LTU) and NUDT18 (3GG6), all cryptic networks of the protein family members were populated by more than 16 probes in FTMap (**Figure 3** and **Supplementary Material** – FTMap).

The result of this initial druggability assessment suggest that NUDIX hydrolases are on average good drug targets with regard to their expected or known active sites. Further, only a few members of the family possess a second druggable site as based on DoGSite and FTMap analyses, and even fewer exhibit conformational flexible sites remote from the identified active site.

User Arm – Step 1: SiteMap Binding Site Prediction and Druggability Assessment

In a second parallel approach we assessed druggability using SiteMap and a Glide-based virtual screening workflow applied to a KNIME filtered fragment library (**Figure 2**). SiteMap, an application to identify binding pockets and predict druggability, is implemented in the Schrödinger small-molecule modeling suite. Binding pockets identified on the protein surface are given a score, the Dscore, which is based on pocket parameters such as size, exposure to solvent, enclosure by protein, ratio of hydrogen bond donors and acceptors and importantly hydrophilicity, hydrophobicity and a determined ratio thereof. This druggability score favors proteins with a higher hydrophobic/hydrophilic ratio and thus allows for an early assessment of pocket polarity as required for binding of small-molecule drugs. Typical Dscores for druggable protein pockets are above 1.108 while Dscores below 0.871 suggest a difficult to drug protein (Halgren, 2007, 2009). In addition, comparing individual pocket parameters allows for

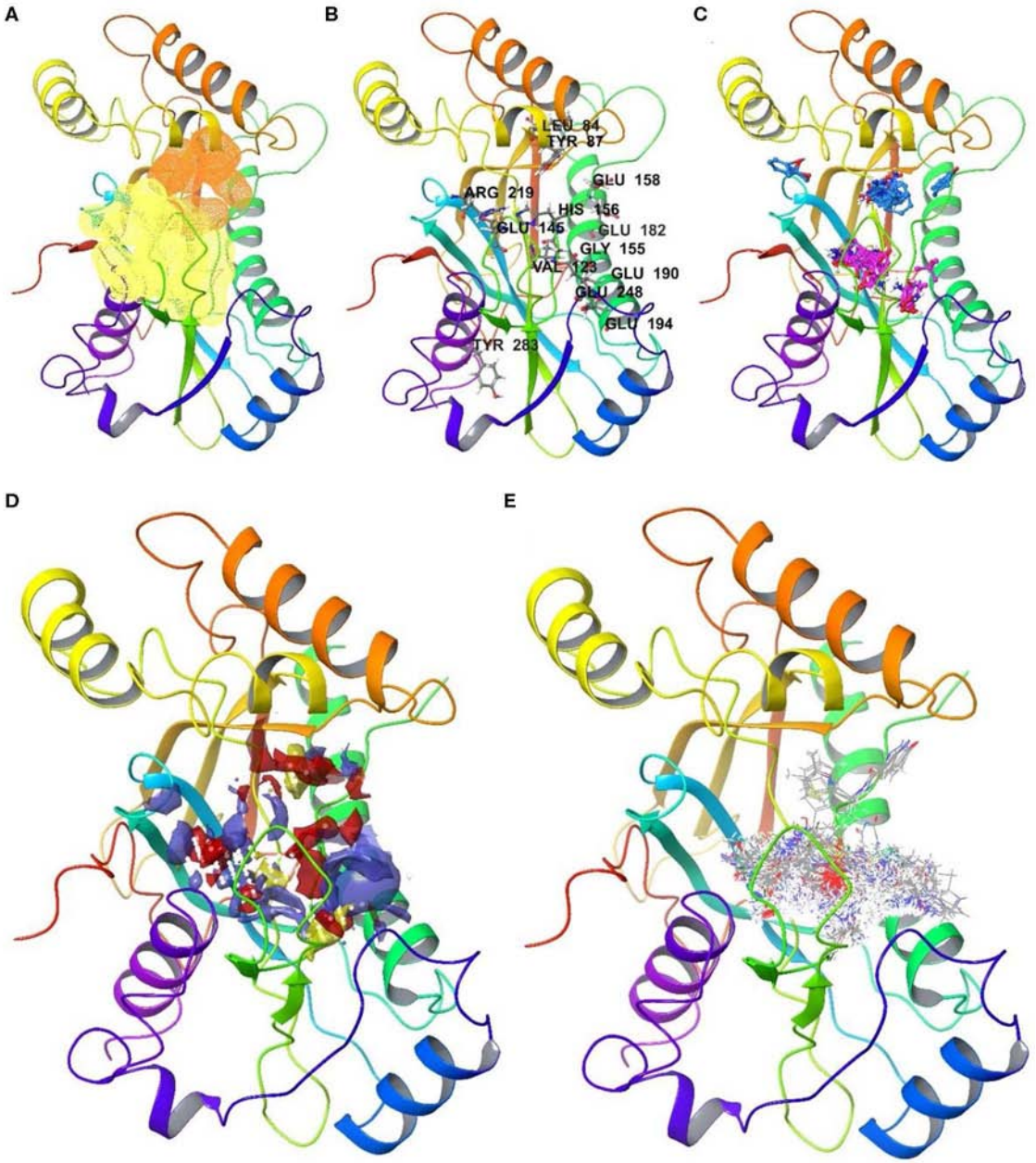


FIGURE 4 | Druggable sites of NUDT22 as identified in the automated and the user arm: **(A)** Rigid assessment with DoGSite identifies a second highly druggable site (orange, DrugScore 0.81) in close proximity to the highest-ranking site (yellow, DrugScore 0.82); **(B)** Flexible assessment with CryptoSite predicts 52 amino acid residues around both pockets as part of an extended 3D network. Highlighted amino acid residues possess cryptic score over 0.2; **(C)** FTMap confirms all cluster hotspots and at least 16 probes within 5 Å radius of either of the binding sites (site 1: pink; site 2: light blue) or high cryptic value sites. **(D)** SiteMap combines the two sites identified by DogSite into a single large binding pocket with an evenly distribution of hydrophobic (yellow) and hydrophilic (red and purple) patches; **(E)** cascade docking of the ZINC fragment library shows a preference for the active site, while the second druggable site is only engaged by members of one chemotype among the top 1000 fragments. The assessment highlights NUDT22 comprising two adjacent druggable sites which in a prospective drug discovery campaign could be targeted separately or in combination.

a detailed picture of druggability and for specific assessment of proteins with similar Dscores and/or sequence.

When SiteMap was applied on the NUDIX hydrolases, the obtained Dscores of the highest-ranking sites were between 0.51 and 1.11 with an average of 0.88 (**Figure 5A** and **Supplementary Material** -SiteMap). Interestingly, except for NUDT4, all identified highest-ranking sites were in overlapping regions or even identical with sites identified with DoGSite

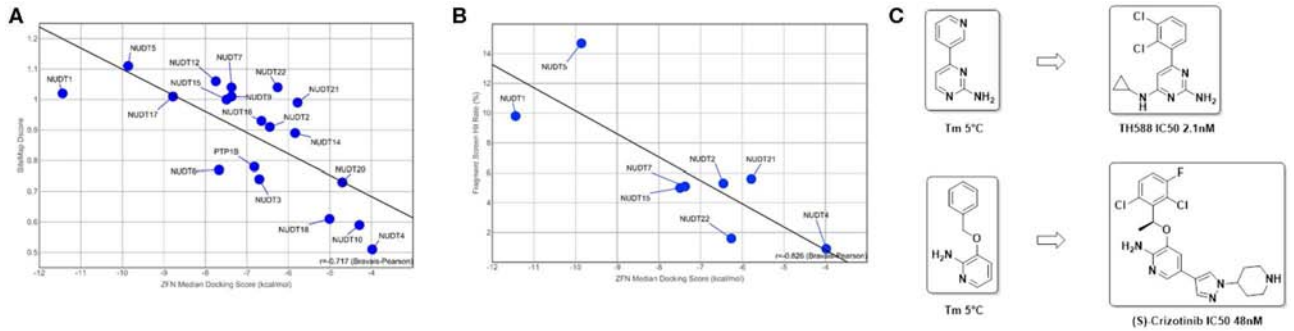


FIGURE 5 | Predictive value of *in silico* assessment and docking for in vitro screening hit rates and suitability of fragment screens for chemical probe generation: **(A)** *in silico* druggability assessment of pockets identified by SiteMap correlates well with the observed ZINC Fragments Now (ZFN) median docking scores of the highest ranking 1000 fragments ($R = -0.717$); **(B)** when translated to *in vitro* to either DSF or X-Ray screens using fragment libraries with overlapping chemical space, a similar correlation can be observed, highlighting the suitability of a purely *in silico* druggability workflow as a standalone method ($R = -0.826$); **(C)** DSF fragment screens reported here yielded strong stabilizing fragment hits that are structural subunits of reported ligands for NUDT1 (Gad et al., 2014; Huber et al., 2014).

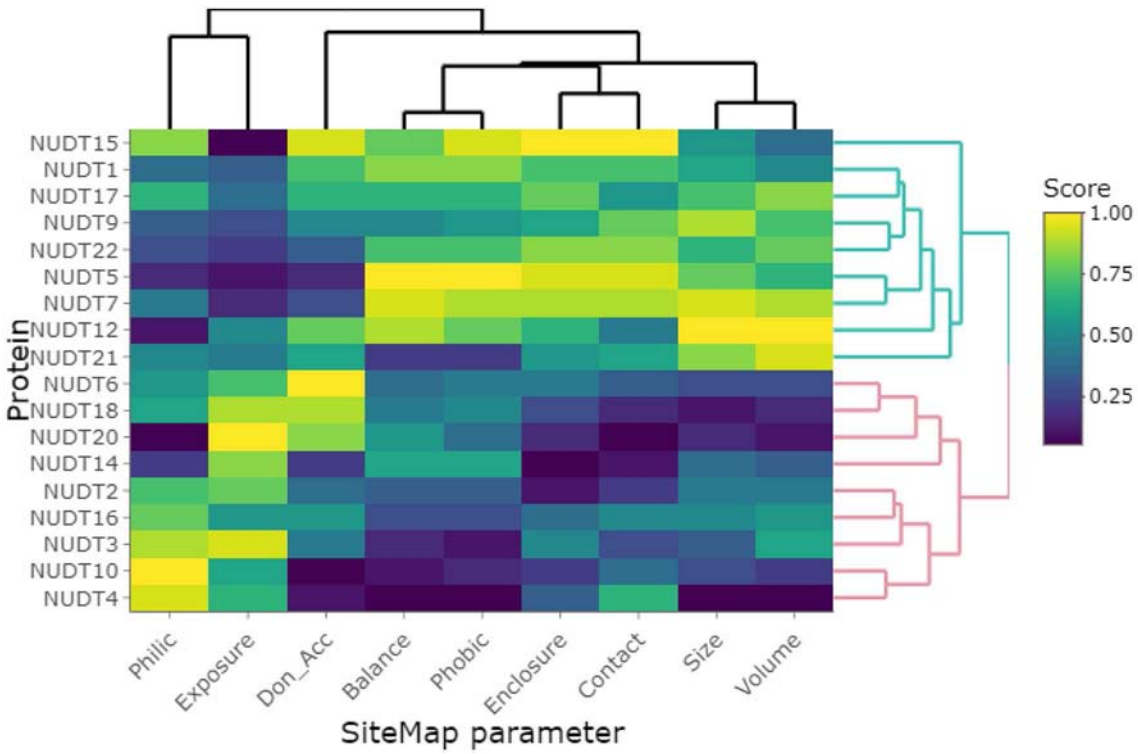


FIGURE 6 | Clustering of investigated NUDIX family members based on the primary SiteMap parameters of the top-ranked sites, resulting in a clear separation of the members deemed druggable (green branches) and those deemed undruggable (red branches).

(**Supplementary Material** - FTMap). Thus, the returned lower Dscore values for NUDT4 (0.51) and NUDT18 (0.61) were consistent between these approaches. In addition, judging by SiteMap, NUDT3 (0.74), NUDT6 (0.77), NUDT10 (0.59), NUDT20 (0.73), and PTP1B (2HNP, 0.78) were classed as difficult drug targets. The highest-ranking members and thus favored drug targets in the family were NUDT1 (1.02), NUDT5 (1.11) NUDT7 (1.04), NUDT9 (1.01), NUDT12 (1.05), NUDT15 (1.00), NUDT17 (1.01), and NUDT22 (1.04, **Figure 3**). Due to the chosen cut-off distance to merge identified pockets (5 Å), SiteMap identified large extended pockets which included several subpockets. Furthermore, as NUDT5, NUDT12, and NUDT15 are functional homodimers, these have two high-ranking pockets. Of these, NUDT12 and NUDT15 contain a third druggable site. NUDT7on the other hand possesses a second high-ranking pocket (Dscore 0.82).

Clustering of the highest-scoring SiteMap pockets using the primary SiteMap parameters shows a clear separation of druggable versus undruggable NUDIX members and allows for comparison of members which are (dis)similar in terms of their active site properties rather than based on sequence (dis)similarities (**Figure 6**). Full-length sequence identity is generally low among the NUDIX family members (see **Supplementary Material**, Percentage Identity), with the exception of NUDT3, NUDT4, NUDT10, and NUDT11. The former 3 being deemed challenging targets and NUDT11 was not evaluated due to lack of structural data. General selectivity issues are thus not anticipated when targeting a specific NUDIX family member. On the other hand, several members have some degree of overlap in their substrate specificity, e.g. NUDT1, 15 and 18 as a subgroup, and NUDT5, NUDT9, NUDT12, and NUDT14 as a second subgroup (Carreras-Puigvert et al., 2017), implying that their active sites share some structural similarity. In this context, the SiteMap parameter profile of NUDT5 is a good reference as it has the highest Dscore of all members. In comparison, NUDT9 has a less favorable balance in hydrophobic and hydrophilic character, while the active site of NUDT12 is somewhat more exposed than NUDT5, but also larger. Although these 3 members have high Dscores, classing them clearly as druggable, they vary in their capacity to accommodate different fragments, as it is reflected by their different median docking scores further down (**Figure 5**). NUDT14 is considered challenging, primarily due to its smaller active site which also is more exposed.

NUDT15 and NUDT18 have been hypothesized to be able to act as back-up enzymes for NUDT1 due to their overlapping substrate specificities. Comparison of their SiteMap parameter profiles shows clear differences despite NUDT1 and NUDT15 being among the NUDIX members with highest Dscores. The active site of NUDT15 is somewhat smaller and more enclosed than for NUDT1 due to an inward movement of a helix (Carter et al., 2015). NUDT18 is considered to be challenging due to its small and more exposed active site. This also results in a poor fragment scoring profile (please see below). Collectively, these differences in site parameters allow for the development of highly selective chemical probes, as witnessed for NUDT1 and NUDT5 (Gad et al., 2014; Page et al., 2018).

User Arm – Step 2: *in silico* Docking of ZINC Library

As a final druggability assessment and potential to identify starting points amenable for a fragment growing-based drug discovery campaign out of commercial fragment space, we performed *in silico* docking campaigns of the ZINC Frag now database (Irwin et al., 2012) against the structures under consideration. The comprehensive fragment library was filtered against unwanted structural motifs and prepared for docking using a KNIME workflow (Berthold et al., 2008). For a detailed description, please refer to the Methods part of this manuscript. Ultimately, 205,891 fragments remained after filtering the original set of 704,041 ZINC fragments. Using the Schrödinger suite, ligand preparation and grid generation for the highest-ranking pocket as identified by SiteMap were performed to enable virtual screening of this subset applying three stages of accuracy. In each stage, the top-ranked 10% of compounds were retained and passed on to the next stage. Finally, the top-ranked 1,000 fragments were used to calculate a median docking score enabling assessment of druggability based on commercially available fragment space. The returned median docking scores, where lower is better, ranged from−4.0 to −11.4 kcal/mol with an average of−6.8 kcal/mol. NUDT1 and NUDT5, both validated drug targets in the literature, scored−11.4 and −9.9 kcal/mol respectively. In addition, and judged by the median docking score, NUDT17 (−8.8 kcal/mol) is a third promising drug target. PTP1B (−6.8 kcal/mol) scores average among the NUDIX family members, while the scores for NUDT4 (−4.0 kcal/mol), NUDT10 (−4.3 kcal/mol), NUDT20 (−4.7 kcal/mol) and NUDT18 (−5.0 kcal/mol) indicate a potentially challenging drug discovery campaign (**Figure 5A**). When using the median docking scores and plotted against their respective SiteMap Dscores, a good inverse correlation ($R = -0.717$, Bravais-Pearson) can be observed (**Figure 5**), suggesting an *in silico*-based prioritization scheme of drug discovery campaigns against NUDIX proteins. Thus, fragment docking against the top-ranked SiteMap pockets recapitulates their druggability potential but additionally provides potential starting points readily accessible for fragment-based drug discovery campaigns.

The hydrophobicity of small-molecule drugs is a property which needs to be delicately balanced since it affects multiple parameters including solubility, permeability, plasma protein binding and metabolism. Druggable binding pockets of target proteins therefore require a certain hydrophobic-hydrophilic balance to accommodate ligands with drug-like properties. When applying a balance of at least 0.5 the SiteMap assessment prefers NUDT1 (3Q93, 0.69), NUDT5 (6GRU, 1.34), NUDT7 (5T3P, 0.72), NUDT15 (5BON, 0.62), NUDT17 (5LF8, 0.50), and NUDT22 (5LF9, 0.58) and disfavors NUDT3 (2FVV, 0.01), NUDT4 (5LTU, 0.00), NUDT10 (3MCF, 0.01), and PTP1B (2HNP, 0.05). With regard to their returned median fragment docking scores, pocket polarity might correlate with either higher or lower scores (**Supplementary Material** – ZINC fragment docking and SiteMap). A possible explanation is, that the library was filtered to fit a drug-like profile and thus preselects for druggable proteins itself, ignoring their respective pocket properties. Importantly, none of the crystal structures used here were bound to high-affinity lead compounds originating from drug discovery campaigns and hence no hydrophobic subpockets induced by such compounds where probed in this study.

When combined, the top-1,000 ranked fragments obtained for the 18 protein targets comprised 13,203 unique fragments, indicating a certain amount of "promiscuity," i.e., fragments binding to 2 or more proteins (36% of fragments). In fact, 73 fragments bound to 6 or more targets (see **Supplementary Material** - Fragment promiscuity), with one fragment hitting 11 out of 18 proteins. It should be noted that the average docking scores were rather poor, ranging from −7.56 to −5.53 kcal/mol. Of interest is the notion that the proteins

TABLE 2 | Summary of NUDIX protein family members in screens against fragment and in biochemical malachite green assays.

Protein	Screening technique	Size screening set	Hit rate
NUDT7	Fragment covalent (Resnick et al., 2019)	993	36 (3.6%)
NUDT4	Fragment diamond Xray[1]	768	7 (0.9%)
NUDT5	Fragment diamond Xray[1]	768	113 (14.7%)
NUDT7	Fragment diamond Xray[1]	768	39 (5.1%)
NUDT21	Fragment diamond Xray[1]	768	43 (5.6%)
NUDT22	Fragment diamond Xray[1]	768	12 (1.6%)
NUDT1	Fragment DSF	450	44 (9.8%)
NUDT2	Fragment DSF	650	35 (5.3%)
NUDT5	Fragment DSF	450	0*
NUDT15	Fragment DSF	1,200	60 (5.0%)
NUDT1	Biochem screen malachite green	5,336	429 (8.0%)
NUDT2	Biochem screen malachite green	5,336 [11,992]	261 (4.9%) [235 (2.0%)**]
NUDT5	Biochem screen malachite green	5,336 [72,004]	3 (0.1%) [527 (0.7%)**]
NUDT15	Biochem screen malachite green	5,336 [17,908]	10 (0.2%) [98 (0.5%)**]
NUDT16	Biochem screen malachite green	5,336	7 (0.1%)

*Please note the biochemical screen library grew over time. For comparison, results beyond the basic set of 5,336 compounds are presented in brackets. *Melting temperature unsuitable for thermal shift screens; **Definition of hit more stringent.*

deemed undruggable by SiteMap appeared to be enriched for promiscuous fragment hits (except NUDT10 and NUDT20), as opposed to druggable proteins (except NUDT15 and NUDT22). A certain degree of promiscuity should be expected when docking 200K fragments to multiple targets, as this is in line with the basic concept of fragment-based drug discovery, i.e., the ability of low-complexity fragments to interact with a multitude of (sub)pockets across a wide range of proteins.

Correlation With Experimental Fragment-Based and Biochemical Screening Data

A number of fragment screens against NUDIX proteins have been performed by others and us[1]. For a list of applied screening techniques, library sets and hit rates, please see **Table 2**. When the hit rates of the fragment screens were compared with the *in silico*-derived median ZFN docking scores a good correlation was observed (Bravais-Pearson 0.826; **Figure 5B**). This underscores the applicability of *in silico* docking for rapid protein druggability assessment. In agreement with most computational assessments, NUDT1 and NUDT5 yield high hit rates of 9.8% and 14.7%, respectively, while the experimental hit rate of 0.9% for NUDT4 confirms its challenging character predicted by computational assessment. Other NUDIX proteins are in the range of common hit rates for fragment screens and between 1.6 and 5.6% (Aretz et al., 2014). This observation holds true for different sets screened by different groups (**Figures 5B**, **7**). Interestingly, the DSF screen against NUDT1 found two structures with a strong thermal stabilization of 5°C. These structures are fragments of the reported NUDT1 inhibitors TH588 (IC50: 2.1 nM) and Crizotinib (IC50: 48 nM) and thus underscore the suitability of DSF to find starting points for lead generation (**Figure 5C**). However, DSF is not feasible for proteins with high native melting points (e.g., NUDT5, 76°C), and here *in silico* fragment screening against druggable sites may be particularly advantageous.

Several biochemical screening campaigns against NUDIX proteins have also been performed in our laboratories. While compound libraries have varied somewhat between targets, reflecting development of the compound libraries over time, there is a small core set of about 5,300 chemically diverse compounds that have been tested for all proteins. It is noteworthy that these screens were performed based on a common screening platform employing a coupled enzymatic assay with a malachite green readout. This cost-effective assay has been frequently employed in our lab, including screens on other nucleotide-processing targets such as dCTPase (Llona-Minguez et al., 2016), ITPase and dUTPase, and with robust performance in compound sets beyond 100,000 compounds (all unpublished). A key reason for this is the low rates of interference with the coupled enzymes and the absorbance readout at 630 nm, as evidenced by a low appearance of common hits. Also, the presence of PAINS and aggregators within hit lists is generally low for this family of proteins (**Supplementary Material** – Screens using malachite green), demonstrating robust screening performance of the recombinantly produced proteins and other assay components. The biochemical screen outcomes are summarized in detail in **Table 2** and in the **Supplementary Material** – Screens using malachite green. In line with assessments of chemical amenability and learnings in the fragment-based screens, the majority of targets generated hits that confirmed activity in follow-up studies, with NUDT1 demonstrating an extreme hit rate in this sub-set. This significant amenability is in line with the publication of hits from multiple groups. A critical outlier in this set was NUDT5, which demonstrated hit rates as low as notoriously challenging targets dUTPase and ITPase, while predictions and fragment-based screening showed the opposite (**Supplementary Material** – Fragment screening hit rates). Already at the time of screening

we had a reason to revisit the screening data for NUDT5 and follow-up studies demonstrated competition between active site hits and a structurally important Mg^{2+} (Costa and Dieckmann, 2011; Vardakou et al., 2014). After correction of the assay buffer, by lowering the $MgCl_2$ concentration 10-fold, we observed significantly higher hit rates in the subsequently applied compound sets and identified compounds that could be further optimized to nM potencies (Page et al., 2018). As a general observation and for the five NUDIX protein members screened, no correlation between ZINC fragment docking scores and observed hit rates from biochemical screening can be observed (**Supplementary Material** – Screens using malachite green). In contrast to the covered fragment library chemical space, a rule-of-five compliant library of few thousand compounds may complement the search for a chemical starting point but may be limited in coverage of chemical space itself. However, in the past we have shown, that embarking on drug discovery campaigns from observed hits in both fragment and biochemical screen lead to successful generation of chemical probes for a number of NUDIX protein family members and other pyrophosphatases (Gad et al., 2014; Llona-Minguez et al., 2016; Page et al., 2018)[2].

CONCLUSION AND SUMMARY

Here we presented a dual *in silico* druggability assessment workflow suitable for large-scale evaluation of proteins and protein families, applied to the NUDIX family. Initially, we introduced a hands-on workflow solely based on the protein crystal structure using the open access server of DogSite (Volkamer et al., 2012) and FTMap (Kozakov et al., 2015) for rigid and CryptoSite (Cimermancic et al., 2016) and FTMap for dynamic assessment. Importantly, before using these servers, thorough manual protein structure verification is necessary to exclude artifacts due to crystal packing, construct used and resolution limits. On the one hand, DogSite returns both identity and score of druggable sites, while FTMap docks small organic solvent molecules. Especially in cases of sparsely evaluated proteins or protein complexes this dual assessment may be beneficial for structural assessment and potential chemical probe generation. On the other hand, CryptoSite identifies conformationally active amino acid residues. In the past, the returned cryptic scores have been correlated with FTMap solvent docking and eased decision on whether or where potential allosteric sites may be situated (Vajda et al., 2018). Timewise, this quick computational assessment may be achieved within days for singular proteins and weeks for small protein families. Depending on local load and choice of sever location the return time is usually minutes for DogSite, hours for FTMap, and 1 day for CryptoSite. The detailed assessment and correlation of data from the different algorithms allows for the rationalization of targeting strategies. In case of the NUDIX proteins, NUDT22 for example showed to have high scores in DogSite and FTMap with CryptoSite confirming flexibility around two closely related sites. Further, in the past we have shown that comparing different crystals structures of the same protein can allow for the observation of targetable conformations more suitable small-molecule development (Michel et al., 2019). With the open access deposition of all screening data by the SGC, a similar albeit more time consuming approach is possible for a number of NUDIX proteins[1].

In a second, user-guided arm we assessed protein druggability employing several implemented functions in Schrödinger's commercial small-molecule modeling suite combined with freely available KNIME (Berthold et al., 2008). First, proteins were interrogated for potential binding pockets and the corresponding DScores using SiteMap. The highest ranking pockets were then

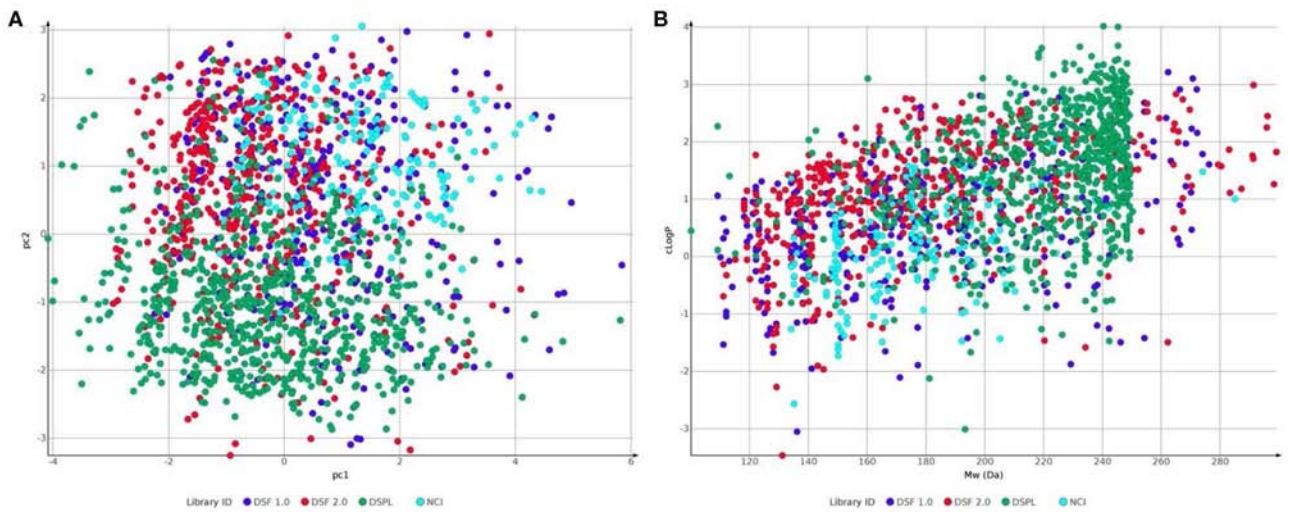


FIGURE 7 | Comparison of fragment libraries used in screens against NUDIX protein family members: Laboratory for Chemical Biology at Karolinska Institutet (LCBKI) DSF fragment library of the first generation (DSF 1.0, blue), additional members of a second generation DSF fragment library (DSF 2.0, red), Nucleobase analogs (NCI, light blue) and Diamond-SGC Poised Library (DSPL, green) with respect to: **(A)** physicochemical property coverage and diversity, expressed as the first two principal components (pc1 and pc2) obtained from a principal component analysis (PCA) on six Lipinski-type properties; **(B)** clogP against Molecular weight.

used to perform cascade docking with a filtered ZINC fragment library (Irwin et al., 2012). Subsequently, the median docking score of the top-1,000 ranked fragments was used as a chemical space-based druggability assessment. Both parameters, DScore and median docking score form the basis of this second *in silico* druggability assessment and require few days per protein, depending on the user set up and the size of the used *in silico* library. The observed docking scores correlate well with predicted DScores (**Figure 5A**) and additionally provide commercially accessible chemical starting points for the development of chemical probes. At last, when compared with experimental fragment screens based on X-ray crystallography, a covalent set and thermal stabilization in DSF, a similar correlation was observed between hit rates and median docking scores (**Figure 5B**), even when using chemically distinct screening sets (**Figure 7**). This supports the applicability of an *in silico* druggability workflow as a standalone method for protein assessment and speaks for the chemical space coverage of fragment libraries generated at CBCS and Diamond/SGC (Michel et al., 2019)[7].

In summary, we report here a fully *in silico* druggability assessment of the NUDIX protein family, that serves as a standalone method and a workflow to identify the most suitable members for a drug discovery campaign. We show that the dual assessment correlates well with experimental results and further allows for the *in silico* identification of secondary druggable sites, alternative targeting strategies and structural basis for fragment growing campaigns. Importantly, the workflow allows for rapid assessment of any protein with reported structures in the protein data bank and as such should be broadly applicable in early drug discovery campaigns.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

[7]*DSI-Poised Library - MX - Diamond Light Source*. Available online at: https://www.diamond.ac.uk/Instruments/Mx/Fragment-Screening/Fragment-Libraries/DSI-poised-library.html (accessed October 7, 2019).

AUTHOR CONTRIBUTIONS

MM and EH compiled the list of PDB structures and performed protein preparation and validation. MM controlled end stage correlation and assessment and performed tasks in the automated arm. EH performed tasks in the User arm. EW, KP, IA, A-LG, and TL performed assays, fragment, and biochemical high-throughput screens. MM and EH rationalized and designed the project. UW and TH supervised the project. MM, EH, AG, and TL wrote the manuscript. MM coordinated collaborations. All authors commented on the manuscript.

FUNDING

Funding was obtained from the Swedish Cancer Society (TH), the Swedish Children's Cancer Foundation (TH), ERC Tarox-695376 (TH) and Swedish Pain Relief Foundation (TH). Chemical Biology Consortium Sweden received funding from the Swedish Research Council, Karolinska Institutet and SciLifeLab during the course of this work.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the national infrastructure of CBCS and LCBKI for supplying the DSF fragment and HTS library and help during development and performance of the HTS assay for different NUDIX protein family members. The Protein Science Facility of Karolinska Institutet was acknowledged for protein production. Further, the operators of CryptoSite at the University of California, San Francisco, DogSite at the Zentrum für Bioinformatik, University of Hamburg and FTMap at Boston University are acknowledged for their computing set up and the open access to program code. Dana Michel was acknowledged for her critical discussion and proofreading of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2020.00443/full#supplementary-material

REFERENCES

Abdelraheim, S. R., Spiller, D. G., and McLennan, A. G. (2003). Mammalian NADH diphosphatases of the nudix family: cloning and characterization of the human peroxisomal NUDT12 protein. *Biochem. J.* 374(Pt. 2), 329–335. doi: 10.1042/bj20030441

An, S., and Fu, L. (2018). Small-molecule PROTACs: an emerging and promising approach for the development of targeted therapy drugs. *EBioMedicine* 36, 553–562. doi: 10.1016/j.ebiom.2018.09.005

Aretz, J., Anumala, U. R., Fuchsberger, F. F., Molavi, N., Ziebart, N., Zhang, H., et al. (2018). Allosteric inhibition of a mammalian lectin. *J. Am. Chem. Soc.* 140, 14915–14925. doi: 10.1021/jacs.8b08644

Aretz, J., Wamhoff, E.-C., Hanske, J., Heymann, D., and Rademacher, C. (2014). Computational and experimental prediction of human C-type lectin receptor druggability. *Front. Immunol.* 5:323. doi: 10.3389/fimmu.2014.00323

Baell, J. B., and Holloway, G. A. (2010). New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 53, 2719–2740. doi: 10.1021/jm901137j

Barford, D., Flint, A. J., and Tonks, N. K. (1994). Crystal structure of human protein tyrosine phosphatase 1B. *Science* 263, 1397–1404. doi: 10.1126/science.8128219

Baykov, A. A., Evtushenko, O. A., and Avaeva, S. M. (1988). A malachite green procedure for orthophosphate determination and its use in alkaline phosphatase-based enzyme immunoassay. *Anal. Biochem.* 171, 266–270. doi: 10.1016/0003-2697(88)90484-8

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., et al. (2008). "KNIME: The Konstanz Information Miner," in *Data Analysis, Machine Learning and Applications; Studies in Classification, Data Analysis, and Knowledge Organization*, eds C. Preisach, H. Burkhardt, L.

Schmidt-Thieme, and R. Decker (Berlin; Heidelberg: Springer). p. 319–326. doi: 10.1007/978-3-540-78246-9_38

Caffrey, J. J., Hidaka, K., Matsuda, M., Hirata, M., and Shears, S. B. (1999). The human and rat forms of multiple inositol polyphosphate phosphatase: functional homology with a histidine acid phosphatase up-regulated during endochondral ossification. *FEBS Lett.* 442, 99–104. doi: 10.1016/S0014-5793(98)01636-6

Caffrey, J. J., Safrany, S. T., Yang, X., and Shears, S. B. (2000). Discovery of molecular and catalytic diversity among human diphosphoinositol-polyphosphate phosphohydrolases. an expanding nudt family. *J. Biol. Chem.* 275, 12730–12736. doi: 10.1074/jbc.275.17.12730

Carreras-Puigvert, J., Zitnik, M., Jemth, A.-S., Carter, M., Unterlass, J. E., Hallström, B., et al. (2017). A comprehensive structural, biochemical and biological profiling of the human NUDIX hydrolase family. *Nat. Commun.* 8:1541. doi: 10.1038/s41467-017-01642-w

Carter, M., Jemth, A.-S., Hagenkort, A., Page, B. D. G., Gustafsson, R., Griese, J. J., et al. (2015). Crystal structure, biochemical and cellular activities demonstrate separate functions of MTH1 and MTH2. *Nat. Commun.* 6:7871. doi: 10.1038/ncomms8871

Choi, S.-Y., Jang, J. H., and Kim, K. R. (2011). Analysis of Differentially Expressed Genes in Human rectal carcinoma using suppression subtractive hybridization. *Clin. Exp. Med.* 11, 219–226. doi: 10.1007/s10238-010-0130-5

Cimermancic, P., Weinkam, P., Rettenmaier, T. J., Bichmann, L., Keedy, D. A., Woldeyes, R. A., et al. (2016). CryptoSite: expanding the druggable proteome by characterization and prediction of cryptic binding sites. *J. Mol. Biol.* 428, 709–719. doi: 10.1016/j.jmb.2016.01.029

Coseno, M., Martin, G., Berger, C., Gilmartin, G., Keller, W., and Doublié, S. (2008). Crystal structure of the 25 KDa subunit of human cleavage factor I_m. *Nucleic Acids Res.* 36, 3474–3483. doi: 10.1093/nar/gkn079

Costa, J. B. D., and Dieckmann, T. (2011). Entropy and Mg2+ control ligand affinity and specificity in the malachite green binding RNA aptamer. *Mol. BioSyst.* 7, 2156–2163. doi: 10.1039/c1mb05075c

Dubianok, Y., Collins, P., Krojer, T., Fairhead, M., MacLean, E., Diaz, S., et al. (2018). *LPDB 6gru Structure Summary ‹ Protein Data Bank in Europe (PDBe) ‹ EMBL-EBI*. Available online at: https://www.ebi.ac.uk/pdbe/entry/pdb/6gru (accessed June 3, 2019).

Ellermann, M., Eheim, A., Rahm, F., Viklund, J., Guenther, J., Andersson, M., et al. (2017). Novel class of potent and cellularly active inhibitors devalidates MTH1 as broad-spectrum cancer target. *ACS Chem. Biol.* 12, 1986–1992. doi: 10.1021/acschembio.7b00370

Farand, J., Kropf, J. E., Blomgren, P., Xu, J., Schmitt, A. C., Newby, Z. E., et al. (2020). Discovery of potent and selective MTH1 inhibitors for oncology: enabling rapid target (In)validation. *ACS Med. Chem. Lett.* 11, 358–364. doi: 10.1021/acsmedchemlett.9b00420

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.* 47, 1739–1749. doi: 10.1021/jm0306430

Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., et al. (2006). Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. *J. Med. Chem.* 49, 6177–6196. doi: 10.1021/jm051256o

Gad, H., Koolmeister, T., Jemth, A.-S., Eshtad, S., Jacques, S. A., Ström, C. E., et al. (2014). MTH1 inhibition eradicates cancer by preventing sanitation of the DNTP pool. *Nature* 508, 215–221. doi: 10.1038/nature13181

Ge, H., Chen, X., Yang, W., Niu, L., and Teng, M. (2013). Crystal structure of wild-type and mutant human Ap4A hydrolase. *Biochem. Biophys. Res. Commun.* 432, 16–21. doi: 10.1016/j.bbrc.2013.01.095

Greenwood, J. R., Calkins, D., Sullivan, A. P., and Shelley, J. C. (2010). Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J. Comput. Aided Mol. Des.* 24, 591–604. doi: 10.1007/s10822-010-9349-1

Halgren, T. (2007). New method for fast and accurate binding-site identification and analysis. *Chem Biol Drug Des.* 69, 146–148. doi: 10.1111/j.1747-0285.2007.00483.x

Halgren, T. A. (2009). Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model* 49, 377–389. doi: 10.1021/ci800324m

Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *J. Med. Chem.* 47, 1750–1759. doi: 10.1021/jm030644s

Huber, K. V. M., Salah, E., Radic, B., Gridling, M., Elkins, J. M., Stukalov, A., et al. (2014). Stereospecific targeting of MTH1 by (S)-crizotinib as an anticancer strategy. *Nature* 508, 222–227. doi: 10.1038/nature13194

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. (2012). ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model* 52, 1757–1768. doi: 10.1021/ci3001277

Kettle, J. G., Alwan, H., Bista, M., Breed, J., Davies, N. L., Eckersley, K., et al. (2016). Potent and selective inhibitors of MTH1 probe its role in cancer cell survival. *J. Med. Chem.* 59, 2346–2361. doi: 10.1021/acs.jmedchem.5b01760

Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., et al. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* 10, 733–755. doi: 10.1038/nprot.2015.043

Krishnan, N., Konidaris, K. F., Gasser, G., and Tonks, N. K. (2018). A potent, selective, and orally bioavailable inhibitor of the protein-tyrosine phosphatase PTP1B improves insulin and leptin signaling in animal models. *J. Biol. Chem.* 293, 1517–1525. doi: 10.1074/jbc.C117.819110

Llona-Minguez, S., Höglund, A., Jacques, S. A., Johansson, L., Calderón-Montaño, J. M., Claesson, M., et al. (2016). Discovery of the first potent and selective inhibitors of human DCTP pyrophosphatase 1. *J. Med. Chem.* 59, 1140–1148. doi: 10.1021/acs.jmedchem.5b01741

Mathea, S., Salah, E., Velupillai, S., Tallant, C., Pike, A. C. W., Bushell, S. R., et al. (2017b). *PDB 5mp0 Structure Summary ‹ Protein Data Bank in Europe (PDBe) ‹ EMBL-EBI.* Available online at: https://www.ebi.ac.uk/pdbe/entry/pdb/5mp0 (accessed June 3, 2019).

Mathea, S., Tallant, C., Salah, E., Wang, D., Velupillai, S., Nowak, R., et al. (2017a). *PDB 5lf8 Structure Summary ‹ Protein Data Bank in Europe (PDBe) ‹ EMBL-EBI*. Available online at: https://www.ebi.ac.uk/pdbe/entry/pdb/5lf8 (accessed June 3, 2019).

Michel, M., Visnes, T., Homan, E. J., Seashore-Ludlow, B., Hedenström, M., Wiita, E., et al. (2019). Computational and experimental druggability assessment of human DNA glycosylases. *ACS Omega* 4, 11642–11656. doi: 10.1021/acsomega.9b00162

Mullard, A. (2018). Phosphatases start shedding their stigma of undruggability. *Nat. Rev. Drug Discov.* 17, 847–849. doi: 10.1038/nrd.2018.201

Niesen, F. H., Berglund, H., and Vedadi, M. (2007). The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat. Protoc.* 2, 2212–2221. doi: 10.1038/nprot.2007.321

Page, B. D. G., Valerie, N. C. K., Wright, R. H. G., Wallner, O., Isaksson, R., Carter, M., et al. (2018). Targeted NUDT5 inhibitors block hormone signaling in breast cancer cells. *Nat. Commun.* 9:250. doi: 10.1038/s41467-017-02293-7

Petrocchi, A., Leo, E., Reyna, N. J., Hamilton, M. M., Shi, X., Parker, C. A., et al. (2016). Identification of potent and selective MTH1 inhibitors. *Bioorg. Med. Chem. Lett.* 26, 1503–1507. doi: 10.1016/j.bmcl.2016.02.026

Resnick, E., Bradley, A., Gan, J., Douangamath, A., Krojer, T., Sethi, R., et al. (2019). Rapid covalent-probe discovery by electrophile-fragment screening. *J. Am. Chem. Soc.* 141:8951–8968. doi: 10.1021/jacs.9b02822

Samaranayake, G. J., Huynh, M., and Rai, P. (2017). MTH1 as a chemotherapeutic target: the elephant in the room. *Cancers* 9:47. doi: 10.3390/cancers9050047

Shelley, J. C., Cholleti, A., Frye, L. L., Greenwood, J. R., Timlin, M. R., and Uchimaya, M. (2007). Epik: a software program for PK prediction and protonation state generation for drug-like molecules. *J. Comput. Aided. Mol. Des.* 21, 681–691. doi: 10.1007/s10822-007-9133-z

Shen, B. W., Perraud, A.-L., Scharenberg, A., and Stoddard, B. L. (2003). The crystal structure and mutational analysis of human NUDT9. *J. Mol. Biol.* 332, 385–398. doi: 10.1016/S0022-2836(03)00954-9

Srikannathasan, V., Nunez, C. A., Tallant, C., Siejka, P., Mathea, S., Kopec, J., et al. (2017b). *PDB 5t3p Structure Summary ‹ Protein Data Bank in Europe (PDBe) ‹ EMBL-EBI*. Available online at: https://www.ebi.ac.uk/pdbe/entry/pdb/5t3p (accessed June 3, 2019).

Srikannathasan, V., Nunez, C. A., Tallant, C., Siejka, P., Mathea, S., Newman, J., et al. (2017a). *PDB 5ltu Structure Summary ‹ Protein Data Bank in Europe (PDBe) ‹ EMBL-EBI.* Available online at: https://www.ebi.ac.uk/pdbe/entry/pdb/5ltu (accessed June 3, 2019).

Tallant, C., Siejka, P., Mathea, S., Shrestha, L., Krojer, T., Srikannathasan, V., et al. (2017). *PDB 5lf9 structure summary ‹ Protein Data Bank in Europe (PDBe) ‹ EMBL-EBI*. Available online at: https://www.ebi.ac.uk/pdbe/entry/pdb/5lf9 (accessed June 3, 2019).

Thorsell, A.-G., Persson, C., Gräslund, S., Hammarström, M., Busam, R. D., and Hallberg, B. M. (2009). Crystal structure of human diphosphoinositol phosphatase 1. *Proteins* 77, 242–246. doi: 10.1002/prot.22489

Trésaugues, L., Lundbäck, T., Welin, M., Flodin, S., Nyman, T., Silvander, C., et al. (2015). Structural basis for the specificity of human nudt16 and its regulation by inosine monophosphate. *PLoS ONE* 10:e0131507. doi: 10.1371/journal.pone.0131507

Tresaugues, L., Moche, M., Arrowsmith, C. H., Berglund, H., Bountra, C., Collinsm, R., et al. (2009a). *PDB 3h95 Structure Summary ‹ Protein Data Bank in Europe (PDBe) ‹ EMBL-EBI*. Available online at: https://www.ebi.ac.uk/pdbe/entry/pdb/3h95 (accessed June 3, 2019).

Tresaugues, L., Moche, M., Arrowsmith, C. H., Berglund, H., Busam, R. D., and Collins, R. (2008). *PDB 3cou Structure Summary ‹ Protein Data Bank in Europe (PDBe) ‹ EMBL-EBI*. Available online at: https://www.ebi.ac.uk/pdbe/entry/pdb/3cou (accessed June 3, 2019).

Tresaugues, L., Siponen, M. I., Arrowsmith, C. H., Berglund, H., Bountra, C., Collins, R., et al. (2011a). *PDB 3q93 structure summary ‹ Protein Data Bank in Europe (PDBe) ‹ EMBL-EBI*. Available online at: https://www.ebi.ac.uk/pdbe/entry/pdb/3q93 (accessed June 3, 2019).

Tresaugues, L., Siponen, M. I., Arrowsmith, C. H., Berglund, H., Bountra, C., Collins, R., et al. (2011b). *PDB 3q91 Structure Summary ‹ Protein Data Bank in Europe (PDBe) ‹ EMBL-EBI*. Available online at: https://www.ebi.ac.uk/pdbe/entry/pdb/3q91 (accessed June 3, 2019).

Tresaugues, L., Siponen, M. I., Lehtio, L., Arrowsmith, C. H., Berglund, H., Bountra, C., et al. (2009b). *PDB 3gg6 Structure Summary ‹ Protein Data Bank in Europe (PDBe) ‹ EMBL-EBI*. Available online at: https://www.ebi.ac.uk/pdbe/entry/pdb/3gg6 (accessed June 3, 2019).

Tresaugues, L., Welin, M., Arrowsmith, C. H., Berglund, H., Bountra, C., Collins, R., et al. (2010). *PDB 3mcf structure summary ‹ Protein Data Bank in Europe (PDBe) ‹ EMBL-EBI*. Available online at: https://www.ebi.ac.uk/pdbe/entry/pdb/3mcf (accessed June 3, 2019).

Tsuzuki, T., Egashira, A., and Kura, S. (2001). Analysis of MTH1 gene function in mice with targeted mutagenesis. *Mutat. Res.* 477, 71–78. doi: 10.1016/S0027-5107(01)00108-7

Vajda, S., Beglov, D., Wakefield, A. E., Egbert, M., and Whitty, A. (2018). Cryptic binding sites on proteins: definition, detection, and druggability. *Curr. Opin. Chem. Biol.* 44, 1–8. doi: 10.1016/j.cbpa.2018.05.003

Vardakou, M., Salmon, M., Faraldos, J. A., and O'Maille, P. E. (2014). Comparative analysis and validation of the malachite green assay for the high throughput biochemical characterization of terpene synthases. *MethodsX* 1, 187–196. doi: 10.1016/j.mex.2014.08.007

Volkamer, A., Kuhn, D., Rippmann, F., and Rarey, M. (2012). DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics* 28, 2074–2075. doi: 10.1093/bioinformatics/bts310

Walters, W. P., and Murcko, M. A. (2002). Prediction of "drug-likeness". *Adv. Drug Deliv. Rev.* 54, 255–271. doi: 10.1016/S0169-409X(02)00003-0

Warpman Berglund, U., Sanjiv, K., Gad, H., Kalderén, C., Koolmeister, T., Pham, T., et al. (2016). Validation and development of MTH1 inhibitors for treatment of cancer. *Ann. Oncol.* 27, 2275–2283. doi: 10.1093/annonc/mdw429

Wenthur, C. J., Gentry, P. R., Mathews, T. P., and Lindsley, C. W. (2014). Drugs for allosteric sites on receptors. *Annu Rev Pharmacol Toxicol* 54, 165–184. doi: 10.1146/annurev-pharmtox-010611-134525

Wright, R. H. G., Lioutas, A., Le Dily, F., Soronellas, D., Pohl, A., Bonet, J., et al. (2016). ADP-ribose-derived nuclear ATP synthesis by NUDIX5 is required for chromatin remodeling. *Science* 352, 1221–1225. doi: 10.1126/science.aad9335

Wu, H., Li, L., Chen, K.-M., Homolka, D., Gos, P., Fleury-Olela, F., et al. (2019). Decapping enzyme NUDT12 partners with BLMH for cytoplasmic surveillance of NAD-capped RNAs. *Cell Rep.* 29, 4422–4434.e13. doi: 10.1016/j.celrep.2019.11.108

Yueh, C., Rettenmaier, T. J., Xia, B., Hall, D. R., Alekseenko, A., Porter, K. A., et al. (2019). Kinase atlas: druggability analysis of potential allosteric sites in kinases. *J. Med. Chem.* 62, 6512–6524. doi: 10.1021/acs.jmedchem.9b00089

Zhang, S., and Zhang, Z.-Y. (2007). PTP1B as a drug target: recent developments in PTP1B inhibitor discovery. *Drug Discov. Today* 12, 373–381. doi: 10.1016/j.drudis.2007.03.011

Zhou, W., Ma, L., Yang, J., Qiao, H., Li, L., and Guo, Q., et al. (2019). Potent and specific MTH1 inhibitors targeting gastric cancer. *Cell Death Dis.* 10:434. doi: 10.1038/s41419-019-1665-3

Conflict of Interest: TH is listed as an inventor of patents describing NUDT1 and NUDT5 inhibitors. EH is listed as an inventor of a patent describing NUDT1 inhibitors. The patents are fully owned by a non-profit public foundation, the Helleday Foundation, and TH and UW are member of the foundation board developing inhibitors toward and in the clinic (NCT03036228). TL is an employee of AstraZeneca, but performed all experimental work associated with this publication while at Chemical Biology Consortium Sweden.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Michel, Homan, Wiita, Pedersen, Almlöf, Gustavsson, Lundbäck, Helleday and Warpman Berglund. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership