COMBINING SIMULATIONS, THEORY, AND EXPERIMENTS INTO MULTISCALE MODELS OF BIOLOGICAL EVENTS

EDITED BY: Fabio Trovato, Peter J. Bond, Joanna Trylska and Peter Guy Wolynes

PUBLISHED IN: Frontiers in Molecular Biosciences

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 978-2-88974-025-3
DOI 10.3389/978-2-88974-025-3

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: frontiersin.org/about/contact

COMBINING SIMULATIONS, THEORY, AND EXPERIMENTS INTO MULTISCALE MODELS OF BIOLOGICAL EVENTS

Topic Editors:
Fabio Trovato, Freie Universität Berlin, Germany
Peter J. Bond, Bioinformatics Institute (A*STAR), Singapore
Joanna Trylska, University of Warsaw, Poland
Peter Guy Wolynes, Rice University, United States

Citation: Trovato, F., Bond, P. J., Trylska, J., Wolynes, P. G., eds. (2021). Combining Simulations, Theory, and Experiments Into Multiscale Models of Biological Events. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-025-3

Table of Contents

05 ***Editorial: Combining Simulations, Theory, and Experiments Into Multiscale Models of Biological Events***
Fabio Trovato, Joanna Trylska, Peter J. Bond and Peter G. Wolynes

08 ***Coarse-Grained Simulation of Mechanical Properties of Single Microtubules With Micrometer Length***
Jinyin Zha, Yuwei Zhang, Kelin Xia, Frauke Gräter and Fei Xia

19 ***Mutations of N1 Riboswitch Affect its Dynamics and Recognition by Neomycin Through Conformational Selection***
Piotr Chyży, Marta Kulik, Suyong Re, Yuji Sugita and Joanna Trylska

31 ***Case Report: Bayesian Statistical Inference of Experimental Parameters via Biomolecular Simulations: Atomic Force Microscopy***
Sotaro Fuchigami, Toru Niina and Shoji Takada

37 ***CUTie2: The Attack of the Cyclic Nucleotide Sensor Clones***
Florencia Klein, Florencia Sardi, Matías R. Machado, Claudia Ortega, Marcelo A. Comini and Sergio Pantano

45 ***Role of Electrostatic Hotspots in the Selectivity of Complement Control Proteins Toward Human and Bovine Complement Inhibition***
Yogesh B. Narkhede, Avneesh K. Gautam, Rohaine V. Hsu, Wilson Rodriguez, Nehemiah T. Zewde, Reed E. S. Harrison, Pablo R. Arantes, Zied Gaieb, Ronald D. Gorham Jr., Chris Kieslich, Dimitrios Morikis, Arvind Sahu and Giulia Palermo

55 ***Bottom-Up Coarse-Grained Modeling of DNA***
Tiedong Sun, Vishal Minhas, Nikolay Korolev, Alexander Mirzoev, Alexander P. Lyubartsev and Lars Nordenskiöld

72 ***Multi-Scale Flexible Fitting of Proteins to Cryo-EM Density Maps at Medium Resolution***
Marta Kulik, Takaharu Mori and Yuji Sugita

85 ***Molecular Basis of Class B GPCR Selectivity for the Neuropeptides PACAP and VIP***
Chenyi Liao, Jacob M. Remington, Victor May and Jianing Li

95 ***Characterization of Weak Protein Domain Structure by Spin-Label Distance Distributions***
Irina Ritsch, Laura Esteban-Hofer, Elisabeth Lehmann, Leonidas Emmanouilidis, Maxim Yulikov, Frédéric H.-T. Allain and Gunnar Jeschke

115 ***The Effect of Cholesterol on Membrane-Bound Islet Amyloid Polypeptide***
Mikkel Christensen, Nils A. Berglund and Birgit Schiøtt

131 ***Molecular Mechanics Study of Flow and Surface Influence in Ligand–Protein Association***
Shivansh Kaushik and Chia-en A. Chang

142 ***HEMNMA-3D: Cryo Electron Tomography Method Based on Normal Mode Analysis to Study Continuous Conformational Variability of Macromolecular Complexes***
Mohamad Harastani, Mikhail Eltsov, Amélie Leforestier and Slavica Jonic

160 ***Nascent Folding of Proteins Across the Three Domains of Life***
Mateusz Chwastyk and Marek Cieplak

169 ***From System Modeling to System Analysis: The Impact of Resolution Level and Resolution Distribution in the Computer-Aided Investigation of Biomolecules***
Marco Giulini, Marta Rigoli, Giovanni Mattiotti, Roberto Menichetti, Thomas Tarenzi, Raffaele Fiorentini and Raffaello Potestio

198 ***Thermodynamic Origin of Differential Excipient-Lysozyme Interactions***
Jas Kalayan, Robin A. Curtis, Jim Warwicker and Richard H. Henchman

211 ***A Primer on the oxDNA Model of DNA: When to Use it, How to Simulate it and How to Interpret the Results***
A. Sengar, T. E. Ouldridge, O. Henrich, L. Rovigatti and P. Šulc

233 ***Elucidating Axonal Injuries Through Molecular Modelling of Myelin Sheaths and Nodes of Ranvier***
Marzieh Saeedimasine, Annaclaudia Montanino, Svein Kleiven and Alessandra Villa

243 ***Quantitative Modeling of Protein Synthesis Using Ribosome Profiling Data***
Vandana Yadav, Inayat Ullah Irshad, Hemant Kumar and Ajeet K. Sharma

253 ***Generating Chromosome Geometries in a Minimal Cell From Cryo-Electron Tomograms and Chromosome Conformation Capture Maps***
Benjamin R. Gilbert, Zane R. Thornburg, Vinson Lam, Fatema-Zahra M. Rashid, John I. Glass, Elizabeth Villa, Remus T. Dame and Zaida Luthey-Schulten

275 ***Investigating Crosstalk Among PTMs Provides Novel Insight Into the Structural Basis Underlying the Differential Effects of Nt17 PTMs on Mutant Httex1 Aggregation***
Anass Chiki, Zhidian Zhang, Kolla Rajasekhar, Luciano A. Abriata, Iman Rostami, Lucien F. Krapp, Driss Boudeffa, Matteo Dal Peraro and Hilal A. Lashuel

292 ***Exploring Covalent Docking Mechanisms of Boron-Based Inhibitors to Class A, C and D β-Lactamases Using Time-Dependent Hybrid QM/MM Simulations***
Łukasz Charzewski, Krystiana A. Krzyśko and Bogdan Lesyng

307 ***Coarse-Grained Modeling and Molecular Dynamics Simulations of Ca^{2+}-Calmodulin***
Jules Nde, Pengzhi Zhang, Jacob C. Ezerski, Wei Lu, Kaitlin Knapp, Peter G. Wolynes and Margaret S. Cheung

frontiers in Molecular Biosciences

EDITORIAL
published: 19 November 2021
doi: 10.3389/fmolb.2021.797754

Editorial: Combining Simulations, Theory, and Experiments into Multiscale Models of Biological Events

Fabio Trovato[1]*, *Joanna Trylska*[2], *Peter J. Bond*[3,4] and *Peter G. Wolynes*[5]

[1]Department of Physics, Freie Universität, Berlin, Germany, [2]Centre of New Technologies, University of Warsaw, Warsaw, Poland, [3]Bioinformatics Institute (A*STAR), Singapore, Singapore, [4]Department of Biological Sciences, National University of Singapore, Singapore, Singapore, [5]Center for Theoretical Biological Physics, Rice University, Houston, TX, United States

Keywords: multiscale modeling, integrative modeling, molecular simulations, chromosome, binding, folding, multi-domain proteins, experiments

Editorial on the Research Topic

Combining Simulations, Theory, and Experiments into Multiscale Models of Biological Events

The number of publications that combine experiments and computer simulations has been growing steadily in the last 10 years. However, several challenges still need to be addressed in order to achieve a systematic integration, especially in the context of multiscale modeling of biological events. Computationally connecting the different scales—or more precisely the different system resolutions—constitutes one such difficulty of multiscale models. A second challenge is to choose the sources of experimental data that are most informative for parameterizing both the structural model and its inter-molecular interactions. This includes the modeling of the experimental conditions in order to avoid the misinterpretation of the results.

Quantum methods (QM) coupled to atomistic models (AT) have been instrumental in addressing, for example, the catalysis within an enzyme binding pocket. Five contributions in this special issue investigate biophysical events using multiscale QM/AT methods. To computationally explore slower events occurring in larger systems, it is often necessary to simplify the structural details of the system, an operation known as coarse-graining (CG). In this special issue, seven contributions employ such reduced models and three couple them with higher-resolution models, in a multiscale fashion. The interested reader is referred to the comprehensive reviews of Sun et al., and Giulini et al., as a guide to understanding the mathematical foundations of coarse-graining and multiscale methods, how to parameterize the inter-molecular forces and the relationship between a model resolution and its ability to reproduce the physicochemical properties of interest.

In this special issue, we have collected 22 manuscripts that emphasize the current efforts of the international scientific community to combine together experiments and computer modeling of biological events at multiple scales. A number of experiments have been performed across the different studies. They are cryo-EM, cryo-ET, FRET, EPR, AFM, CD, 3C[1] and biological assays. Computational methods include MD, MC, NMA, MSM, ChK[2], in silico mutations and whole-cell modeling. The investigated systems are represented at one or multiple among the QM, AT or CG resolutions or as a continuum, as in the Finite Element (FE) method.

OPEN ACCESS

Edited and reviewed by:
Massimiliano Bonomi, Institut Pasteur, France

***Correspondence:**
Fabio Trovato
fabiotrovato@gmail.com

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *19 October 2021*
Accepted: *02 November 2021*
Published: *19 November 2021*

Citation:
Trovato F, Trylska J, Bond PJ and Wolynes PG (2021) Editorial: Combining Simulations, Theory, and Experiments into Multiscale Models of Biological Events. Front. Mol. Biosci. 8:797754. doi: 10.3389/fmolb.2021.797754

[1]MD, Molecular Dynamics; NMA, Normal Mode Analysis; MC, Monte Carlo; MSM, Markov State Modeling; ChK, Chemical Kinetics.

[2]cryo-EM, Cryogenic Electron Microscopy; cryo-ET, Cryogenic Electron Tomography; FRET, Förster Resonance Energy Transfer; EPR, Electron Paramagnetic Resonance; AFM, Atomic Force Microscopy; CD, Circular Dichroism; 3C, Chromosome Conformation Capture.

Two research groups benchmark their methods for generating biomolecular structures. Fuchigami et al. combine AFM with MD simulations of a CG nucleosome model, modified to account for the AFM tip. With the aim of finding the nucleosome structure consistent with AFM images, the authors simulate nucleosome conformers while simultaneously inferring the geometric AFM tip radius on which the model (and the images) depend.

Kulik et al. combine a multiscale AT/CG protocol and mid-resolution cryo-EM maps to search the conformational space of multi-domain proteins and protein complexes sizing up to 2,211 residues. Their protocol is able to fit a flexible molecular model to cryo-EM maps with reasonable accuracy, therefore capturing states of a biomolecule that differ by large-scale movements, such as in the case of the DNA polymerase. The authors observe that their multiscale protocol is more beneficial for large biomolecules.

Harastani et al. present a cryo-ET sub-tomogram data analysis approach, called HEMNMA-3D, to study continuous conformational variability of the nucleosome. By generating a reference structure from the sub-tomograms and its normal modes, the authors retrieve known collective modes of a nucleosome. At variance with NMA-based methods, which fit an atomic structure to a cryo-EM map, HEMNMA-3D is designed to obtain a low-dimensional representation of the heterogeneity of a given set of EM maps, such as sub-tomograms.

Gilbert et al. devise a whole-cell modeling strategy and apply it to generate single-cell chromosome conformations of a genetically minimal bacterial cell, called Syn3A. Cryo-ET images are used to determine the cell size and ribosome distribution, within which the circular CG chromosome is embedded. This work is interesting in several aspects. A number of observations and model assumptions, grounded in the careful analysis of gene expression profiles and cryo-ET images, lead the authors to suggest that the chromosome of Syn3A cells has no substantial supercoiling or interactions between DNA segments that are distant in sequence. These initial assumptions are validated using contact maps calculated from preliminary 3C experiments and they generate a final cell model in which the circular chromosome is organized into territories and is not separated from the other cellular components, as instead occurs in *E. coli*. This study and that of Zha et al. emphasize that the generation of an accurate spatial model is often a delicate, preliminary step before any simulations can be conducted. In order to calculate the various model parameters, in fact, it is of utmost importance to choose informative biophysical and complementary methodologies to validate the computational results.

Several other DNA models exist in the literature, as reviewed by Sun et al.. Notable among them is oxDNA, which Sengar et al. describe in detail, discussing when it is worthwhile to use it and providing guidance to setting up and analyze the simulations.

Two additional large-scale models are presented by Zha et al., Saeedimasine et al.. The first research group tests an ultra coarse-graining strategy to generate microtubule models from 1 to 12 microns. Because these models reproduce known experimental mechanical properties, they are usable in future investigations aimed at understanding the supramolecular organization and small deformations of bundles of microtubules Saeedimasine et al.. propose a CG/FE multiscale model of the axolemma, which provides an explanation for the susceptibility of neurons to rupture under an applied mechanical stress. At the smallest scale, this model includes the myelin sheath and nodes of Ranvier lipid bilayers, represented at a CG resolution. Combining the results of the CG simulations with a FE model of the entire axon, the authors conclude that rupture is more likely to occur at the nodes of Ranvier instead of myelin, because myelin is observed to be more mechanically resistant.

Protein co-translational folding is the focus of the works of Chwastyk and Cieplak, and Yadav et al. The first research group studies how the geometry of the ribosome influences the co-translational folding of three model proteins, by performing simulations with different ribosome structures. The simulations suggest that protein synthesis is more obstructed in eukaryotes than in archaea and bacteria, which is consistent with constriction sites found in different positions of the exit tunnel of these ribosomes. Whether these constriction mechanisms are active in cells is unclear from the simulations alone because the ribosomes are approximated as rigid bodies. To predict the rate of protein synthesis or the populations of different nascent polypeptide states from an mRNA sequence, it is advisable to use mathematical and chemical kinetics modeling Yadav et al.. Since the predictions are sensitive to the codon translation rates used, Yadav et al. discuss how to estimate these kinetic parameters accurately from ribosome profiling data.

The characterization of intrinsic disorder in proteins is an important biophysical problem with many potential medical applications. An especially challenging aspect is the characterization of weak deviations from the random-coil state, due to protein residues that interact non-randomly and weakly. Towards this aim, Ritsch et al. combine distance distributions obtained from multiple EPR experiments with protein ensemble modeling. Because the method to analyze the raw EPR data typically influences the shape of the distance distributions used for structure generation, a model that attempts to detect weak deviations from a random-coil may contain a bias. To reduce such bias, the authors propose an overlap criterion for the distance distributions and a two-step modeling strategy to introduce the distance constraints. Under such controlled conditions, 19 distance distribution restraints are sufficient to observe deviations of a 133-residue segment from random-coil behavior.

Protein intrinsic disorder is also modulated by Post Translational Modifications (PTMs) and the presence of lipid membranes. Chiki et al. study the effects of PTMs on the aggregation of the first 17 amino acids of the Huntingtin protein. Key to the experimental success of their study is the production of phosphorylated T3 or acetylated K6 mutants within a background of oxidized M8. MD simulations suggest that the modified peptides explore a dynamic ensemble of disordered and helical conformations, with the helical conformations possessing a high fraction of short N-term helices. The increased content of N-terminal helical structure retards aggregation, in agreement with time-dependent AFM images. Consistent with a heterogeneous structural ensemble of the modified peptides, the distance distributions calculated from the simulations show a wide, often multi-peaked shape. Given the potential to detect weak deviations, Ritsch et al. EPR experiments performed on the Huntingtin protein could be used in the future to confirm these results.

Using molecular simulations, Christensen et al. study the effects of lipid bilayers of varying composition on the insertion mechanism of IAPP, a disordered peptide involved in glucose

metabolism and type-II diabetes. The authors observe stabilization of IAPP helical states near the membrane and a dependence of IAPP insertion on the lipid composition. The presence of cholesterol causes the membrane to become more ordered and therefore less accessible to IAPP. The authors propose this mechanism to be protective with regard to amyloid formation.

Klein et al. extend their previous computational method to design a FRET sensor for the detection of cyclic guanosine 3′-5′-monophosphate (cGMP). As a first step, the authors generate the structure of the FRET sensor, which includes a linker and a FRET pair made of two fluorescent proteins. Subsequently, CG MD simulations are performed to exhaustively sample the conformations of the model sensor. The good agreement between the calculated and *in vitro* measured FRET efficiencies makes this method an interesting tool for the development of small molecule sensors.

Calmodulin (CaM) has been extensively investigated both computationally and experimentally. Nde et al. generate a CG model of CaM, which includes the effects of Ca^{2+} binding, albeit implicitly. The authors re-parameterize the AWSEM force field in order to reproduce CaM's radius of gyration and the ratio of CaM's open and closed states. Such a method might be applicable to study other flexible multi-domain proteins that respond to the binding of small molecules.

In the remainder of this editorial, we highlight six investigations on the theme of small molecule or protein binding to either proteins or RNA. Kaushik and Chang study the association between the HIV protease and xk263, under conditions of continuous flow of ligands close to a non-polar monolayer surface. Such conditions are reminiscent of those used in Surface Plasmon Resonance (SPR) for measuring binding constants. Besides the results of the Brownian dynamics simulations, this work could be applied to study the effects of solvent flow and surfaces on the association of any protein-ligand system, therefore paralleling and complementing SPR measurements.

Using enhanced MD simulations, Chyży et al. study the binding of neomycin to four single-mutants of the N1 riboswitch. The nucleobase mutations in the apical loop, although distant from the binding pocket, significantly affect the neomycin-riboswitch interactions. By carefully comparing their results with previous experimental work, the authors provide insights into the riboswitch structure-dynamics–activity relationship.

Narkhede et al. investigate the contribution of electrostatic forces to the selectivity of viral proteins SPICE and VCP towards the C3b complement protein, a key component of the complement immune system. By combining electrostatic calculations, in-silico alanine scanning and hotspot analysis, the authors identify sites resistant to local perturbation, where the electrostatic potential is likely to be evolutionary conserved. Their calculations are further supported by cofactor activity assays.

Liao et al. study the interactions between the neuropeptides PACAP and VIP with several receptors. Using a combination of homology modeling, enhanced MD simulations for the binding process and Markov State modeling, the authors find that different pathways are explored depending on the specific receptor/neuropeptide pair. The predicted receptor conformations and transition rates along the different binding pathways might be a useful dataset for guiding future drug design efforts.

Kalayan et al. investigate the influence of excipients on protein precipitation, which is relevant to the field of pharmaceutical protein formulations. The authors apply molecular simulations and their Energy-Entropy Multiscale Cell Correlation method to understand why lysozyme precipitates when interacting with the polyanion tripolyphosphate (TPP) but not with citrate. Overall, solvent and protein stabilization upon TPP binding are expected to drive binding via cross-linking with other proteins and therefore precipitation. By coarse-graining the current model, future simulations might enable observation of spontaneous cross-linking events, therefore validating this putative mechanism of precipitation.

Charzewski et al. propose a time-dependent QM/MM multiscale methodology to study the covalent docking between beta-lactamase and boron-based inhibitors. The authors apply this methodology to covalently dock three beta-lactamase enzymes and two boron-based inhibitors, obtaining several insights into the intermediate states. The simulation results suggest that this methodology could be effective in finding docking pathways by other boron inhibitors as well.

We conclude this editorial with a few additional messages. The first regards the importance of carefully estimating the experimental conditions. For example, when using FRET/EPR data, the effects of the fluorescent/spin label should be assessed either with control experiments Ritsch et al. or by directly incorporating the effects of the label into the model Klein et al. Similarly, the inclusion of geometrical features of an AFM tip can improve the model based on AFM images Fuchigami et al. The second message concerns the use of complementary experimental data when a single experimental technique is not sufficient Ritsch et al. or when it is necessary to validate a complex model Gilbert et al. Lastly, we underline that building a multiscale model is beneficial to test new hypotheses on complex events such as brain injury Saeedimasine et al. or to have access to multiple time and length scales for calculating different experimental quantities accurately Zha et al.

AUTHOR CONTRIBUTIONS

FT wrote the manuscript. JT, PB, and PW revised the manuscript.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Trovato, Trylska, Bond and Wolynes. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 15 February 2021
doi: 10.3389/fmolb.2020.632122

Coarse-Grained Simulation of Mechanical Properties of Single Microtubules With Micrometer Length

*Jinyin Zha[1], Yuwei Zhang[1], Kelin Xia[2,3], Frauke Gräter[4,5,6] and Fei Xia[1,7]**

[1]School of Chemistry and Molecular Engineering, East China Normal University, Shanghai, China, [2]Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, Singapore, [3]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore, [4]Interdisciplinary Centre for Scientific Computing (IWR), Heidelberg University, Heidelberg, Germany, [5]Heidelberg Institute for Theoretical Studies (HITS), Schloß-Wolfsbrunnenweg 35, Heidelberg, Germany, [6]Max Planck School Matter to Life, Jahnstraße 29, Heidelberg, Germany, [7]Shanghai Engineering Research Center of Molecular Therapeutics and New Drug Development, NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai, China*

Microtubules are one of the most important components in the cytoskeleton and play a vital role in maintaining the shape and function of cells. Because single microtubules are some micrometers long, it is difficult to simulate such a large system using an all-atom model. In this work, we use the newly developed convolutional and K-means coarse-graining (CK-CG) method to establish an ultra-coarse-grained (UCG) model of a single microtubule, on the basis of the low electron microscopy density data of microtubules. We discuss the rationale of the micro-coarse-grained microtubule models of different resolutions and explore microtubule models up to 12-micron length. We use the devised microtubule model to quantify mechanical properties of microtubules of different lengths. Our model allows mesoscopic simulations of micrometer-level biomaterials and can be further used to study important biological processes related to microtubule function.

Keywords: microtubule, persistence length, ultra-coarse-grained model, mechanical property, convolutional and K-means coarse-graining

OPEN ACCESS

Edited by:
Fabio Trovato,
Freie Universität Berlin, Germany

Reviewed by:
Gregory Voth,
University of Chicago, United States
Adolfo Poma,
Institute of Fundamental Technological Research, Polish Academy of Sciences, Poland

****Correspondence:***
Fei Xia
fxia@chem.ecnu.edu.cn

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *22 November 2020*
Accepted: *30 December 2020*
Published: *15 February 2021*

Citation:
Zha J, Zhang Y, Xia K, Gräter F and Xia F (2021) Coarse-Grained Simulation of Mechanical Properties of Single Microtubules With Micrometer Length. Front. Mol. Biosci. 7:632122. doi: 10.3389/fmolb.2020.632122

INTRODUCTION

Microtubule (Fratzl and Weinkamer, 2007; Conde and Caceres, 2009) is one of the primary components of the cytoskeleton which is responsible for the morphology of eukaryotic cells. The building blocks of microtubules are α- and β-tubulins which assemble into a hollow cylinder structure with grooves. The length of a single microtubule can reach 10 μm and they are associated with each other in a hierarchical organization such as in the axons (Lazarus et al., 2015; Soheilypour et al., 2015). A microtubule is generally composed of 13 protofilaments (Kollman et al., 2010) lying side by side along the longitudinal direction, where the position of each protofilament shifts slightly to each other. The mechanical properties (Franz et al., 2020) of microtubules play important roles in sustaining the cell shape, executing the mitotic division and related physiological functions. For instance, they serve as tracks for the motor proteins kinesin (Gittes et al., 1996) to transport the substances required for cells.

In order to uncover the microscopic mechanisms of microtubules for transporting substances in cells, much effort has been put into measuring the biomechanical properties (Dye et al., 1993; Gittes et al., 1993; Venier et al., 1994; Kurachi et al., 1995; Felgner et al., 1996; Vinckier, 1996; Wagner et al.,

1999; Kis et al., 2002; Le Goff et al., 2002; VanBuren et al., 2002; de Pablo et al., 2003; Janson and Dogterom, 2004; Pampaloni et al., 2006; Schaap et al., 2006; Hawkins et al., 2013; Cross, 2019; Szatkowski et al., 2019) of microtubules associated with their function, such as Young's modulus E, the flexural rigidity κ, and the persistence length l_p. Previous studies have exploited a variety of advanced techniques including thermal fluctuation (Gittes et al., 1993), hydrodynamic flow (Venier et al., 1994), atomic force microscopy (Vinckier, 1996), and optical tweezers (Felgner et al., 1996) to measure these mechanical quantities. So far, reported Young's modulus E, the flexural rigidity κ, and persistence length l_p for microtubules with different lengths span a large range of 0.31–7.0 GPa (Kurachi et al., 1995; Vinckier, 1996; Kis et al., 2002; de Pablo et al., 2003; Schaap et al., 2006; Kis et al., 2008), $0.37–4.0 \times 10^{-23}$ Nm^2 (Gittes et al., 1993; Venier et al., 1994; Kurachi et al., 1995; Felgner et al., 1996), and 2.0–6.3 mm (Gittes et al., 1993; Venier et al., 1994; Janson and Dogterom, 2004; Pampaloni et al., 2006), respectively. Kis et al. (Kis et al., 2002) first reported that the observed phenomenon of the measured persistence length was length-dependent, caused by the anisotropic association of heterodimeric α- and β-tubulins. This finding was further demonstrated by Pampaloni et al. (Pampaloni et al., 2006) using the single-particle tracking method and they observed that the persistence length l_p increased for short lengths varying from 2.6 to 47.5 μm. It approached a constant value of l_p^{∞} = 6.3 mm for lengths beyond 21 μm. All measured results unambiguously imply that microtubules do not follow the behavior of an ideal elastic rod and fail to be described with the worm-like chain model (Kroy and Frey, 1996; Taute et al., 2008) based on the isotropic hypothesis.

Theoretical modeling of microtubules (Kasas et al., 2004a; Kasas et al., 2004b; Molodtsov et al., 2005; VanBuren et al., 2005; Schaap et al., 2006; Deriu et al., 2007; Heuvel et al., 2007; Deriu et al., 2008; Dima and Joshi, 2008; Deriu et al., 2010; Sept and MacKintosh, 2010; Wells and Aksimentiev, 2010; Grafmuller and Voth, 2011; Ji and Feng, 2011a; b; Peter and Mofrad, 2012; Grafmuller et al., 2013; Theisen et al., 2013; Kononova et al., 2014; Lazarus et al., 2015; Soheilypour et al., 2015; Havelka et al., 2017; Stevens, 2017; Szatkowski et al., 2019; Tong and Voth, 2020) faces a huge challenge due to the micrometer length. All-atom molecular dynamics (MD) simulations have been primarily used to investigate the atomistic details of the interactions between the α- and β-tubulins (Deriu et al., 2007; Deriu et al., 2008; Sept and MacKintosh, 2010; Wells and Aksimentiev, 2010; Grafmuller and Voth, 2011; Grafmuller et al., 2013; Havelka et al., 2017) for short microtubules. Recent all-atom simulations (Tong and Voth, 2020) reached a length scale of a 1 μm microtubule including 3 protofilaments. Additionally, coarse-grained (CG) models (Dima and Joshi, 2008; Deriu et al., 2010; Ji and Feng, 2011a; Ji and Feng, 2011b; Theisen et al., 2013; Kononova et al., 2014; Lazarus et al., 2015; Soheilypour et al., 2015; Stevens, 2017; Szatkowski et al., 2019) have also been used to model longer microtubules. For instance, Deriu et al. (Deriu et al., 2010) constructed a 350 nm long microtubule to estimate its mechanical properties based on the elastic network model (Bahar et al., 1997; Atilgan et al., 2001; Xia and Lu, 2012; Xia et al., 2013a). Nevertheless, their model was still far shorter than the microtubules measured in experiments (Pampaloni et al., 2006). In some special CG models, each tubulin monomer of a microtubule was even represented by one CG bead (Ji and Feng, 2011a; Ji and Feng, 2011b; Lazarus et al., 2015) so that the whole microtubule model could reach the micrometer length. Alternatively, other theoretical methods based on continuous mechanics such as the finite element method (Kasas et al., 2004a; Kasas et al., 2004b; Schaap et al., 2006) were also used to evaluate the mechanical properties of macromaterials. However, these methods had an obvious disadvantage that they could not reflect the correct surface contour of microtubules, especially for the functional grooves serving as binding sites to interact with other proteins.

Ultra-coarse-grained (UCG) models (Dama et al., 2013; Davtyan et al., 2014; Dama et al., 2017; Zhang et al., 2017a; Jin et al., 2018; Jin and Voth, 2018) have been rapidly developed in recent years and applied to simulate the functions of biomolecular complexes such as the mechanical properties of F-actin (Katkar et al., 2018), the assembly of HIV capsids (Grime et al., 2016), and two-state conformational changes of proteins (Zhang et al., 2020). The bead resolution of UCG models is lower than the high-resolution models, with one UCG bead usually representing more than a few residues or even secondary structure elements. Previously, Voth and coworkers (Zhang et al., 2008; Zhang et al., 2009) developed the essential-dynamic coarse-graining (ED-CG) method to optimize the UCG representation (Sinitskiy et al., 2012) of proteins and later extended the framework to the density-based ED-CG method (Zhang and Voth, 2010). Xia and coworkers (Li et al., 2016a; Li et al., 2016b; Zhang et al., 2017b) developed a stepwise local iterative optimization-based fluctuation-maximization coarse-graining method to derive the optimal UCG representation (Wu et al., 2020) of large biomolecules. Further, Zhang et al. (Zhang et al., 2019) developed the convolutional and K-means coarse-graining (CK-CG) method to derive the UCG representation of biomolecules directly from the low-resolution cryo-EM density data (Ming et al., 2002; Esquivel-Rodriguez and Kihara, 2013), which allows the construction of models without any atomistic details needed further. Next, the UCG beads could be parameterized using the energy potential of the Multiscale Virtual Particle-based Anisotropic Network Model (MVP-ANM) (Xia et al., 2013b; Opron et al., 2014; Xia, 2017) for further functional studies. In the previous study (Zhang et al., 2019), we have validated the combined CK-CG and MVP-ANM methods in constructing UCG model for a correct estimation of the biomechanical properties of short segments F-actin and collagen fibrils. In this work, we attempt to verify the usefulness of UCG models in simulating the mechanical properties such as the length-dependent persistence length of long microtubules and expect to apply it to the simulation more complicated and hierarchically organized biomaterials (Zapp et al., 2020).

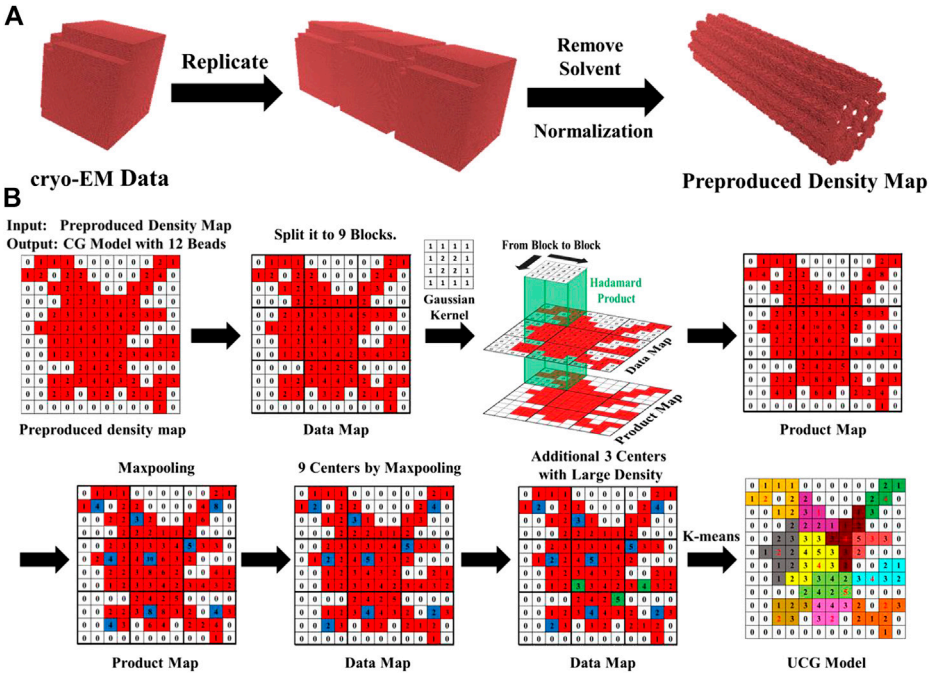


FIGURE 1 | (A) The data reprocessing includes replicating the density, removing solvent, and normalizing density. **(B)** The process of coarse-graining cryo-EM data into 12 UCG beads using the CK-CG method. The density map is first divided into 9 blocks. The Hadamard product for each block is obtained by using a Gaussian kernel and the subsequent max-pooling generates the initial positions of 9 UCG beads. The positions of the remaining 3 UCG beads are determined by other maximal densities in the blocks. Then, the 12 positions serve as the centers for clustering to yield the final UCG beads.

COMPUTATIONAL AND SIMULATION DETAILS

Data Preprocessing and CK-CG Method

The density file EMD-4043(Grange et al., 2017) of microtubule used for model construction was downloaded from the cryo-EM data bank online. The original length of microtubule in the file EMD-4043 is only 32 nm, in a resolution of 18.5 Å. In order to construct long microtubules, we replicated the density data shown in **Figure 1A** along the longitudinal direction and extended the 32 nm length to the different lengths ranging from 1 to 12 μm. Then, we removed the density of solvents from the raw data with a technique described in the previous work (Zhang et al., 2019) and evaluated the normalized density μ^s (r) according to μ^s (r) = $(\mu(r) - \mu_{min})/(\mu_{max} - \mu_{min})$, where μ (r) denotes the cryo-EM density data. After the preprocessing on data, the remaining densities are positive and ready as the input for coarse-graining by using CK-CG method.

Coarse-graining density data is actually to cluster the adjacent discretized density data into a group of single UCG particles. In the conventional K-means method, the positions of center points in initial guess are generated randomly, which renders the convergence quite slow. In the CK-CG method, we utilized an idea similar to the Convolutional Neutral Network model (Li et al., 2019) to generate the initial guess for clustering, including the steps of the domain partition, density convolution, and max-pooling, displayed in **Figure 1B**. In more detail, the cryo-EM density data is firstly divided into a series of 3D block matrixes with the voxel b_{ijl}, where the subscripts i, j, and l denote dimensions. In the next convolution operation, we used a 3D Gaussian kernel matrix with the voxel k_{ijl} to execute a Hadamard product based convolution operation and saved the matrix values h_{ijl}.

In the subsequent max-pooling, the position corresponding to the largest density value in each block was chosen as the initial position of UCG bead for clustering. It has to be mentioned that if the expected number of UCG beads N is larger than the number of divided blocks M, the excess number of (N-M) beads might be automatically assigned into the blocks with largest values. After generating the initial guess, the clustering using K-means was carried out until the convergence was achieved. More details about the procedure of CK-CG workflow are described in the previous work (Zhang et al., 2019).

Multiscale Virtual Particle-Based Anisotropic Network Model

UCG beads derived from CK-CG could be connected with harmonic springs and parameterized to be the MVP-ANM model developed by Xia et al. (Xia et al., 2013b; Opron et al., 2014; Xia, 2017). The energy function $E_{MVP-ANM}$ is a sum of harmonic interactions as shown in **Eq. 1**, where γ_{ij} is a

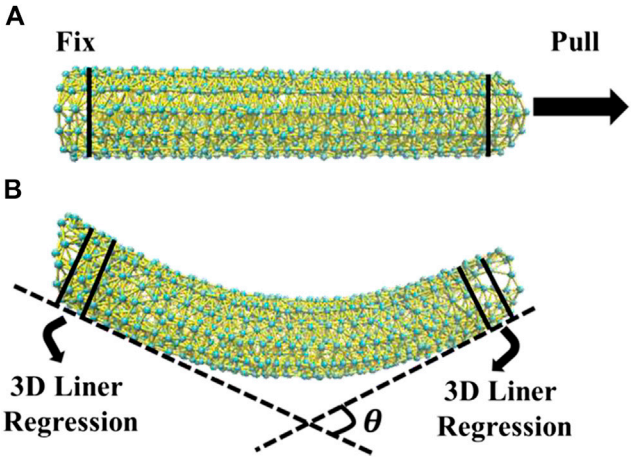


FIGURE 2 | (A) An illustration of force-clamped MD simulation for measuring Young's modulus of microtubule. **(B)** The estimation of persistence length using a method of 3D linear regression based on thermal fluctuation.

heterogeneous force constant, d_{ij} and d_{ij}^0 denote the instantaneous and equilibrium distances of pairwise UCG beads i and j, and $H(d_c - d_{ij}^0)$ is a Heaviside function depending on a cutoff distance d_c.

$$E_{\mathrm{MVP-ANM}} = \sum_i \sum_{j>i} \gamma_{ij} \left(d_{ij} - d_{ij}^0\right)^2 \cdot H\left(d_c - d_{ij}^0\right) \tag{1}$$

The parameter γ_{ij} contains both the contributions from the density and distance of pairwise UCG beads i and j, with the expression shown in **Eq. 2**:

$$\gamma_{ij} = A \cdot (1 + m_i)(1 + m_j) \cdot e^{-d_{ij}^2/R^2} \tag{2}$$

The parameter A is a scaling factor to control the mechanical properties of whole model and can be determined by fitting the experimental Young's modulus using force-clamped MD simulation (Israelewitz et al., 2001; Gräter and Grubmuller, 2007) as shown in **Figure 2A**. The term of $(1 + m_i)(1 + m_j)$ is a function representing the mass contributions from the beads i and j. The distance contribution is defined as a Gaussian function in which R is a parameter to control the bond strength of pairwise beads i and j, and it is only taken into account if the equilibrium distance d_{ij}^0 is less than the chosen cutoff d_c. A minimal cutoff can be determined through MD simulations by judging whether the structures of microtubules remain stable all the time. Please see the details for determining a minimal d_c described in *Parameterization of UCG Model* below. Additionally, the parameter γ_{ij} decays significantly as $d_{ij} > 3R$. Thus, we define the magnitude of R to be one-third of the minimal cutoff of d_c.

Evaluation of Persistence Length From UCG MD Simulation

We performed the UCG MD simulation of microtubules at 300 K and calculated the persistence length according the definition in **Eq. 3**:

$$\langle \cos\theta \rangle = e^{-L/\mathrm{lp}} \tag{3}$$

where $\langle \cos\theta \rangle$ means the average value of the cosine of θ ($0 \le \theta \le \pi/2$) and θ denotes the intersection angle of two tangent lines at the positions of the head and tail of microtubules. In order to calculate it, we recorded the Cartesian coordinates of all the beads every 12 ps in UCG MD simulation, and obtained a number of frames for further analysis. For the structure in each frame, we used a 3D liner regression method in **Figure 2B** to solve **Eq. 3** and derived the persistence length from the calculated average value of $\langle \cos\theta \rangle$. All the UCG MD simulations mentioned above were carried out with using the software LAMMPS (Plimpton, 1995) in the NVT ensembles at 300K. The Langevin dynamics simulation (Grønbech-Jensen, 2019) was performed under the nonperiodic condition with the damping factor being 10.0 ps. The time step of simulation is chosen to be 10.0 fs for the simulation of Young's modulus. For the estimation of persistence length from a long time UCG simulation, since the lightest bead in the UCG model is around 6.0 kg/mol, nearly 6000 times greater than H atom, we

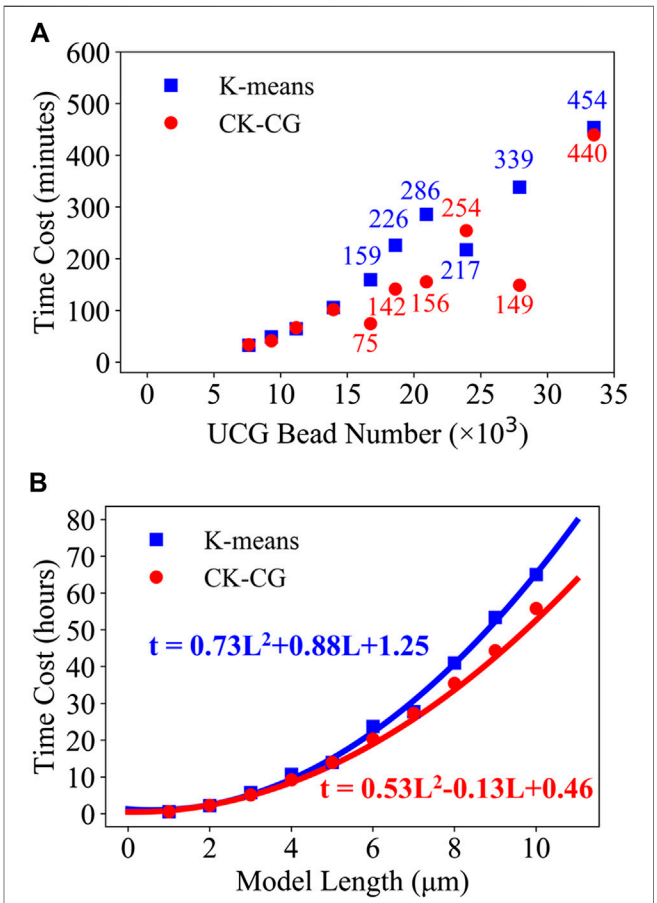


FIGURE 3 | The time cost for building models with **(A)** different numbers of UCG beads and **(B)** different lengths. The blue squares and red dots represent the results from the K-means and CK-CG methods, respectively. The integers in plot **(A)** denote the exact time cost in minutes. In the plot (b), the set of data are fitted with two quadratic curves, respectively.

chose a moderate value of 3.0 ps for the integration steps. The UCG MD simulation for a single microtubule with micrometer length could reach the time scale of 60 ms.

It has to be mentioned that the persistence length of microtubules was frequently estimated from the bending measurements in experiments (Kurachi et al., 1995; Felgner et al., 1996; Kis et al., 2008) and simulations (Kononova et al., 2014; Wu and Adnan, 2018). However, our UCG model is not suitable for being used in a direct simulation of bending deformation, since the MVP-ANM potential used in our model is harmonic and invalid in large deformation. A simulation of the complicated process such as the dissociation and association of tubulins from microtubule requires a more complex potential such as a Gō-like potential (Zhang et al., 2017a; Poma et al., 2017; Poma et al., 2018; Zhang et al., 2020).

RESULTS AND DISCUSSIONS

Coarse-Graining Microtubules With Different Bead Resolutions and Lengths

Previously, we have constructed a UCG model for a short 160 nm microtubule (Zhang et al., 2019). In this work, we attempt to coarse-grain the microtubules with the lengths from 1 to 10 μm which are much longer than the previous one by 1 or 2 orders of magnitude. In order to demonstrate the efficiency of CK-CG over the conventional K-means methods, we test its time costs in coarse-graining the microtubules into different numbers of beads and lengths.

The first test given is the time cost for coarse-graining a 1 μm microtubule UCG model into different numbers of UCG beads. The test was done by running C++ language codes with one CPU on a CentOS platform. As shown in **Figure 3A**, the derived numbers of UCG beads for a fixed-length microtubule range from 7609 to 33480 beads per micrometer, corresponding to the averaged masses of UCG beads from 22.0 to 5.0 kg/mol. **Figure 3A** shows that the performances of the CK-CG and K-means methods are similar to each other in deriving the smaller numbers of UCG beads such as 7009, 9300, 11160, and 13950, whereas they exhibit distinct difference for large UCG numbers. The time cost of coarse-graining into 16740, 18600, 20925, and 27900 UCG beads by using CK-CG is almost half of that of K-means. For instance, coarse-graining into 27900 beads by CK-CG needs merely 149 minutes, while the K-means method costs 339 min. Also, it is seen that the time costs of both methods for deriving 23914 and 33480 UCG beads are comparable to each other. However, no case in which K-means was obviously superior to CK-CG was observed in our ten tests. Thus, the performance of CK-CG in coarse-graining a fixed-length microtubule into different beads outperforms the K-means method by 47.4 ± 44.4 hours with 95% confidence interval, from the statistical perspective. In this test, no simple correlation such as the linear relationship was found to exist between the time cost and the number of coarse-grained beads.

Next, we tested how the time cost changes with the increased lengths of microtubules under the same computational condition. The UCG models were built with the constraint of 7609 beads per micrometer, with the microtubule lengths changing from 1 to 10 μm. **Figure 3B** shows the comparison of time cost of using CK-CG and K-means methods for coarse-graining. It is surprisingly observed that the time cost exhibits a regular quadratic relationship with the increased length. It is found that the performances of two methods are nearly the same as the microtubules are short below 5 μm. Beyond that, the difference of time cost between them gradually becomes large with the increased length of microtubule. At the length of 10 μm, the CK-CG method saves 10 h more than K-means to optimize a total number of 76090 UCG beads. This comparison indicates that a good initial guess of center positions is important for accelerating the convergence of data clustering, especially for the biomolecules with the length at large scale.

Determination of a Minimal Model for Microtubule

After validating the efficiency of CK-CG for coarse-graining micrometer microtubules, we are ready to construct a minimal model of microtubule for its mechanical simulation. The so-call "minimal model" must meet two requirements for functional study: one is that the number of UCG beads for a given length is so minimal that it can reduce the simulation time to the maximal extent. Secondly, it is expected that this model is capable of maintaining the geometric feature of microtubule, especially for the functional grooves. According to the two requirements, we designed a minimal model with the possible minimal number of beads, which is illustrated in **Figure 4**.

As shown in **Figure 4A**, the groove along the longitudinal direction of a microtubule is formed by two adjacent protofilaments. The simplest and minimal UCG model to characterize the feature of a groove is to represent each tubulin with two connected beads, with one shared by two neighboring tubulins. Since the average molar mass of tubulin is 55 kg/mol (Fratzl and Weinkamer, 2007), the estimated average molar mass of UCG beads is 22.5 kg/mol. To derive the positions of UCG beads, we set the average mass with the value of 22 kg/mol in the coarse-graining by CK-CG, which lead to a number of 7609 beads per 1 μm in *Coarse-Graining Microtubules With Different Bead Resolutions and Lengths*. **Figure 4B** shows the side and transverse views of the resulting microtubule model, where each UCG bead is virtualized with solid dots and the 13 protofilaments are in different colors.

Figure 4C shows the zoomed top and side views of a segment of density map of 8 tubulins, labeled with the upper cases A-H, where the density points belonging to the same UCG bead centers are in the same color. From the top and the side views, we can see clearly that the bottoms of most grooves are well represented by one UCG bead, such as the density regions connecting the adjacent "AB", "DE", or "BC" tubulins. Meanwhile, two adjacent tubulins such as "DG", "BE", or "EH" along a protofilament share the same beads. From the view of UCG beads, it can be judged that the UCG beads representing the groove bottoms of tubulins usually reside in a lower position, while the beads representing the tubulins along the

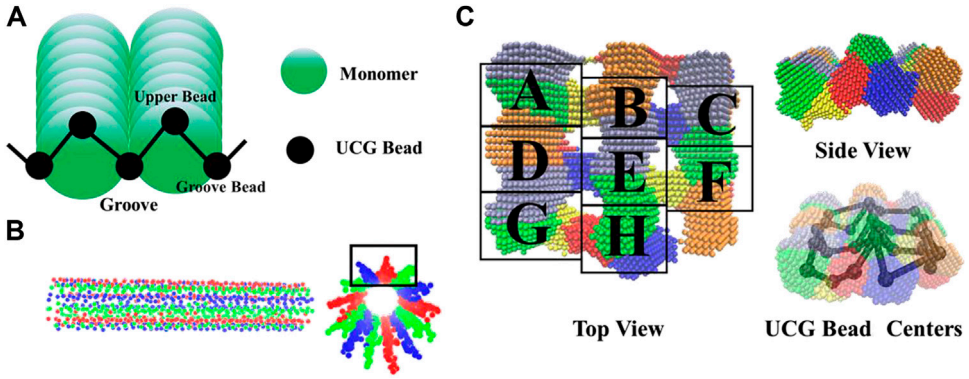


FIGURE 4 | (A) An illustration of the minimal number of UCG beads required for maintaining the grooves of microtubule. **(B)** The transverse and longitude views of the built models. **(C)** Zoomed top and side views of the density map of 8 adjacent tubulins, labeled with the upper cases A-H. The different colors denote the density regions which are clustered into the UCG beads, with the corresponding UCG bead centers shown.

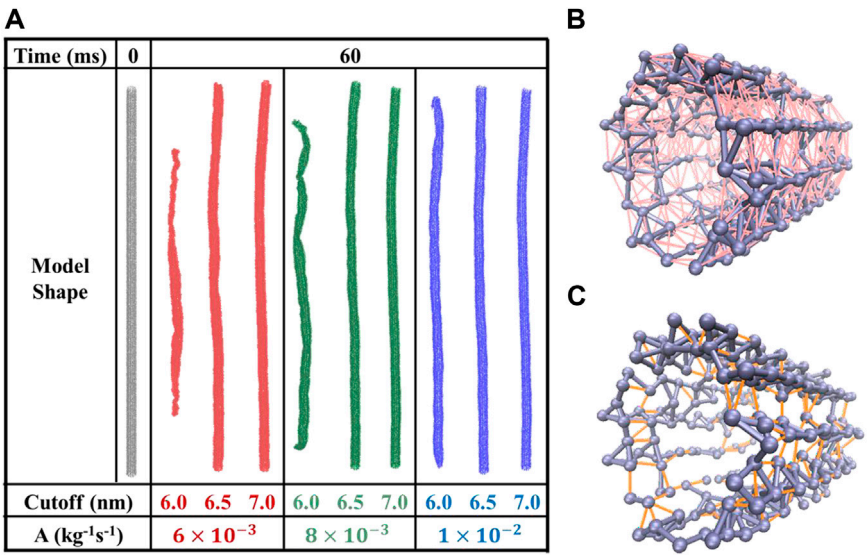


FIGURE 5 | (A) The structures of constructed UCG models for microtubules at the initial simulation time and after 60 ms. The cutoff distances adopt the values of 6.0, 6.5, and 7.0 nm for three models, with the tested parameter *A* adopting 6×10^{-3}, 8×10^{-3}, and 1×10^{-2} kg^{-1} s^{-1}, respectively. **(B)** The UCG bond distances below 4.4 nm are colored in gray, while those bonds in 4.4–6.6 nm are in pink. **(C)** The UCG bonds below 4.0 nm in gray and those in 4.0–4.4 nm are colored in orange.

protofilaments stay in a high position. Thus, the coarse-grained beads between two adjacent protofilaments maintain a geometric "V" shape along the contour surface of the microtubules, which is potentially important for functional studies including interacting proteins.

Parameterization of Ultra-Coarse-Grained Model

Using the potential energy of MVP-ANM model of **Eq. 1** to describe the interaction between pairwise UCG beads requires determining the two parameters *R* and *A* in **Eq. 2**. As mentioned in *Multiscale Virtual Particle-Based Anisotropic Network Model*, the parameter *R* could be defined as one-third of minimal cutoff distance d_c. In a word, we need to determine a minimal cutoff at first. As shown in **Figure 5A**, we constructed the UCG models with the cutoff set as the tentative values of 6.0, 6.5, and 7.0 nm, respectively. Meanwhile, the parameter *A* adopts the empirical values of 6.0×10^{-3}, 8.0×10^{-3}, and 1.0×10^{-2} $kg^{-1} \cdot s^{-1}$. After the 60 ms MD simulation, we found that all the three microtubule models with the cutoff of 6.0 nm have collapsed and their structures could not keep stable anymore. Besides, the UCG models with the cutoffs of 6.5 and 7.0 nm are stable in the long time MD simulation and maintain their tube shape well. The simulation results also indicate that the stabilities of microtubules did not depend on the choices of magnitudes of the parameter *A*.

To rationalize the choice of a cutoff of 6.5 nm, we analyze the number of harmonic bonds formed between pairwise UCG beads.

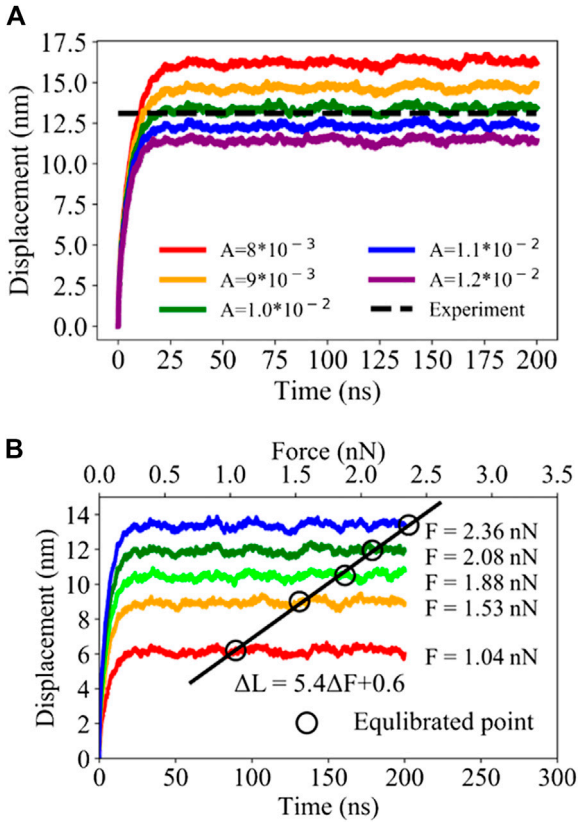


FIGURE 6 | (A) The relationship of the terminal displacements of microtubule models with the different values used for the parameter *A* in the units of kg^{-1} s^{-1}. The black dashed line represents the experimental displacement and the colorful curves denote the displacements of microtubule models changing with time under an external force of 2360 pN in force-clamp MD simulations. **(B)** The curves of the displacements of microtubule models under the different forces ranging from 1.04 to 2.36 nN, where the black hollow circles denote the average values of the end-to-end displacements sampled from 100–200 ns. The black line denotes the linear fitting of the circles.

In **Eq. 2**, the strength of harmonic bonds is weakened with an increasing distance *R*. When d_c adopts the cutoff 6.5 nm, we find that numerous bonds that describe strong interactions are defined between UCG beads, shown as gray lines in **Figure 5B** (Ji and Feng, 2011a; Kononova et al., 2014). Within the distance between 4.4 and 6.6 nm (2*R*–3*R*), bonds describing nonbonded interactions are found (shown in pink), and they are roughly 40 to 150 times weaker than the bonds within 4.4 nm. In contrast, **Figure 5C** shows the formed bonds within 4.0 nm in gray and the bonds in 4.0–6.0 nm in orange, as the cutoff adopts 6.0 nm. It is obvious that the two adjacent protofilaments in the UCG model do not feature sufficient bonding to keep the whole cylinder structure stable. Thus, the reason of collapse with the cutoff 6.0 nm lies in the lack of sufficient pairwise interactions to maintain the microtubule's stability.

Eventually, we chose the cutoff to be 6.5 nm for all models and the resulting *R* value is 2.2 nm. The scaling factor *A* is determined by fitting the longitudinal Young's modulus of a 100 nm short microtubule. There are two reasons that we need to fit Young's modulus to derive the parameter *A*. One is that the longitudinal Young's modulus is an independent property and has no influence on the measurement of the persistence length according to **Eq. 3**. The second is that the longitudinal Young's modulus is independent of the length *L* of microtubule and only intrinsically determined by the strength of force constant γ_{ij}. Thus, the parameter *A* in **Eq. 2** is tightly correlated with the longitudinal Young's modulus.

Kis et al. (2002) reported that a microtubule had a longitude Young's modulus of 100 MPa under 300 K. In our force-clamped MD simulation, one terminal of the 100 nm microtubule is fixed, while the other one is pulled with a constant force of 2.36 nN. In **Figure 6A**, different values of the parameter *A* are used for UCG models to measure the difference between the simulated displacements and the experimental result around 13.0 nm. As the parameter *A* adopts the value of 1.0×10^{-2} $kg^{-1} \cdot s^{-1}$, the simulated displacement of microtubule reaches the experimental value well. So far, both the parameters *R* and *A* for the models of microtubules are determined and ready for the subsequent simulation of mechanical properties. In addition, we also performed force-clamped UCG MD simulations with the same

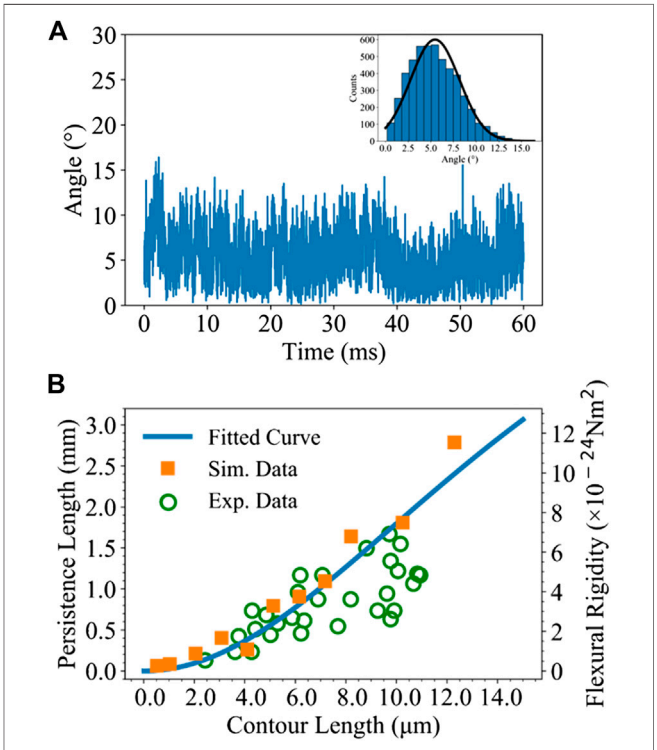


FIGURE 7 | (A) The angle fluctuation of a 2 μm microtubule model within 60 ms. The inset displays the histogram distribution of angles with a black fitted Gaussian curve. **(B)** The orange solid squares show the relationship of the simulated persistence length and flexural rigidity with respect to the contour length, fitted with a blue curve based on **Eq. 4**. The green hollow circles denote the experimental measured results (Pampaloni et al., 2006). The simulation time ranges from microseconds to milliseconds for the different microtubules.

parameters (A = 1.0 × 10^{-2} $kg^{-1}·s^{-1}$ and R = 2.2 nm) under different pulling forces ranging from 1.04 nN to 2.36 nN, as shown in **Figure 6B**. It can be observed that the equilibrated displacements have a good linear relationship with the applied forces. Based on the slope of the linear relationship, Young's modulus is calculated to be 99.2 MPa, which reproduces Young's modulus under different forces and verifies the validity of UCG model (Kis et al., 2002).

Evaluation of Persistence Length

The persistence length of microtubule is evaluated by using MD simulation based on its thermal fluctuation. It should be noted that the experimental measurement of thermal fluctuation usually occurred at time scale of seconds. However, it is impossible to perform such a long time MD simulation in silicon, even by employing the UCG model. The previous simulation performed by Ding et al. (Ding and Xu, 2011) for evaluating persistence length from thermal fluctuation was 50 μs. In our UCG MD simulation, we used an integration step of 3 ps and the simulation time for microtubules with micrometer length could reach milliseconds. A number of MD trajectories were saved for further analysis. **Figure 7A** shows the recorded intersection angle of a 2 μm microtubule changing with time in MD simulation. It can be seen that the angle fluctuates between 0.0 and 17.0 degrees and the angles almost exhibit a Gaussian distribution in the inset. The data in **Figure 7A** suggest that our MD samplings are sufficient to evaluate persistence length.

Based on the sampled structures of microtubules with different lengths, we calculated the persistence lengths corresponding to different lengths according to **Eq. 3**. **Figure 7B** shows the change of calculated persistence length l_p changing with the contour length, with the longest microtubule being 12 μm. The calculated flexural rigidity κ according to the formula $\kappa = l_p k_B T$ (Pampaloni et al., 2006) is plotted at the right side of y-axis, where k_B is the Boltzmann constant and T is the simulation temperature. The simulated data, denoted by the yellow squares, clearly verify that the persistence length increases nonlinearly with the length. During the simulated range of length from 0.5 to 12 μm, the variation tendency of l_p is in a good agreement with the experimental measurements (Pampaloni et al., 2006). In order to evaluate the magnitude of the persistence length l_p^{∞}, we fitted the data of calculated l_p with the microtubule length L according to **Eq. 4** derived by Pampaloni et al. (Pampaloni et al., 2006). As shown in **Eq. 4** below, the l_p has a complicated functional dependence on the l_p^{∞} at infinity:

$$l_p = l_p^{\infty}\left(1 + \frac{3EI}{GkAL^2}\right)^{-1} \tag{4}$$

where E is the longitude Young's modulus and adopts the experimental value of 100 MPa (Kis et al., 2002) here, A is the area of cross section with 25 nm^2 for a microtubule composed of 13 protofilaments, I is the area moment of the cross section with 16670 nm^4, k is a geometric factor adopting 0.72, and G is the transverse sheer modulus. **Figure 7B** shows the fitted blue curve based on **Eq. 4**, similar to the fitted one in the previous work (Pampaloni et al., 2006). The persistence length l_p^{∞} derived from our fitting is 7.0 mm, consistent with the experimental value of 6.3 ± 0.8 mm for long microtubules. They also calculated the term of $(3EI/GkA)^{0.5}$ to be 21 μm (Pampaloni et al., 2006), which was considered as a threshold between short and long microtubules. We further compare our simulation results with the experimental data denoted by the hollow green circles below 12 μm (Pampaloni et al., 2006). It appears that our simulation results fit quiet well with the experimental measurements within the contour length of 10 μm. Besides, in the experimental measurement of microtubules up to 47.5 μm, the persistence length l_p seems to approach a constant beyond 21 μm. At present, our data obtained from UCG MD simulation based on the microtubule models with the lengths of 0.5–12 μm generated good results for l_p and l_p^{∞} agreeable with known experimental results. In addition, the calculated values of the flexural rigidity κ for the length of 2–12 μm are in the range of 0.89–11.5 × 10^{-24} Nm^2, falling in the range of the reported experimental values 3.7–40.0 × 10^{-24} Nm^2 (Gittes et al., 1993; Venier et al., 1994; Kurachi et al., 1995; Felgner et al., 1996). However, a simulation of microtubules beyond 21 μm still faces a severe challenge memory requirement for calculating the interactions between pairwise particles in the UCG models.

CONCLUSION

In this work, we introduce a systematic strategy to build a UCG model for microtubule with micrometer length from the cryo-EM density data. For the large microtubule with micrometer length, the CK-CG serves as an efficient coarse-graining method to derive its UCG representation. The comparison of time cost in coarse-graining the microtubules with the lengths from 1 to 10 μm demonstrates that CK-CG is superior to the conventional K-means in coarse-graining large systems. A reasonable minimal model was then determined for modeling the mechanical properties of microtubules. By means of using the MVP-ANM force fields, we parameterized the UCG beads and constructed the corresponding models for the study of mechanical properties.

We estimated the persistence length of microtubules with different lengths by using the force-clamp MD simulation based on thermal fluctuation. The obtained relationship between the persistence length and contour length accords well with the previous theoretical and experimental results, which demonstrates the validity of our method in modeling the mechanical properties of biomaterials. It is highlighted that our method has the advantage of being capable of constructing models directly from the low-resolution density data of cryo-EM, without any atomistic details required. Thus, it is particularly suitable for the functional study of macrobiomolecules for which only low-resolution cryo-EM data are available. Our model relies on low-resolution cryo-EM data, which is sufficient given our choice for an ultra-coarse-grained nature of our model. However, it remains to be examined if higher resolution cryo-EM structures can further improve models, even those on the very coarse scale such as the UCG model presented here. However, the current UCG model is still a sort of ENMs and it is limited to be valid in

the study of biomechanical properties of biomaterials with small deformations under external forces. Future work will be invested into developing more physically meaningful energy potentials for UCG model and using our method to investigate longer microtubules and eventually the mechanical properties of a bundle of microtubules such as those present in axons.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

FX designed the project. JZ performed the simulation and analyzed the data. All people discussed and wrote the paper together.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 21773065 and 22073029), Nanyang Technological University Startup Grant M4081842, Singapore Ministry of Education Academic Research fund Tier 1 RG31/18 and RG109/19, Tier 2 MOE2018-T2-1-033. FG acknowledges support by the Klaus Tschira Foundation and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–2082/1–390761711.

ACKNOWLEDGMENTS

The authors acknowledge the support of the NYU-ECNU Center for Computational Chemistry at NYU Shanghai. They also thank the ECNU Public Platform for Innovation (001) for providing computer time.

REFERENCES

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J.* 80 (1), 505–515. doi:10.1016/S0006-3495(01)76033-X

Bahar, I., Atilgan, A. R., and Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des.* 2 (3), 173–181. doi:10.1016/s1359-0278(97)00024-2

Conde, C., and Cáceres, A. (2009). Microtubule assembly, organization and dynamics in axons and dendrites. *Nat Rev Neurosci.* 10 (5), 319–332. doi:10.1038/nrn2631

Cross, R. A. (2019). Microtubule lattice plasticity. *Curr Opin Cell Biol.* 56, 88–93. doi:10.1016/j.ceb.2018.10.004

Dama, J. F., Jin, J., and Voth, G. A. (2017). The theory of ultra-coarse-graining. 3. Coarse-grained sites with rapid local equilibrium of internal states. *J Chem Theory Comput.* 13 (3), 1010–1022. doi:10.1021/acs.jctc.6b01081

Dama, J. F., Sinitskiy, A. V., McCullagh, M., Weare, J., Roux, B., Dinner, A. R., et al. (2013). The theory of ultra-coarse-graining. 1. General principles. *J Chem Theory Comput.* 9 (5), 2466–2480. doi:10.1021/ct4000444

Davtyan, A., Dama, J. F., Sinitskiy, A. V., and Voth, G. A. (2014). The theory of ultra-coarse-graining. 2. Numerical implementation. *J Chem Theory Comput.* 10 (12), 5265–5275. doi:10.1021/ct500834t

de Pablo, P. J., Schaap, I. A., MacKintosh, F. C., and Schmidt, C. F. (2003). Deformation and collapse of microtubules on the nanometer scale. *Phys Rev Lett.* 91 (9), 098101–098104. doi:10.1103/PhysRevLett.91.098101

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 (21), 8864–8872. doi:10.1007/s10853-007-1784-6

Deriu, M.A., Soncin, M., and Redaelli, A. (2008). Mechanical model of the tubulin dimer based on molecular dynamics simulations. *J Biomech Eng.* 130 (4), 41008–41014. doi:10.1115/1.2913330

Deriu, M. A., Soncini, M., Orsi, M., Patel, M., Essex, J. W., Montevecchi, F. M., et al. (2010). Anisotropic elastic network modeling of entire microtubules. *Biophys J.* 99 (7), 2190–2199. doi:10.1016/j.bpj.2010.06.070

Dima, R. I., and Joshi, H. (2008). Probing the origin of tubulin rigidity with molecular simulations. *Proc Natl Acad Sci USA.* 105 (41), 15743–15748. doi:10.1073/pnas.0806113105

Ding, Y., and Xu, Z. (2011). Mechanics of microtubules from a coarse-grained model. *BioNanoSci.* 1 (4), 173–182. doi:10.1007/s12668-011-0027-0

Dye, R. B., Fink, S. P., and Williams, J. (1993). Taxol-induced flexibility of microtubules and its reversal by MAP-2 and Tau. *J Biol Chem.* 268 (10), 6847–6850. doi:10.1016/s0021-9258(18)53113-6

Esquivel-Rodríguez, J., and Kihara, D. (2013). Computational methods for constructing protein structure models from 3D electron microscopy maps. *J Struct Biol.* 184 (1), 93–102. doi:10.1016/j.jsb.2013.06.008

Felgner, H., Frank, R., and Schliwa, M. (1996). Flexural rigidity of microtubules measured with the use of optical tweezers. *J Cell Sci.* 109 (2), 509–516. doi:10.1083/jcb.132.4.741

Franz, F., Daday, C., and Gräter, F. (2020). Advances in molecular simulations of protein mechanical properties and function. *Curr Opin Struct Biol.* 61, 132–138. doi:10.1016/j.sbi.2019.12.015

Fratzl, P., and Weinkamer, R. (2007). Nature's hierarchical materials. *Progress in Materials Science.* 52 (8), 1263–1334. doi:10.1016/j.pmatsci.2007.06.001

Gittes, F., Meyhöfer, E., Baek, S., and Howard, J. (1996). Directional loading of the kinesin motor molecule as it buckles a microtubule. *Biophys J.* 70 (1), 418–429. doi:10.1016/S0006-3495(96)79585-1

Gittes, F., Mickey, B., Nettleton, J., and Howard, J. (1993). Flexural rigidity of microtubules and actin filaments measured from thermal fluctuations in shape. *J Cell Biol.* 120 (4), 923–934. doi:10.1083/jcb.120.4.923

Grafmüller, A., Noya, E. G., and Voth, G. A. (2013). Nucleotide-dependent lateral and longitudinal interactions in microtubules. *J Mol Biol.* 425 (12), 2232–2246. doi:10.1016/j.jmb.2013.03.029

Grafmüller, A., and Voth, G. A. (2011). Intrinsic bending of microtubule protofilaments. *Structure.* 19 (3), 409–417. doi:10.1016/j.str.2010.12.020

Grange, M., Vasishtan, D., and Grünewald, K. (2017). Cellular electron cryo tomography and in situ sub-volume averaging reveal the context of microtubule-based processes. *J Struct Biol.* 197 (2), 181–190. doi:10.1016/j.jsb.2016.06.024

Gräter, F., and Grubmüller, H. (2007). Fluctuations of primary ubiquitin folding intermediates in a force clamp. *Journal of Structural Biology.* 157 (3), 557–569. doi:10.1016/j.jsb.2006.11.005

Grime, J. M., Dama, J. F., Ganser-Pornillos, B. K., Woodward, C. L., Jensen, G. J., Yeager, M., et al. (2016). Coarse-grained simulation reveals key features of HIV-1 capsid self-assembly. *Nat Commun.* 7, 11568–11578. doi:10.1038/ncomms11568

Grønbech-Jensen, N. (2019). Complete set of stochastic verlet-type thermostats for correct langevin simulations. *Molecular Physics.* 118 (8), e1662506–e1662525. doi:10.1080/00268976.2019.1662506

Havelka, D., Deriu, M. A., Cifra, M., and Kučera, O. (2017). Deformation pattern in vibrating microtubule: structural mechanics study based on an atomistic approach. *Sci Rep.* 7 (1), 4227–4237. doi:10.1038/s41598-017-04272-w

Hawkins, T. L., Sept, D., Mogessie, B., Straube, A., and Ross, J. L. (2013). Mechanical properties of doubly stabilized microtubule filaments. *Biophys J.* 104 (7), 1517–1528. doi:10.1016/j.bpj.2013.02.026

Heuvel, M. G. L. v. d., Bolhuis, S., and Dekker, C. (2007). Persistence length measurements from stochastic single-microtubule trajectories. *Nano Lett*. 7 (10), 3138–3144. doi:10.1021/nl071696y

Isralewitz, B., Gao, M., and Schulten, K. (2001). Steered molecular dynamics and mechanical functions of proteins. *Curr Opin Struct Biol*. 11 (2), 224–2230. doi:10.1016/S0959-440X(00)00194-9

Janson, M. E., and Dogterom, M. (2004). A bending mode analysis for growing microtubules: evidence for a velocity-dependent rigidity. *Biophys J*. 87 (4), 2723–2736. doi:10.1529/biophysj.103.038877

Ji, X. Y., and Feng, X. Q. (2011a). Coarse-grained mechanochemical model for simulating the dynamic behavior of microtubules. *Phys Rev E Stat Nonlin Soft Matter Phys*. 84 (3), 031933–031941. doi:10.1103/PhysRevE.84.031933

Ji, X. Y., and Feng, X. Q. (2011b). Mechanochemical modeling of dynamic microtubule growth involving sheet-to-tube transition. *PLoS One*. 6 (12), e29049–e29059. doi:10.1371/journal.pone.0029049

Jin, J., Han, Y., and Voth, G. A. (2018). Ultra-coarse-grained liquid state models with implicit hydrogen bonding. *J Chem Theory Comput*. 14 (12), 6159–6174. doi:10.1021/acs.jctc.8b00812

Jin, J., and Voth, G. A. (2018). Ultra-coarse-grained models allow for an accurate and transferable treatment of interfacial systems. *J Chem Theory Comput*. 14 (4), 2180–2197. doi:10.1021/acs.jctc.7b01173

Kasas, S., Cibert, C., Kis, A., De Los Rios, P., Riederer, B. M., Forró, L., et al. (2004a). Oscillation modes of microtubules. *Biol Cell*. 96 (9), 697–700. doi:10.1016/j.biolcel.2004.09.002

Kasas, S., Kis, A., Riederer, B. M., Forró, L., Dietler, G., and Catsicas, S. (2004b). Mechanical properties of microtubules explored using the finite elements method. *Chemphyschem*. 5 (2), 252–257. doi:10.1002/cphc.200300799

Katkar, H. H., Davtyan, A., Durumeric, A. E. P., Hocky, G. M., Schramm, A. C., De La Cruz, E. M., et al. (2018). Insights into the cooperative nature of ATP hydrolysis in actin filaments. *Biophys J*. 115 (8), 1589–1602. doi:10.1016/j.bpj.2018.08.034

Kis, A., Kasas, S., Babić, B., Kulik, A J., Benoît, W., Briggs, G. A., et al. (2002). Nanomechanics of microtubules. *Phys Rev Lett*. 89 (24), 248101–248104. doi:10.1103/PhysRevLett.89.248101

Kis, A., Kasas, S., Kulik, A. J., Catsicas, S., and Forró, L. (2008). Temperature-dependent elasticity of microtubules. *Langmuir*. 24 (12), 6176–6181. doi:10.1021/la800438q

Kollman, J. M., Polka, J. K., Zelter, A., Davis, T. N., and Agard, D. A. (2010). Microtubule nucleating gamma-TuSC assembles structures with 13-fold microtubule-like symmetry. *Nature*. 466 (7308), 879–882. doi:10.1038/nature09207

Kononova, O., Kholodov, Y., Theisen, K. E., Marx, K. A., Dima, R. I., Ataullakhanov, F. I., et al. (2014). Tubulin bond energies and microtubule biomechanics determined from nanoindentation in silico. *J Am Chem Soc*. 136 (49), 17036–17045. doi:10.1021/ja506385p

Kroy, K., and Frey, E. (1996). Force-extension relation and plateau modulus for wormlike chains. *Phys Rev Lett*. 77 (2), 306–309. doi:10.1103/PhysRevLett.77.306

Kurachi, M., Hoshi, M., and Tashiro, H. (1995). Buckling of a single microtubule by optical trapping forces: direct measurement of microtubule rigidity. *Cell Motil Cytoskeleton*. 30 (3), 221–228. doi:10.1002/cm.970300306

Lazarus, C., Soheilypour, M., and Mofrad, M. R. (2015). Torsional behavior of axonal microtubule bundles. *Biophys J*. 109 (2), 231–239. doi:10.1016/j.bpj.2015.06.029

Le Goff, L., Hallatschek, O., Frey, E., and Amblard, F. (2002). Tracer studies on f-actin fluctuations. *Phys Rev Lett*. 89 (25), 258101–258104. doi:10.1103/PhysRevLett.89.258101

Li, M., Zhang, J. Z., and Xia, F. (2016a). Constructing optimal coarse-grained sites of huge biomolecules by fluctuation maximization. *J Chem Theory Comput*. 12 (4), 2091–2100. doi:10.1021/acs.jctc.6b00016

Li, M., Zhang, J. Z., and Xia, F. (2016b). A new algorithm for construction of coarse-grained sites of large biomolecules. *J Comput Chem*. 37 (9), 795–804. doi:10.1002/jcc.24265

Li, Z., Cai, X., Liu, Y., and Zhu, B. (2019). A novel gaussian-bernoulli based convolutional deep belief networks for image feature extraction. *Neural Process Lett*. 49 (4), 305–319. doi:10.1007/s11063-017-9751-y

Ming, D., Kong, Y., Lambert, M. A., Huang, Z., and Ma, J. (2002). How to describe protein motion without amino acid sequence and atomic coordinates. *Proc Natl Acad Sci USA*. 99 (13), 8620–8625. doi:10.1073/pnas.082148899

Molodtsov, M. I., Ermakova, E. A., Shnol, E. E., Grishchuk, E. L., McIntosh, J. R., and Ataullakhanov, F. I. (2005). A Molecular-Mechanical Model of the Microtubule. *Biophys J*. 88 (5), 3167–3179. doi:10.1529/biophysj.104.051789

Opron, K., Xia, K., and Wei, G. W. (2014). Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis. *J Chem Phys*. 140 (23), 234105–234123. doi:10.1063/1.4882258

Pampaloni, F., Lattanzi, G., Jonás, A., Surrey, T., Frey, E., and Florin, E. L. (2006). Thermal fluctuations of grafted microtubules provide evidence of a length-dependent persistence length. *Proc Natl Acad Sci USA*. 103 (27), 10248–10253. doi:10.1073/pnas.0603931103

Peter, S. J., and Mofrad, M. R. (2012). Computational modeling of axonal microtubule bundles under tension. *Biophys J*. 102 (4), 749–757. doi:10.1016/j.bpj.2011.11.4024

Plimpton, S. (1995). Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*. 117 (1), 1–19. doi:10.1006/jcph.1995.1039

Poma, A. B., Chwastyk, M., and Cieplak, M. (2017). Elastic moduli of biological fibers in a coarse-grained model: crystalline cellulose and β-amyloids. *Phys Chem Chem Phys*. 19 (41), 28195–28206. doi:10.1039/c7cp05269c

Poma, A. B., Li, M. S., and Theodorakis, P. E. (2018). Generalization of the elastic network model for the study of large conformational changes in biomolecules. *Phys Chem Chem Phys*. 20 (25), 17020–17028. doi:10.1039/c8cp03086c

Schaap, I. A., Carrasco, C., de Pablo, P. J., MacKintosh, F. C., and Schmidt, C. F. (2006). Elastic Response, Buckling, and Instability of Microtubules under Radial Indentation. *Biophys J*. 91 (4), 1521–1531. doi:10.1529/biophysj.105.077826

Sept, D., and MacKintosh, F. C. (2010). Microtubule elasticity: connecting all-atom simulations with continuum mechanics. *Phys Rev Lett*. 104 (1), 018101–018104. doi:10.1103/PhysRevLett.104.018101

Sinitskiy, A. V., Saunders, M. G., and Voth, G. A. (2012). Optimal number of coarse-grained sites in different components of large biomolecular complexes. *J Phys Chem B*. 116 (29), 8363–8374. doi:10.1021/jp2108895

Soheilypour, M., Peyro, M., Peter, S. J., and Mofrad, M. R. (2015). Buckling behavior of individual and bundled microtubules. *Biophys J*. 108 (7), 1718–1726. doi:10.1016/j.bpj.2015.01.030

Stevens, M. J. (2017). The long persistence length of model tubules. *J Chem Phys*. 147 (4), 044902–044907. doi:10.1063/1.4994913

Szatkowski, L., Merz, D. R., Jr., Jiang, N., Ejikeme, I., Belonogov, L., Ross, J. L., et al. (2019). Mechanics of the microtubule seam interface probed by molecular simulations and in vitro severing experiments. *J Phys Chem B*. 123 (23), 4888–4900. doi:10.1021/acs.jpcb.9b03059

Taute, K. M., Pampaloni, F., Frey, E., and Florin, E. L. (2008). Microtubule dynamics depart from the wormlike chain model. *Phys Rev Lett*. 100 (2), 028102–028105. doi:10.1103/PhysRevLett.100.028102

Theisen, K. E., Desai, N. J., Volski, A. M., and Dima, R. I. (2013). Mechanics of severing for large microtubule complexes revealed by coarse-grained simulations. *J Chem Phys*. 139 (12), 121926–121937. doi:10.1063/1.4819817

Tong, D., and Voth, G. A. (2020). Microtubule simulations provide insight into the molecular mechanism underlying dynamic instability. *Biophys J*. 118 (12), 2938–2951. doi:10.1016/j.bpj.2020.04.028

VanBuren, V., Cassimeris, L., and Odde, D. J. (2005). Mechanochemical model of microtubule structure and self-assembly kinetics. *Biophys J*. 89 (5), 2911–2926. doi:10.1529/biophysj.105.060913

VanBuren, V., Odde, D. J., and Cassimeris, L. (2002). Estimates of lateral and longitudinal bond energies within the microtubule lattice. *Proc Natl Acad Sci USA*. 99 (8), 6035–6040. doi:10.1073/pnas.040639310110.1073/pnas.092504999

Venier, P., Maggs, A. C., Carlier, M. F., and Pantaloni, D. (1994). Analysis of microtubule rigidity using hydrodynamic flow and thermal fluctuations. *J Biol Chem*. 269 (18), 13353–13360. doi:10.1016/0009-3084(94)02321-210.1016/s0021-9258(17)36840-0

Vinckier, A. (1996). Dynamical and mechanical study of immobilized microtubules with atomic force microscopy. *J. Vac. Sci. Technol. B*. 14 (2), 1427–1431. doi:10.1116/1.589113

Wagner, O., Zinke, J., Dancker, P., Grill, W., and Bereiter-Hahn, J. (1999). Viscoelastic properties of f-actin, microtubules, f-actin/alpha-actinin, and f-actin/hexokinase determined in microliter volumes with a novel nondestructive method. *Biophys J*. 76 (5), 2784–2796. doi:10.1016/S0006-3495(99)77432-1

Wells, D. B., and Aksimentiev, A. (2010). Mechanical properties of a complete microtubule revealed through molecular dynamics simulation. *Biophys J.* 99 (2), 629–637. doi:10.1016/j.bpj.2010.04.038

Wu, Y. T., and Adnan, A. (2018). Damage and failure of axonal microtubule under extreme high strain rate: an in-silico molecular dynamics study. *Sci Rep.* 8 (1), 12260. doi:10.1038/s41598-018-29804-w

Wu, Z., Zhang, Y., Zhang, J. Z., Xia, K., and Xia, F. (2020). Determining optimal coarse-grained representation for biomolecules using internal cluster validation indexes. *J Comput Chem.* 41 (1), 14–20. doi:10.1002/jcc.26070

Xia, F., and Lu, L. (2012). Multiscale coarse-graining via normal mode analysis. *J Chem Theory Comput.* 8 (11), 4797–4806. doi:10.1021/ct3006387

Xia, F., Tong, D., and Lu, L. (2013a). Robust heterogeneous anisotropic elastic network model precisely reproduces the experimental b-factors of biomolecules. *J Chem Theory Comput.* 9 (8), 3704–3714. doi:10.1021/ct4002575

Xia, K., Opron, K., and Wei, G. W. (2013b). Multiscale multiphysics and multidomain models--flexibility and rigidity. *J Chem Phys.* 139 (19), 194109–194124. doi:10.1063/1.4830404

Xia, K. (2017). Multiscale virtual particle based elastic network model (MVP-ENM) for normal mode analysis of large-sized biomolecules. *Phys Chem Chem Phys.* 20 (1), 658–669. doi:10.1039/c7cp07177a

Zapp, C., Obarska-Kosinska, A., Rennekamp, B., Kurth, M., Hudson, D. M., Mercadante, D., et al. (2020). Mechanoradicals in tensed tendon collagen as a source of oxidative stress. *Nat Commun.* 11 (1), 2315–2322. doi:10.1038/s41467-020-15567-4

Zhang, Z., Lu, L., Noid, W. G., Krishna, V., Pfaendtner, J., and Voth, G. A. (2008). A systematic methodology for defining coarse-grained sites in large biomolecules. *Biophys J.* 95 (11), 5073–5083. doi:10.1529/biophysj.108.139626

Zhang, Z., Pfaendtner, J., Grafmüller, A., and Voth, G. A. (2009). Defining coarse-grained representations of large biomolecules and biomolecular complexes from elastic network models. *Biophys J.* 97 (8), 2327–2337. doi:10.1016/j.bpj.2009.08.007

Zhang, Y., Cao, Z., and Xia, F. (2017a). Construction of ultra-coarse-grained model of protein with a Gō-like potential. *Chemical Physics Letters.* 681, 1–6. doi:10.1016/j.cplett.2017.05.039

Zhang, Y., Cao, Z., Zhang, J. Z., and Xia, F. (2017b). Performance comparison of systematic methods for rigorous definition of coarse-grained sites of large biomolecules. *J Chem Inf Model.* 57 (2), 214–222. doi:10.1021/acs.jcim.6b00683

Zhang, Y., Cao, Z., Zhang, J. Z., and Xia, F. (2020). Double-well ultra-coarse-grained model to describe protein conformational transitions. *J Chem Theory Comput.* 16 (10), 6678–6689. doi:10.1021/acs.jctc.0c00551

Zhang, Y., Xia, K., Cao, Z., Gräter, F., and Xia, F. (2019). A new method for the construction of coarse-grained models of large biomolecules from low-resolution cryo-electron microscopy data. *Phys Chem Chem Phys.* 21 (19), 9720–9727. doi:10.1039/c9cp01370a

Zhang, Z., and Voth, G. A. (2010). Coarse-grained representations of large biomolecular complexes from low-resolution structural data. *J Chem Theory Comput.* 6 (9), 2990–3002. doi:10.1021/ct100374a

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 © 2021 Zha, Zhang, Xia, Gräter and Xia. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 18 February 2021
doi: 10.3389/fmolb.2021.633130

Mutations of N1 Riboswitch Affect its Dynamics and Recognition by Neomycin Through Conformational Selection

Piotr Chyży[1,2], Marta Kulik[1,3]†, Suyong Re[4,5], Yuji Sugita[4,6,7]† and Joanna Trylska[1]†*

[1]Centre of New Technologies, University of Warsaw, Warsaw, Poland, [2]Faculty of Physics, University of Warsaw, Warsaw, Poland, [3]Department of Chemistry, Biological and Chemical Research Centre, University of Warsaw, Warsaw, Poland, [4]RIKEN Center for Biosystems Dynamics Research (BDR), Kobe, Japan, [5]National Institutes of Biomedical Innovation, Health and Nutrition, Ibaraki, Japan, [6]RIKEN Cluster for Pioneering Research (CPR), Wako, Japan, [7]RIKEN Center for Computational Science, Kobe, Japan

OPEN ACCESS

Edited by:
Vojtech Spiwok,
University of Chemistry and Technology in Prague, Czechia

Reviewed by:
Susmita Roy,
Indian Institute of Science Education and Research Kolkata, India
Julia Weigand,
Technical University of Darmstadt, Germany
Sandro Bottaro,
University of Copenhagen, Denmark
Jiri Sponer,
Academy of Sciences of the Czech Republic, Czechia

***Correspondence:**
Marta Kulik
m.kulik@uw.edu.pl

†ORCID:
Marta Kulik
orcid.org/0000-0003-2381-7665
Yuji Sugita
orcid.org/0000-0001-9738-9216
Joanna Trylska
orcid.org/0000-0002-1464-5323

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: 24 November 2020
Accepted: 07 January 2021
Published: 18 February 2021

Citation:
Chyży P, Kulik M, Re S, Sugita Y and Trylska J (2021) Mutations of N1 Riboswitch Affect its Dynamics and Recognition by Neomycin Through Conformational Selection. Front. Mol. Biosci. 8:633130. doi: 10.3389/fmolb.2021.633130

Short, structured fragments of non-coding mRNA may act as molecular switches upon binding specific ligands, regulating the translation of proteins encoded downstream this mRNA sequence. One switch, called riboswitch N1, is regulated by aminoglycosides such as neomycin. Nucleobase mutations in the apical loop, although distant from the binding pocket, significantly affect neomycin affinity and riboswitch regulatory efficiency. To explain this influence, we conducted molecular dynamics simulations using generalized replica exchange with solute tempering (gREST). Translation assay of a reporter protein in a yeast system shows that mutating A17 to G in the riboswitch apical loop reduces 6-fold the translation regulation efficiency of the mutant. Indeed, simulations of the unbound riboswitch show that G17 frequently stacks with base 7, while base 8 is stabilized towards the binding site in a way that it may interfere with the conformational selection mechanism and decrease riboswitch regulatory activity. In the riboswitch complexes, this single-point A to G mutation disrupts a strong hydrogen bond between nucleotides 5 and 17 and, instead, a new hydrogen bond between residue 17 and neomycin is created. This change forces neomycin to occupy a slightly shifted position in the binding pocket, which increases neomycin flexibility. Our simulations of the U14C mutation suggest that the riboswitch complex with neomycin is more stable if cytosine 14 is protonated. A hydrogen bond between the RNA phosphate and protonated cytosine appears as the stabilizing factor. Also, based on the cell-free translation assay and isothermal titration calorimetry experiments, mutations of nucleotides 14 and 15 affect only slightly the riboswitch ability to bind the ligand and its activity. Indeed, the simulation of the unbound U15A mutant suggests conformations preformed for ligand binding, which may explain slightly higher regulatory activity of this mutant. Overall, our results corroborate the *in vivo* and *in vitro* experiments on the N1 riboswitch-neomycin system, detail the relationship between nucleobase mutations and RNA dynamics, and reveal the conformations playing the major role in the conformational selection mechanism.

Keywords: RNA, riboswitch, aminoglycosides, neomycin, molecular dynamics simulations, replica exchange with solute tempering

1 INTRODUCTION

Riboswitches are regulatory RNA elements, typically found in 5′-untranslated regions (5′-UTR) of mRNA and positioned upstream of the regulated coding sequence. They control gene expression by directly binding ions and metabolites (McCown et al., 2017) or responding to changes in pH or temperature (Bastet et al., 2011). Since additional protein factors are not necessary for their activity, riboswitches are a valuable tool for gene regulation in synthetic biology (Serganov and Nudler, 2013; Berens et al., 2015; Sinumvayo et al., 2018). For example, two riboswitches can be engineered to form a NOR logic gate (Boussebayle et al., 2019).

Riboswitches that bind ligands typically contain a ligand binding sensory domain (aptamer) and expression platform. Interactions between the aptamer and ligand determine the conformation of the expression platform. Structural rearrangements of this platform upon ligand binding influence the expression of genes located downstream the same mRNA (Berens et al., 2015). Regulation of gene expression or repression may occur at different levels: transcription, translation, splicing or mRNA cleavage.

Aptamers are single stranded oligonucleotides that selectively bind small ligands. They have been designed even before the discovery of the first natural riboswitch. Aptamers binding a specific ligand are usually isolated by an *in vitro* evolution and selection process (SELEX) (Ellington and Szostak, 1990) involving screening large libraries of nucleic acid oligomers (Banerjee and Nilsen-Hamilton, 2013). Since aptamers are the regulatory parts of riboswitches, they could be designed to constitute synthetic riboswitches. The activities of such synthetic riboswitches can be tested, e.g., by introducing the aptamer to the 5′-UTR of mRNA in a cell or cell-free system producing a reporter protein. Then, the regulatory mechanism of the riboswitch is due to mechanical blocking of scanning the mRNA by the small ribosomal subunit (Etzel and Mörl, 2017).

A synthetic aminoglycoside-sensing N1 riboswitch (Weigand et al., 2008) was also discovered through screening of aptamers that bind aminoglycoside antibiotics. The N1 riboswitch is a 27-nucleotide-long RNA and is the smallest synthetic riboswitch found active *in vivo* (Gottstein-Schmidtke et al., 2014). Its activity is induced by binding neomycin and was confirmed both *in vitro* and *in vivo* (Weigand et al., 2008). This riboswitch forms a hairpin with two flexible functional regions: a bulge and apical loop (**Figure 1**). Apart from neomycin, N1 riboswitch binds also paromomycin, tobramycin and ribostamycin. The N1 structure in the complex with ribostamycin (Duchardt-Ferner et al., 2010) and paromomycin (Duchardt-Ferner et al., 2016) was determined by NMR spectroscopy.

The N1 riboswitch is active in yeast if inserted into the 5′-UTR of an mRNA of a reporter gene (Weigand et al., 2008) expressing green fluorescent protein. Fluorescence measurements showed that the expression of green fluorescent protein in yeast depends on the presence of neomycin that supposedly binds the N1 riboswitch inserted into 5′-UTR. Furthermore, the N1 riboswitch works also for ribostamycin though with slightly lower regulatory activity (Weigand et al., 2008). Surprisingly, paromomycin does not inhibit gene expression although it differs from neomycin by only one chemical group, namely, amino versus hydroxyl at 6′ position of ring I (**Supplementary Figure S1**). Using molecular dynamics simulations with replica exchange (Kulik et al., 2018), we have previously proposed the reasons for different activities of these ligands by comparing the dynamics of their complexes. As shown in

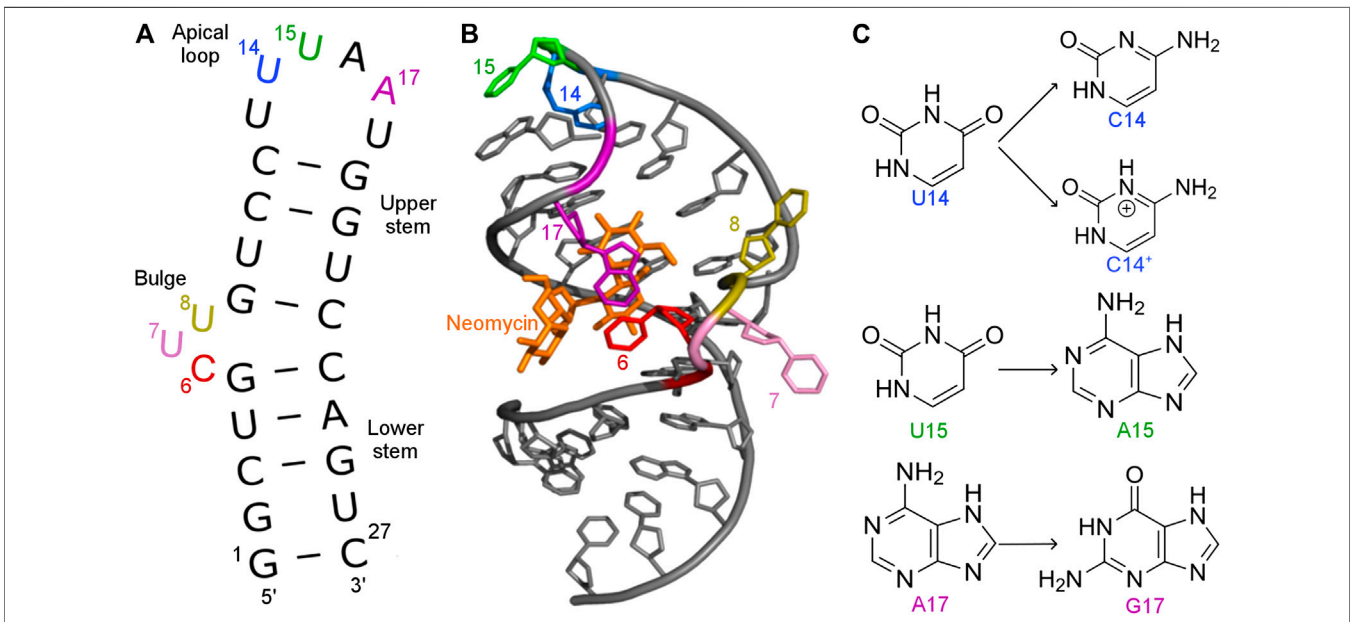


FIGURE 1 | (A) Secondary and **(B)** tertiary structure of the N1 riboswitch in the complex with ribostamycin (PDB ID: 2KXM (Duchardt-Ferner et al., 2010)). **(C)** Mutations of nucleobases 14, 15, and 17 of the N1 riboswitch.

TABLE 1 | Dissociation constants (K_d) determined from ITC experiments and regulatory factors (R.f.) calculated as the ratio of relative fluorescence of GFP expression in the absence and presence of neomycin (Weigand et al., 2014; Duchardt-Ferner et al., 2016). The smaller the regulatory factor, the lower the activity of the riboswitch.

	Ribostamycin		Neomycin		Paromomycin	
	K_d [nM]	R.f.	K_d [nM]	R.f.	K_d [nM]	R.f
N1	330 ± 30	2.3	9.2 ± 1.3	8.6	5130 ± 26	1.0
U14C	–	–	24.2 ± 1.5	3.0	–	–
U15A	–	–	11.0± 0.3	10.0	–	–
A17G	–	–	130.9 ± 16.8	1.5	–	–
A17C	1530 ± 140	1.1	–	2.9	11890 ± 1550	1.0

Table 1, dissociation constants for neomycin, ribostamycin and paromomycin to the isolated riboswitch determined using isothermal titration calorimetry (ITC) (Duchardt-Ferner et al., 2016; Weigand et al., 2014) agree with the translation blocking efficiencies obtained *in vivo*.

Stopped-flow fluorescence assays of the N1 riboswitch labeled with cytidine analog at positions 6 and 8 showed that neomycin binding kinetics is fast and proceeds according to the two-step binding model, in which the initial non-specific ligand binding is followed by a specific binding with minor conformational changes. Also, the study identified the conformational selection mechanism as dominating in the neomycin binding (Gustmann et al., 2019).

Several mutations were introduced in the isolated N1 riboswitch to elucidate the importance of the aptamer sequence for aminoglycoside binding. Mutations of C6, U7, and U8 nucleotides revealed that the sequence of the bulge is crucial for the riboswitch activity (Weigand et al., 2011). Various mutations in the apical loop also hinder riboswitch activity, apart from U15A (Weigand et al., 2014). Mutational analysis of several minimal neomycin aptamers similar to N1 have shown that the formation of the upper stem, between the bulge and apical loop (**Figure 1**), is critical for regulatory properties of the riboswitch, while the sequence of the lower stem may be gently modified without the loss of function (Weigand et al., 2008). Apart from activity studies, also dissociation constants show that N1 mutations affect ligand binding (**Table 1**). For example, the dissociation constant of neomycin - U14C mutant is two-fold smaller than of neomycin - N1 riboswitch, although regulatory activity of the U14C mutant in yeast is preserved (Weigand et al., 2014) (the relative expression of green fluorescent protein with and without neomycin is similar as for the non-mutated N1 riboswitch). The experimental K_d of neomycin for the U15A mutant is close to the reference N1 riboswitch, but the measured regulatory activity of the U15A mutant in yeast is even better. Finally, the A17G change drastically increases K_d and decreases the riboswitch regulatory activity.

To elucidate the link between the nucleobase mutations, riboswitch dynamics and ligand binding, we performed generalized replica exchange with solute tempering (gREST) simulations of selected riboswitch mutants and their complexes with neomycin. Specifically, we explored the dynamics of the A17G mutant, which affects the riboswitch regulatory activity and K_d to the largest extent. For the U14C mutant, we investigated why in the neomycin complex the protonated cytosine state is preferred over the deprotonated state (Weigand et al., 2014). Lastly, we determined the conformational ensemble of N1 and the U15A mutant in unbound forms to verify possible enrichment of riboswitch states preformed for neomycin binding, as suggested by experiments (Weigand et al., 2014).

2 MATERIALS AND METHODS

2.1 Enhanced sampling methods

Parallel simulations of system copies at different temperatures in replica-exchange molecular dynamics (REMD) (Sugita and Okamoto, 1999) increase the sampling of conformations, which is crucial for elucidating the dynamics of flexible RNA systems. In our previous work, we applied the REMD method with 32 replicas to N1 riboswitch with different ligands and confirmed a significant gain in conformational sampling as compared to classical MD (Kulik et al., 2018). Here, to reduce the number of replicas without losing the sampling efficiency, we employ a recently developed method, called generalized replica exchange with solute tempering (gREST) (Kamiya and Sugita, 2018). This method stems from REST or REST2 (Wang et al., 2011; Liu et al., 2005; Moors et al., 2011; Terakawa et al., 2011), in which the acceptance probability for the replica exchange events does not depend on the number of explicit water molecules in the system. This is achieved by dividing the system into a pre-defined solute region, in which the temperature is exchanged between replicas, and the solvent region, kept at room temperature during the simulations. In gREST, the solute selection strategy is more flexible and the solute can include parts of molecules and selected potential energy function terms, such as the dihedral-angle term.

2.2 Structure preparation

The structure of N1 riboswitch in the complex with paromomycin was taken from the RCSB Protein Data Bank (PDB ID: 2MXS (Duchardt-Ferner et al., 2016)). Paromomycin was changed to neomycin by replacing 6′-OH with ammonium group (**Supplementary Figure S1**). This change does not influence the binding pose of the ligand as a similar binding pose is also adopted by ribostamycin in the NMR structure with PDB ID: 2N0J (Duchardt-Ferner et al., 2016). Paromomycin, ribostamycin and neomycin are similar, except for a different substituent at 6′ position of ring I (paromomycin: 6′-OH, ribostamycin: 6′-NH_3^+, neomycin: 6′-NH_3^+) and lack of ring IV in ribostamycin that is present in paromomycin and neomycin. Since paromomycin and ribostamycin exhibit similar binding mode in NMR-derived N1 structures, then neomycin, the combination of the former two, should have a similar mode. The mutant structures were prepared using Pymol (Schrödinger, LLC, 2010) by replacing appropriate nucleobases. Since the unbound riboswitch structures were not determined experimentally, we built them by deleting the ligands. All amino groups in neomycin were protonated (Kulik et al., 2018). Structure optimization and calculation of ESP charges of

neomycin were carried out in Gaussian 09 with HF/6-31G* level of theory (Gaussian Inc., 2009). Initial structures were solvated with a 20 Å shell of TIP3P water molecules around RNA (Jorgensen et al., 1983). 52 Na^+ and 26 Cl^- ions were added using the LEaP program from Amber14 (Case et al., 2014) to neutralize each unbound system and achieve 100 mM NaCl concentration, which is the salt concentration used in the UV-melting experiments of the N1 riboswitch (Weigand et al., 2011). For the systems with neomycin, 46 Na^+ and 26 Cl^- ions were added, while when cytosine was protonated, one more Cl^- ion was necessary. The number of added ions was calculated based on the number of water molecules in the system. The ion parameters were taken from (Joung and Cheatham, 2008). Even though the Mg^{2+} ions play an important role in the RNA structure-function (Hayes et al., 2012; Roy et al., 2017a, Roy et al., 2017b, Roy et al., 2019) and were present in the NMR experiments, their positions were not determined so they were not considered in the simulations. GAFF (Wang et al., 2004) parameters for neomycin were used together with Amber ff99 (Cornell et al., 1995; Wang et al., 2000) force field for RNA. The force field parameters for neomycin were validated previously in (Kulik et al., 2018). The parmbsc0 (Pérez et al., 2007) and χOL3 (Zgarbova et al., 2011) corrections were applied to the RNA force field (Šponer et al., 2018), previously used in the REMD simulations of RNA (Kulik et al., 2018; Bergonzo et al., 2014).

2.3 MD and gREST simulations

Energy minimizations, MD and gREST simulations were run with SPDYN from the GENESIS (v. 1.3 and 1.4) suite of programs (Jung et al., 2015; Kobayashi et al., 2017). The particle mesh Ewald method (Essmann et al., 1995) was used to calculate long-range electrostatic interactions. Lennard-Jones interactions were truncated at 12 Å cutoff distance and the pair list distance set to 13.5 Å. Water molecules were kept rigid with the SETTLE algorithm (Miyamoto and Kollman, 1992). Energy minimizations were carried out for 5000 steps with steepest descent algorithm with positional restraints of 10 kcal/mol/Å^2 on heavy atoms of RNA and aminoglycosides. The systems were equilibrated at 310.15 K for 3 ns in the NVT ensemble with positional restraints gradually decreasing every 500 ps, which was followed by additional 2 ns simulation in the NPT ensemble without restraints. The atomic coordinates were saved after the last round of equilibrations to use them as the starting point for the gREST procedure. Next, for each system, a 30 ns MD production simulation in the NPT ensemble was performed to calculate the average box size for gREST simulations. The 100-ns MD production simulation was done in the NVT ensemble. The previously saved atomic coordinates were parametrized in LEaP (Case et al., 2014), energy minimized with the new box size with 5.0 kcal/mol/Å^2 harmonic restraints on positions of heavy atoms of RNA and aminoglycosides, with cutoff distance of Lennard-Jones interactions at 10 Å and Verlet pair list distance at 11.5 Å. Subsequently, the restart files were used in gREST equilibration in the NVT ensemble for 150 ps with 5 kcal/mol/Å^2 positional restraints and in the NPT ensemble for 150 ps without restraints. Next, in the gREST production, lasting 300 ns, 8 replicas were used with exchange attempts every 2000 steps. The solute region was defined as RNA, aminoglycoside, counterions added to neutralize the system, and the dihedrals, Coulombic and Lennard-Jones potential energy terms. The temperatures of the solute region ranged from 310.15 K to 370.00 K, while the rest of the system was calculated at 310.15 K (referred to as 310 K hereafter). The RESPA integrator (Tuckerman et al., 1992) with 2.5 fs time step and SHAKE algorithm (Ryckaert et al., 1977) were applied to effectively describe bonds with hydrogens. In conventional MD simulations Langevin thermostat and barostat (Quigley and Probert, 2004) were used, but, in gREST, Bussi temperature and pressure control were applied (Bussi et al., 2007, Bussi et al., 2009). The production simulations are gathered in **Supplementary Table S1**. The level of convergence of the gREST simulations was verified by calculating the overlap between the covariant matrices of the final gREST trajectory fragments at 310 K (**Supplementary Table S2**). The overlap of the potential energy of replicas and the random walk in the temperature space in **Supplementary Figure S2** confirm that replicas visit the full range of temperatures. The acceptance ratio is at least 17% (**Supplementary Table S3**) and is sufficient for our simulations (Kamiya and Sugita, 2018).

2.4 Data analysis

For analysis, only the gREST simulation trajectories at 310 K were used, which is the temperature used in the ITC experiments (Weigand et al., 2014). The root mean square deviation (RMSD) was calculated using GENESIS analysis tool (rmsd_analysis) (Kobayashi et al., 2017) using the initial structure as a reference. RMSD for neomycin was calculated for all ligand atoms including hydrogens, after the least squares fitting of the RNA non-hydrogen atoms to the reference starting structure. The root mean square fluctuations (RMSF) around average atom positions, χ torsion angles (defined by O4′–C1′–N9–C4 atoms) and pseudo-dihedral angles were analyzed in Gromacs 2019.4 (Abraham et al., 2019). The pseudo-dihedral angles were defined according to Song *et al.* (Song et al., 2009) and were calculated to quantify the flipping of bases with respect to the RNA backbone. The k-means algorithm (Forgy, 1965) implemented in the GENESIS analysis tools (Kobayashi et al., 2017) was used for clustering with a 98% convergence level to obtain 10 clusters. Hydrogen bonds and stacking interactions between ligands and RNA were calculated with cpptraj (Roe and Cheatham, 2013) and MINT (Górska et al., 2015). Hydrogen bond criteria were: the maximum of 3.5 Å as the distance between non-hydrogen atoms and the minimum of 150° for the acceptor-hydrogen-donor angle. First 50 ns of each gREST simulation was omitted in the analysis, except for the RMSD calculations. If not stated otherwise, only heavy atoms of RNA and aminoglycosides were taken into account. Figures were generated with VMD (Humphrey et al., 1996) and Pymol (Schrödinger, LLC, 2010).

3 RESULTS AND DISCUSSION

3.1 Nucleotide substitutions in the apical loop affect the dynamics of the bulge

The RMSD plots as a function of the simulation time are shown in **Supplementary Figure S3**. As expected, bound neomycin

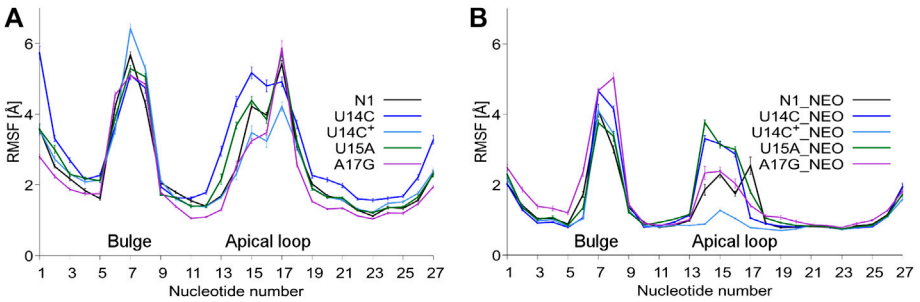


FIGURE 2 | RMSF values per nucleotide at 310 K in: **(A)** unbound and **(B)** neomycin-bound riboswitches. RMSF was calculated for all nucleotide atoms including hydrogen atoms. The block average method was used to calculate the standard error of the mean by dividing the simulation into 10 ns blocks.

TABLE 2 | Selected stacking interactions and hydrogen bonds observed in the trajectories of unbound riboswitches with standard deviations (in parentheses) calculated based on the two halves of the trajectory. "–" stands for undetected interaction. The nucleotide sequence is shown in **Figure 1**. For the full list of interactions, see **Supplementary Tables S4–S6**.

		Percentage of simulation time [%]				
Base 1	**Base 2**	**N1**	**U14C**	**U14C$^+$**	**U15A**	**A17G**
Stacking interactions						
C6	U7	11 (7)	1 (1)	1 (1)	1 (1)	40 (4)
C6	A17	1 (1)	11 (11)	1 (0)	35 (2)	8 (6)
U7	A17	35 (4)	2 (2)	10 (2)	2 (2)	65 (5)
U8	C12	–	1 (1)	1 (1)	7 (1)	54 (4)
Hydrogen bonds						
U8:O2'	C11:N4	–	1 (1)	–	1 (1)	52 (2)
U10:OP2	U8:O2'	–	5 (5)	1 (1)	4 (1)	51 (2)
U13:O4	U8:N3	–	–	–	1 (1)	37 (2)

reduces the RMSD of RNA. In all cases, the RMSD stabilizes after the initial 50 ns of trajectories. **Figure 2** shows that consistently the highest RMSF are in the bulge and apical loop regions of riboswitches. The fluctuations of the terminal nucleotides are expected but not considered or discussed because under cellular conditions the termini are connected with other fragments of mRNA. Although mutations were introduced in the apical loop of the riboswitch, the fluctuations in the bulge region also change. For instance, in the bound state, the A17G mutation causes higher fluctuations of the bulge as compared to the unmutated N1 riboswitch. There is also a difference in the apical loop fluctuations between the U14C and U14C$^+$ systems both in the unbound and bound states. The amplitude of both RMSD and RMSF values for the N1 riboswitch without mutations corresponds well with the previous results in the temperature REMD simulations (Kulik et al., 2018), implying that similar behaviour of the riboswitch and the same level of conformational sampling have been achieved in both methods.

3.2 The C6:A17 contact joins the bulge and apical loop in the riboswitch complexes

To examine how the dynamics of the apical loop affects the fluctuations of the bulge we compared the representative structures from gREST simulations (**Figure 3 and Supplementary Figure S4**). In the unbound systems, all colored bases in the bulge and apical loop acquire various conformations, while the upper and lower stem bases are stably paired. In the simulations with neomycin, U7 and U8 adopt conformations outside the hairpin, while the apical loop nucleotides mostly remain stable due to the C6:A17 stacking. Since this C6:A17 interaction mimics a clip that fastens the bulge and apical loop together, we call it paperclip C6:A17.

The paperclip C6:A17 is rarely found in the trajectories of the unbound riboswitch (**Table 2**), for which the experimental structures are not known. On the other hand, it is visible in more than 82% of simulation time in all complexes (**Table 3**). It is also present in all experimental NMR models for the riboswitch complexes with ribostamycin (Duchardt-Ferner et al., 2010) and paromomycin (Duchardt-Ferner et al., 2016). This finding implies that the C6:A17 stacking interaction is mainly formed after neomycin binding and is stabilized by the presence of the ligand. Assuming the two-step binding mechanism (Gustmann et al., 2019), the initial neomycin binding is likely followed by the formation of the C6:A17 paperclip with a minor conformational change that adjusts the binding pocket to the presence of the ligand.

The most frequent conformations observed in the trajectories of unbound riboswitches (named Cluster 1, Cluster 2, Cluster 3 in **Supplementary Figure S5–S9**) are characterized by higher structural variety than those in the bound state. The RNA binding pocket is often wide open, without any contacts between the bulge and apical loop. The characteristic paperclip C6:A17 appears in 35% of the simulation frames of the U15A riboswitch, including 17% of the conformations in the second most populated cluster (**Figure 3** and Cluster 2 in **Supplementary Figure S8A**). This paperclip is not found in the representative unbound structures of the N1 riboswitch and U14C, U14C$^+$ and A17G mutants.

3.3 The A17G mutant in the unbound state forms paperclips involving the U7 base

In the A17G unbound mutant in Clusters 1 and 3 (representing 17 and 16% of the simulation frames, respectively) a strong contact between the bulge U7 and apical loop G17 bases was

TABLE 3 | Selected stacking interactions and hydrogen bonds observed in the trajectories of neomycin-bound riboswitches with standard deviations (in parentheses) calculated based on the two halves of the trajectory. "–" stands for undetected interaction. The nucleotide sequence is shown in **Figure 1**. For the full list of interactions, see **Supplementary Tables S7, S8**.

		Percentage of simulation time [%]				
Base 1	**Base 2**	**N1_NEO**	**U14C_NEO**	**U14C$^+$_NEO**	**U15A_NEO**	**A17G_NEO**
Stacking interactions						
C6	U7	–	35 (4)	18 (17)	10 (2)	26 (2)
C6	U8	27 (3)	4 (2)	5 (1)	10 (7)	7 (2)
C6	A17	87 (3)	93 (1)	94 (4)	94 (4)	82 (7)
U7	A17	–	9 (1)	5 (4)	2 (1)	25 (14)
Hydrogen bonds						
G5:O2'	A17:N6*	70 (1)	76 (2)	85 (0)	57 (3)	1 (0)
G5:OP1	C6:N4	23 (1)	26 (6)	18 (2)	17 (5)	3 (1)
G5:OP1	A17:C2*	–	–	–	–	31 (4)
G5:OP2	C6:N4	32 (2)	34 (7)	38 (9)	37 (1)	8 (2)
C6:O2'	U7:OP1	32 (1)	19 (6)	30 (1)	21 (5)	4 (1)
U14:N3	A17:OP2	70 (1)	–	67 (1)	47 (5)	62 (12)

**in the A17G–mutant A17:C2 changes to G17:NH_2, and A17:N6 to G17:O6.*

found, called paperclip U7:G17 (**Figure 3** and **Supplementary Figure S9A**). This paperclip exists in the N1 riboswitch without mutations but it is not common–the U7:A17 stacking is seen in 35% of simulation frames (**Table 2**). On the other hand, the U7:G17 stacking interaction is present in 65% of frames. Additional van der Waals interaction with C6, visible in Cluster 3 of the A17G mutant, forms a triple stacking interaction, named paperclip C6:U7:G17. This interaction is present for 25% of simulation time in the A17G system and was not detected in other systems (**Supplementary Table S5**). These paperclips break and form in time, as illustrated in **Supplementary Figure S10**.

The A17G mutation changes the dynamics of the C6, U7, U8 and G17 bases, which are able to flip in or out from the riboswitch. A good way to quantify the flipped-in and flipped-out states is to analyze pseudo-dihedral angles formed by selected bases and phosphate groups, which can be carried out according to the scheme proposed by Song *et al.* (Song et al., 2009). The definition of an exemplary pseudo-dihedral angle is shown in **Supplementary Figure S11**. Of a particular interest are the high peaks for the A17G system at 123° for the A17 base and at 170° for the U7 base, shown in **Figure 4A,C**, which correspond to the paperclip U7:G17 arrangement. **Figure 4D,E** presents the structural view of the latter paperclip in N1 and A17G systems. The results of the pseudo-dihedral angle analysis for all systems are shown in **Supplementary Figures S12, S13**. Stable conformations of C6 and A17 bases in the paperclip C6:A17 in

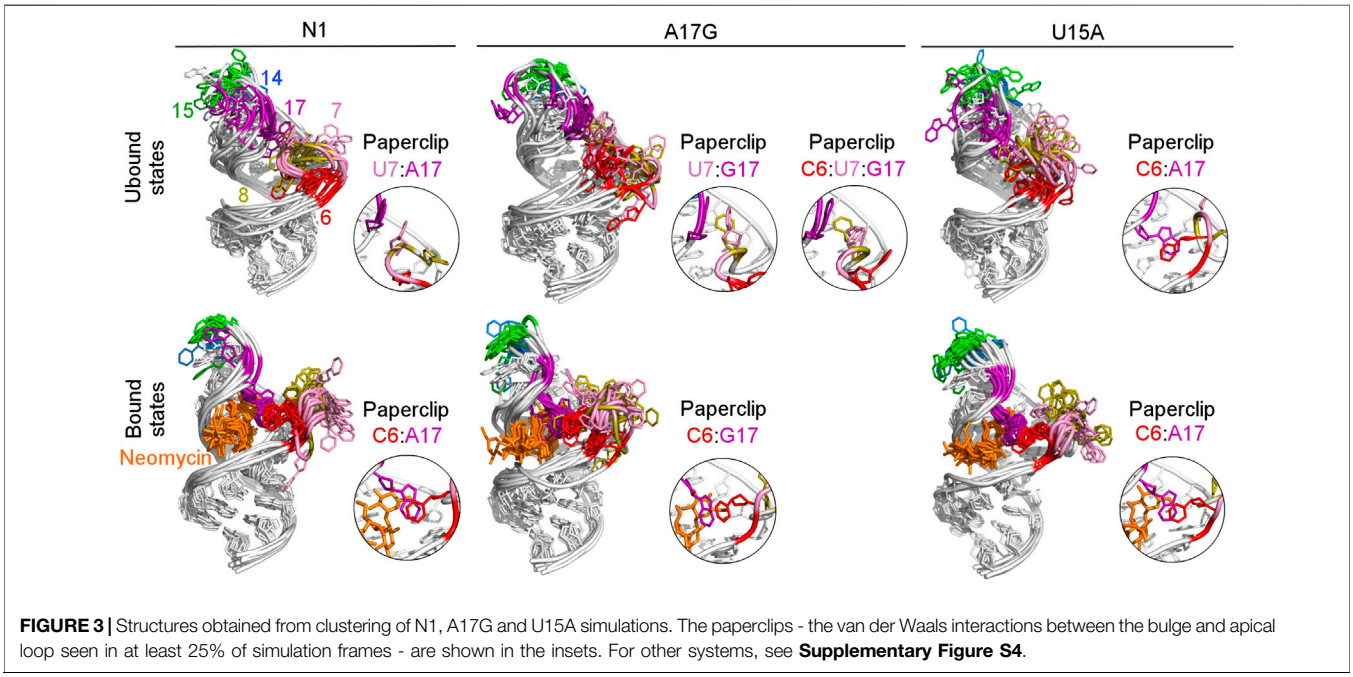


FIGURE 3 | Structures obtained from clustering of N1, A17G and U15A simulations. The paperclips - the van der Waals interactions between the bulge and apical loop seen in at least 25% of simulation frames - are shown in the insets. For other systems, see **Supplementary Figure S4**.

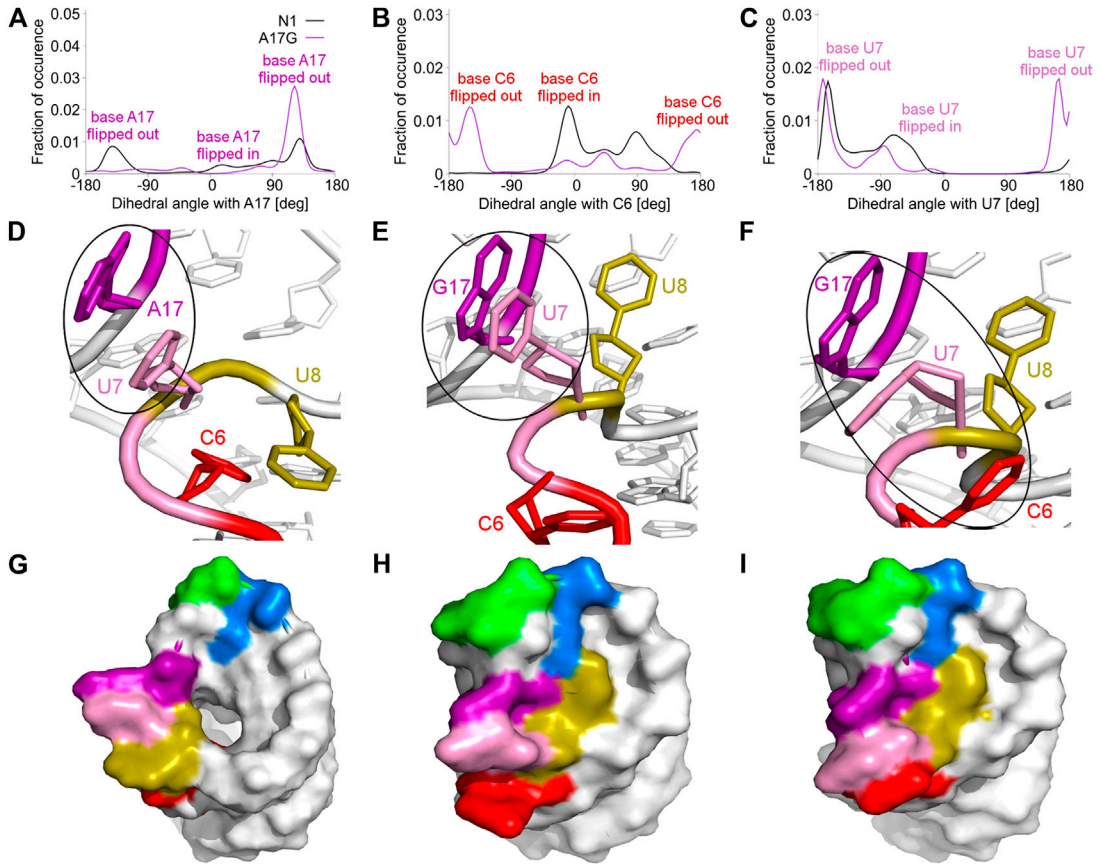


FIGURE 4 | Pseudo-dihedral angles and types of paperclips observed in the unbound N1 and A17G riboswitches. **(A–C)** The distribution of pseudo-dihedral angles for A17, C6 and U7 bases. **(D)** The structure of the paperclip U7:A17 contact in N1 riboswitch. **(E, F)** Structures of paperclips U7:G17 and C6:U7:G17 in the A17G mutant. **(G–I)** Corresponding molecular surfaces.

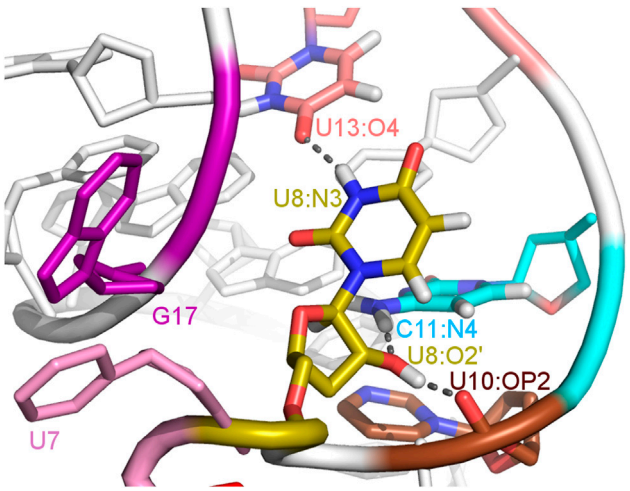


FIGURE 5 | The U8 nucleotide surrounding in the unbound state of the A17G system with hydrogen bonds marked as dashed lines.

the riboswitches with neomycin correspond to the high peaks in their pseudo-dihedral angle distribution (**Supplementary Figure S13**). In contrast, the flexible bases in the unbound state reveal a wide distribution of pseudo-dihedral angles.

3.4 The U8 base is directed towards the binding site in the unbound A17G riboswitch

The paperclip U7:A17 conformation in the N1 riboswitch is associated with the C6 pseudo-dihedral angle of −9° (**Figure 4B**). In this conformation, the C6 base is directed towards the riboswitch and, at the same time, U8 is directed outside. Paperclip U7:G17 in the A17G mutant is also described by the U7 pseudo-dihedral angle at 170°, but the C6 and U8 bases locate differently as compared to N1, i.e., C6 is directed outside and U8 inside the riboswitch (**Figure 4E**). In turn, the structure described by the dihedral angle of -149° for C6, shows a conformation that occurs only in A17G and is characterized by the presence of C6:U7:G17 triple stacking, as shown in **Figure 4F**. The U8 base acquires the same position as in

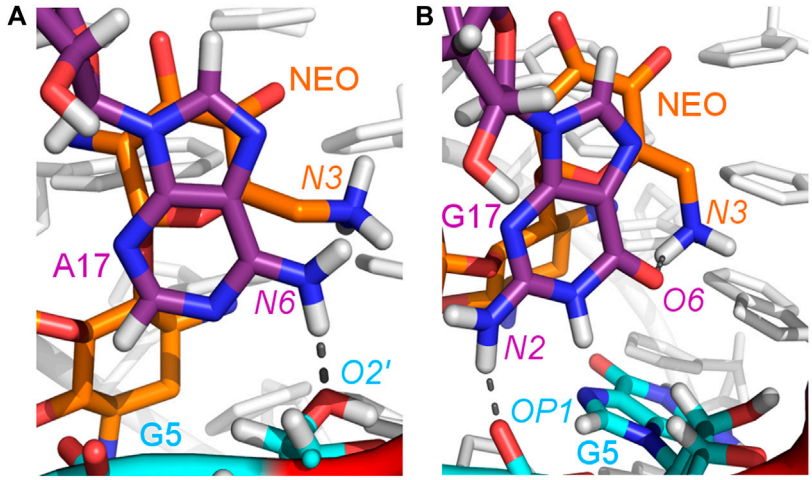


FIGURE 6 | Location of the G5–A17/G17 and G17–N3 hydrogen bonds in: **(A)** N1_NEO and **(B)** A17G_NEO systems.

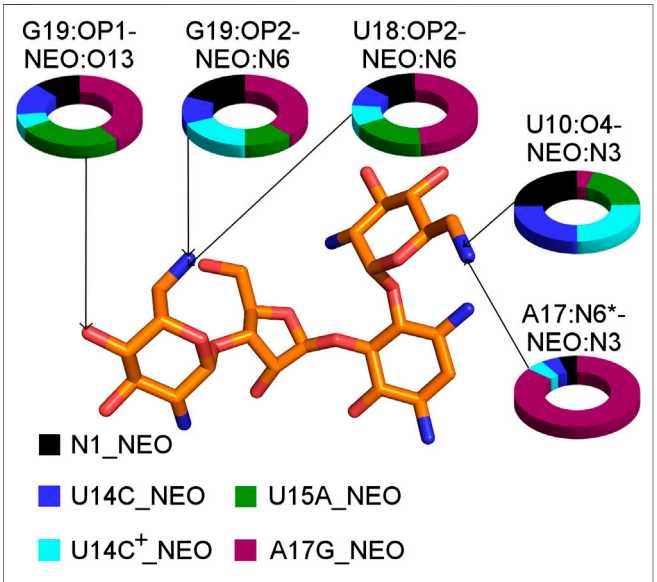


FIGURE 7 | The shares of selected RNA–neomycin interactions.
* indicates N6 in A17 and O6 in G17. All RNA–neomycin interactions are listed in **Supplementary Table S9**.

paperclip U7:G17, namely inside the binding site. Molecular surfaces of the paperclip structures show that the conformation of the U8 base differs between the N1 and A17G systems. In the N1 riboswitch (**Figure 4G**), the binding site is accessible, but in the A17G mutant this site is occupied by U8 (**Figure 4H,I**). In the A17G system, the conformation of the U8 base in the binding site is stabilized by three hydrogen bonds (U8:O2'–U10:OP2, U8:O2'–C11:N4 and U8:N3–U13:O4) shown in **Figure 5**. The occurrence of those hydrogen bonds varies between 37% and 52% of the simulation time in the A17G system and is negligible in all other systems (**Table 2**). Such position of U8 in the binding site may withhold neomycin from entering the riboswitch according to the conformational selection mechanism and thus result in the depleted activity of the mutated riboswitch. However, this is just a hypothesis since the presence of U8 in the A17G binding pocket was not investigated experimentally.

3.5 The A17G mutant changes the hydrogen bond network around neomycin

The A17G mutation changes the interaction network also in the neomycin-bound riboswitch. The A to G substitution breaks the hydrogen bond between G5:O2' and A17:N6 that occurs in the N1–NEO structure for 70% of the simulation time (**Figure 6A** and **Table 3**). In the mutant, the N6 amino group is superseded by a carbonyl group, so a new hydrogen bond is possible with the G5 phosphate oxygen (**Figure 6B**). However, the latter hydrogen bond is less stable, as shown by hydrogen bond frequency analysis (**Table 3**). In the N1_NEO, U14C_NEO, U14C$^+$_NEO and U15A_NEO systems, the riboswitch interacts with the N3 group of neomycin through a hydrogen bond with U10:O4 for more than half of the simulation time (**Figure 7**). The amino-to-carbonyl replacement in A17G also results in a new interaction — G17:O6—NEO:N3. This interaction with neomycin stabilizes the G17 base for 55% of the simulation time and is not present in other systems (**Supplementary Table S9**). The U7 stacking interaction with G17 also becomes possible in the A17G complex with neomycin (**Supplementary Table S7**). Those interactions point to partial destabilization of G17 in the A17G complex. The RMSF of base 17 in **Figure 2B** is higher in N1 than in the A17G mutant due to a single event of riboswitch opening in the N1 complex. This is visible as a peak in the RMSD plot between 180 and 200 ns (**Supplementary Figure S3B**, black line), which reflects an unusual flipped-out position of A17 shown in one of the representative structures of N1 bound state in **Figure 3**, present in only 7% of conformations. Overall, the RMSD values are on average 0.5 Å higher for the bound A17G structure than for the N1 complex (**Supplementary Figure S3**). In step with that, neomycin bound to A17G has some

additional conformational freedom as its RMSD after RNA fitting is at some points about two times higher than in other complexes (**Supplementary Figure S14**). Due to different interaction pattern around neomycin, in the A17G complex the ligand is slightly shifted as compared to the complex without mutation (**Supplementary Figure S15**).

The ITC and fluorescence assays for the N1 and A17G mutant with neomycin are gathered in **Table 1**. However, the NMR experiments of the riboswitch were conducted with ribostamycin only, due to superior spectral resolution of the ligand resonances and similar binding modes to the riboswitch mutants (Duchardt-Ferner et al., 2010). Neomycin possesses additional ring IV, whose presence decreases the dissociation constant and increases the regulatory activity of N1 riboswitch with respect to ribostamycin (compare the values for the N1 complexes in **Table 1**). The imino proton solvent exchange rate measurements report that the terminal loop and the U13:U18 base pair in the A17G–ribostamycin complex are less stable than in the N1 complex (Weigand et al., 2014). According to our simulation results, in A17G_NEO, the neomycin ring IV interacts with U18 and G19 phosphate groups. These additional neomycin–RNA hydrogen bonds might contribute to partial U13:U18 base pair stabilization in the A17G complex. As a result, we observe the U13:O2'–U18:N3 hydrogen bond in over 63% and 84% of frames in N1 and A17G complexes, respectively (**Supplementary Tables S8, S9**). Thus, we anticipate that the simulations of the A17G–ribostamycin complex would lack the stabilizing interactions between RNA and ring IV of the ligand, leading to the lower stability of ribostamycin in the complex, larger dissociation constant and lower regulatory activity in comparison to the A17G–neomycin complex.

Furthermore, the NOESY spectrum of the ^{13}C-guanine-labeled RNA suggests the *syn* conformation of G17 in the A17G complex with ribostamycin, in contrast to the *anti* conformation of A17 in the N1–ribostamycin complex (Weigand et al., 2014). In our simulations of the unbound A17G mutant the distribution of G17 χ torsion angles shows the *syn* conformation, while the other unbound riboswitch variants sample the *anti* conformation (**Supplementary Figure S16**). However, in the complexes with neomycin, all the systems prefer the *anti* conformation, with a slight shift in the orientation of the G17 base position due to a different interaction network as pointed out earlier. The sampling of the *syn* conformation by G17 seen in NMR experiments led the authors to suggest that the N3 group of ribostamycin ring I of ribostamycin does not interact with G17, the C6:G17 stacking interaction is prevented, which increases ligand flexibility (Weigand et al., 2014). Taken together, our results point to other reasons for the observed higher flexibility of the ligand in the simulations; the C6:G17 stacking is present and a strong hydrogen bond between G17:O6 and NEO:N3 directs neomycin to bind in a slightly shifted and less stable position in the A17G binding site with respect to the N1 complex (**Supplementary Figure S15**).

3.6 The gREST simulations suggest the protonation of cytosine 14 in the U14C mutant

The U-turn motif is characterized by the presence of these hydrogen bonds U14:O2'–A16:N7, U14:N3–A17:P and A16:O2'–U18:P (Weigand et al., 2014). Two of these hydrogen bonds, between U14–A16 and A16–U18 are present in all the simulations of the neomycin–riboswitch complexes. In the U14C complex without protonated C, the Watson-Crick edge lacks the proton underlying the C14:N3–A17:P hydrogen bond. Therefore, in **Table 3**, the C14:N3–A17:OP2 hydrogen bond in the U14C_NEO system is absent and there are no other C14–A17 hydrogen bonds, which destabilizes the U-turn. On the other hand, the protonated cytosine can form C14$^+$:N3–A17:OP2 interaction, shown in **Figure 8**, and restore the stability of the U-turn. This hydrogen bond is present for 67% of the simulation time in the U14C$^+$_NEO system, which is close to 70% of simulation frames for the hydrogen bond with U14 in N1_NEO (**Table 3**). The U to C function-retaining replacements in the U-turn motifs have been reported previously and confirmed in the context of the N1 riboswitch (Gottstein-Schmidtke et al., 2014; Krepl et al., 2018).

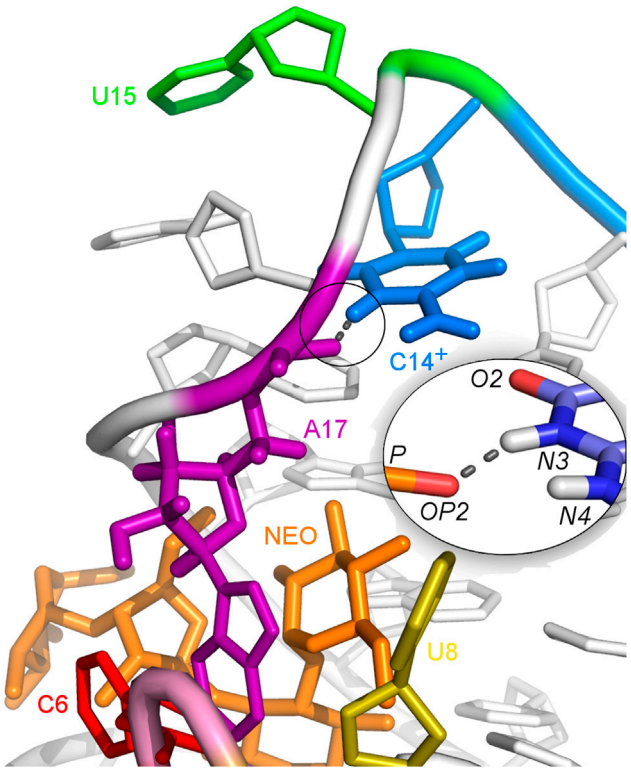


FIGURE 8 | Hydrogen bonds between C14$^+$ and A17 observed in the U14C$^+$_NEO system. The proton connected to the N3 atom is missing in the U14C mutant without cytosine protonation.

In the unbound state of the riboswitch, the above-mentioned hydrogen bonds that form the U-turn are rare, even in the U14C$^+$ case (**Supplementary Table S6**). Thus, when no ligand is present, the protonation of C14 is not enough for the U-turn formation. Lack of U-turn causes large fluctuations of the apical loop in all systems (**Figure 2A**), especially in the U14C system, while A17 in the U14C$^+$ mutant fluctuates less than in other systems. Also, the χ torsion angles of A17 in the U14C$^+$ mutant sample the *syn* conformation, similarly to the A17G system, while the N1

riboswitch samples the *anti* conformation. In the unbound riboswitches, the highest peaks in the pseudo-dihedral angle distributions of C6, U7, U8, and A17 bases observed in the N1 system coincide more often with the highest peaks in the U14C than in the $U14C^+$ system (**Supplementary Figure S12**). Those deviations of the $U14C^+$ system from the N1 system prove that the protonation of C14 cannot mimic the interactions and dynamics of U14 in the N1 riboswitch. This means that the $U14C^+$ system is closer to inactive A17G mutant, which may significantly reduce the number of conformations preformed for ligand binding. According to the experimental data, the U14C mutation reduces the N1 riboswitch activity 3-fold, but it is not as small as for the A17G system (**Table 1**). Also, the dissociation constant of the neomycin -U14C mutant is 3 times larger than for the N1 system but almost 6 times smaller than for the A17G system. Thus, the cytosine in the U14C mutant in the unbound state is probably protonated to a low extent, which would agree with the imino proton NMR signals for the U14C mutant (Weigand et al., 2014).

3.7 The U15A mutant displays conformations preformed for ligand binding

Mutating U15 to A only slightly affects the dynamics of the riboswitch, which is visible in the similar fluctuations in the unbound and bound states of the N1 and U15A variants in **Figure 2**. The only difference with respect to the unmutated structure is the presence of the bound state conformations in the unbound state, such as the paperclip C6:A17 that appears in the second most populated cluster in the unbound U15A trajectory - Cluster 2 in **Supplementary Figure S8A**. The stacking interaction C6:A17 is not visible in the N1 simulation (**Table 2**). Also, the pseudo-dihedral angle distributions of C6, U8 and A17 in the unbound state resemble those from the complex (**Supplementary Figures S12, S13**). Comparison of 1D imino-proton NMR signals of the ligand-free N1 and U15A systems indicates stronger signals for the U13/18 and U10/21 resonances in the U15A system (Weigand et al., 2014), confirming the higher degree of preformation in the latter system. This is in agreement with the frequency of U13:O2—U18:N3 and U10:O2—U21:N3 hydrogen bonds in our simulations, which is more than 15% higher in the U15A simulations than in N1 simulations (**Supplementary Table S6**). Thus, the U15A mutation facilitates the conformational selection and, as a result, increases the regulatory activity of the N1 riboswitch toward neomycin as evidenced by the fluorescence assay (Weigand et al., 2014).

4 CONCLUSION AND OUTLOOK

We performed the gREST simulations for a set of single-point N1 riboswitch mutants in the unbound and neomycin-bound states. The dynamics of the interactions unveiled the reasons behind the hindered activity of the A17G mutant. The A17G mutation affects the unbound riboswitch dynamics because additional bulge-apical loop contacts are formed that act as paperclips and narrow the set of conformations necessary for neomycin entering the riboswitch. This is in line with the conformational selection mechanism suggested for the N1 riboswitch. Furthermore, the dynamical studies of the U14C mutant in different protonation states of C14 support the protonated state of this cytosine in the bound state and deprotonated one in the unbound state. Additionally, the dynamics of the U15A mutant resembles the dynamics of the riboswitch without mutation, except that in the unbound state the U15A mutant displays more conformations preformed for ligand binding than N1. This corresponds to the slightly higher regulatory activity of this mutant compared to the unmutated riboswitch (Weigand et al., 2014). Overall, the U15A mutant is the most effective among the studied riboswitch sequences because the dynamics of this mutant in the unbound state facilitates conformational selection for neomycin binding. Our studies of the mutations introduced to the N1 apical loop corroborated the experimental data and provided insight into the riboswitch dynamics–activity relationship. In the future it would be interesting to investigate the conformational dynamics of N1 replacing also the bulge nucleobases, which could unveil the mutations compensating the apical loop ones analyzed in this work.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

MK, YS, and JT conceived the work. MK, PC, and SR prepared the simulation input files. PC performed and analyzed the simulations and wrote the manuscript draft. MK, SR, YS, and JT co-wrote the manuscript.

FUNDING

National Science Centre Poland (DEC-2017/26/M/NZ1/00827, HARMONIA to JT, PC, and MK); simulations were performed at CeNT and ICM U. Warsaw (GB77-7). Funding for open access charge: National Science Centre Poland (DEC-2017/26/M/NZ1/00827, HARMONIA to JT, PC, and MK).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.633130/full#supplementary-material.

REFERENCES

Abraham, M. J., van der Spoel, D., Lindahl, E., and Hess, B. (2019). *GROMACS Source code version 2019.*

Banerjee, J., and Nilsen-Hamilton, M. (2013). Aptamers: multifunctional molecules for biomedical research. *J. Mol. Med.* 91, 1333–1342. doi:10.1007/s00109-013-1085-2

Bastet, L., Dubé, A., Massé, E., and Lafontaine, D. A. (2011). New insights into riboswitch regulation mechanisms. *Mol. Microbiol.* 80, 1148–1154. doi:10.1111/j.1365-2958.2011.07654.x

Berens, C., Groher, F., and Suess, B. (2015). RNA aptamers as genetic control devices: the potential of riboswitches as synthetic elements for regulating gene expression. *Biotechnol. J.* 10, 246–257. doi:10.1002/biot.201300498

Bergonzo, C., Henriksen, N. M., Roe, D. R., Swails, J. M., Roitberg, A. E., and Cheatham, T. E. (2014). Multidimensional replica exchange molecular dynamics yields a converged ensemble of an RNA tetranucleotide. *J. Chem. Theory Comput.* 10, 492–499. doi:10.1021/ct400862k

Boussebayle, A., Torka, D., Ollivaud, S., Braun, J., Bofill-Bosch, C., Dombrowski, M., et al. (2019). Next-level riboswitch development-implementation of Capture-SELEX facilitates identification of a new synthetic riboswitch. *Nucl. Acids Res.* 47, 4883–4895. doi:10.1093/nar/gkz216

Bussi, G., Donadio, D., and Michele, P. (2007). Canonical sampling through velocity rescaling. *J.Chem. Phys.* 126, 014101. doi:10.1063/1.2408420

Bussi, G., Zykova-Timan, T., and Michele, P. (2009). Isothermal-isobaric molecular dynamics using stochastic velocity rescaling. *J.Chem. Phys.* 130, 074101. doi:10.1063/1.3073889

Case, D. A., Babin, V., Berryman, J. T., Betz, R. M., Cai, Q., Cerutti, D. S., et al. (2014). Technical report. AMBER 14. Available at: http://ambermd.org/contributors (Accessed May, 2001).

Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., et al. (1995). A second generation force field for the dimulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.* 117, 5179–5197. doi:10.1021/ja955032e

Duchardt-Ferner, E., Gottstein-Schmidtke, S. R., Weigand, J. E., Ohlenschlager, O., Wurm, J. P., Hammann, C., et al. (2016). What a difference an OH makes: conformational dynamics as the basis for the ligand specificity of the neomycin-sensing riboswitch. *Angew. Chem. Int. Ed. Engl.* 55, 1527–1530. doi:10.1002/anie.201507365

Duchardt-Ferner, E., Weigand, J. E., Ohlenschlager, O., Schmidtke, S. R., Suess, B., and Wöhnert, J. (2010). Highly modular structure and ligand binding by conformational capture in a minimalistic riboswitch. *Angew. Chem. Int. Ed. Engl.* 49, 6216–6219. doi:10.1002/anie.201001339

Ellington, A. D., and Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature* 346, 818–822. doi:10.1038/346818a0

Essmann, U., Perera, L., Darden, T., Lee, H., and Pedersen, L. (1995). A smooth particle mesh Ewald method. *J. Chem. Phys.* 103, 8577–8593. doi:10.1063/1.470117

Etzel, M., and Mörl, M. (2017). Synthetic riboswitches: From plug and pray toward plug and play. *Biochemistry* 56, 1181–1198. doi:10.1021/acs.biochem.6b01218

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* 21, 768–769.

Gaussian Inc (2009). Gaussian 09. Available at: https://gaussian.com/g09citation/ (Accessed November 01, 2016).

Górska, A., Jasiński, M., and Trylska, J. (2015). MINT: software to identify motifs and short-range interactions in trajectories of nucleic acids. *Nucl. Acids Res.* 43, e114. doi:10.1093/nar/gkv559

Gottstein-Schmidtke, S. R., Duchardt-Ferner, E., Groher, F., Weigand, J. E., Gottstein, D., Suess, B., et al. (2014). Building a stable RNA U-turn with a protonated cytidine. *RNA* 20, 1163–1172. doi:10.1261/rna.043083.113

Gustmann, H., Segler, A. J., Gophane, D. B., Reuss, A. J., Grünewald, C., Braun, M., et al. (2019). Structure guided fluorescence labeling reveals a two-step binding mechanism of neomycin to its RNA aptamer. *Nucl. Acids Res.* 47, 15–28. doi:10.1093/nar/gky1110

Hayes, R. L., Noel, J. K., Mohanty, U., Whitford, P. C., Hennelly, S. P., Onuchic, J. N., et al. (2012). Magnesium fluctuations modulate rna dynamics in the sam-i riboswitch. *J. Am. Chem. Soc.* 134, 12043–12053. doi:10.1021/ja301454u

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

Jorgensen, W., Chandrasekhar, J., Madura, J., and Klein, M. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79, 926–935. doi:10.1063/1.445869

Joung, I. S., and Cheatham, T. E. (2008). Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J Phys Chem B* 112, 9020–9041. doi:10.1021/jp8001614

Jung, J., Mori, T., Kobayashi, C., Matsunaga, Y., Yoda, T., Feig, M., et al. (2015). GENESIS: a hybrid-parallel and multi-scale molecular dynamics simulator with enhanced sampling algorithms for biomolecular and cellular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 5, 310–323. doi:10.1002/wcms.1220

Kamiya, M., and Sugita, Y. (2018). Flexible selection of the solute region in replica exchange with solute tempering: Application to protein-folding simulations. *J. Chem. Phys* 149, 072304. doi:10.1063/1.5016222

Kobayashi, C., Jung, J., Matsunaga, Y., Mori, T., Ando, T., Kamura, K., et al. (2017). GENESIS 1.1: A hybrid-parallel molecular dynamics simulator with enhanced sampling algorithms on multiple computational platforms. *J. Comput. Chem.* 38, 2193–2206. doi:10.1002/jcc.24874

Krepl, M., Vögele, J., Kruse, H., Duchardt-Ferner, E., Wöhnert, J., and Šponer, J. (2018). An intricate balance of hydrogen bonding, ion atmosphere and dynamics facilitates a seamless uracil to cytosine substitution in the U-turn of the neomycin-sensing riboswitch. *Nucl. Acids Res.* 46, 6528–6543. doi:10.1093/nar/gky490

Kulik, M., Mori, T., Sugita, Y., and Trylska, J. (2018). Molecular mechanisms for dynamic regulation of N1 riboswitch by aminoglycosides. *Nucl. Acids Res.* 46, 9960–9970. doi:10.1093/nar/gky833

Liu, P., Kim, B., Friesner, R. A., and Berne, B. J. (2005). Replica exchange with solute tempering: a method for sampling biological systems in explicit water. *Proc. Natl. Acad. Sci. U.S.A* 102, 13749–13754. doi:10.1073/pnas.0506346102

McCown, P. J., Corbino, K. A., Stav, S., Sherlock, M. E., and Breaker, R. R. (2017). Riboswitch diversity and distribution. *RNA* 23, 995–1011. doi:10.1261/rna.061234.117

Miyamoto, S., and Kollman, P. A. (1992). Settle: an analytical version of the shake and rattle algorithm for rigid water models. *J. Comput. Chem.* 13, 952–962. doi:10.1002/jcc.540130805

Moors, S., Michielssens, S., and Ceulemans, A. (2011). Improved replica exchange method for native-state protein sampling. *J. Chem. Theory Comput.* 7, 231–237. doi:10.1021/ct100493v

Pérez, A., Marchan, 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

Quigley, D., and Probert, M. (2004). Langevin dynamics in constant pressure extended systems. *J.Chem. Phys.* 120, 11432–11441. doi:10.1063/1.1755657

Roe, D. R., and Cheatham, T. E. (2013). Ptraj and cpptraj: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comp.* 9, 3084–3095. doi:10.1021/ct400341p

Roy, S., Hennelly, S. P., Lammert, H., Onuchic, J. N., and Sanbonmatsu, K. Y. (2019). Magnesium controls aptamer-expression platform switching in the SAM-I riboswitch. *Nucl. Acids Res.* 47, 3158–3170. doi:10.1093/nar/gky1311

Roy, S., Lammert, H., Hayes, R. L., Chen, B., LeBlanc, R., Dayie, T. K., et al. (2017a). A magnesium-induced triplex pre-organizes the SAM-II riboswitch. *PLoS Comput. Biol.* 13, e1005406. doi:10.1371/journal.pcbi.1005406

Roy, S., Onuchic, J. N., and Sanbonmatsu, K. Y. (2017b). Cooperation between magnesium and metabolite controls collapse of the SAM-I Riboswitch. *Biophys J* 113, 348–359. doi:10.1016/j.bpj.2017.06.044

Ryckaert, J., Ciccotti, G., and Berendsen, H. J. (1977). Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 23, 327–341. doi:10.1016/0021-9991(77)90098-5

Schrödinger, LLC (2010). The PyMOL Molecular Graphics System. Version 1.3r1. https://www.schrodinger.com/citations

Serganov, A., and Nudler, E. (2013). A decade of riboswitches. *Cell* 152, 17–24. doi:10.1016/j.cell.2012.12.024

Sinumvayo, J. P., Zhao, C., and Tuyishime, P. (2018). Recent advances and future trends of riboswitches: attractive regulatory tools. *World J. Microbiol. Biotechnol.* 34, 171. doi:10.1007/s11274-018-2554-0

Song, K., Campbell, A., Bergonzo, C., de Los Santos, C., Grollman, A., and Simmerling, C. (2009). An improved reaction coordinate for nucleic acid

base flipping studies. *J. Chem. Theory Comput.* 5, 3105–3113. doi:10.1021/ct9001575

Šponer, J., Bussi, G., Krepl, M., Banáš, P., Bottaro, S., Cunha, R., et al. (2018). RNA structural dynamics a s captured by molecular simula tions: a comprehensive overview. *Chem. Rev.* 118, 4177–4338. doi:10.1021/acs.chemrev.7b00427

Sugita, Y., and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314, 141–151. doi:10.1016/S0009-2614(99)01123-9

Terakawa, T., Kameda, T., and Takada, S. (2011). On easy implementation of a variant of the replica exchange with solute tempering in GROMACS. *J. Comput. Chem.* 32, 1228–1234. doi:10.1002/jcc.21703

Tuckerman, M., Berne, B. J., and Martyna, G. J. (1992). Reversible multiple time scale molecular dynamics. *J. Chem. Phys.* 97, 1990–2001. doi:10.1063/1.463137

Wang, J., Cieplak, P., and Kollman, P. (2000). How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of or ganic and biological molecules? *J. Comput. Chem.* 21, 1049–1074. doi:10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F

Wang, J., Wolf, R., Caldwell, J., Kollamn, P., and Case, D. (2004). Development and testing of a general Amber force field. *J. Comput. Chem.* 25, 1157–1174. doi:10.1002/jcc.20035

Wang, L., Friesner, R. A., and Berne, B. J. (2011). Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *J.Phys.Chem.B* 115, 9431–9438. doi:10.1021/jp204407d

Weigand, J. E., Gottstein-Schmidtke, S. R., Demolli, S., Groher, F., Duchardt-Ferner, E., Wöhnert, J., et al. (2014). Sequence elements distal to the ligand binding pocket modulate the efficiency of a synthetic riboswitch. *Chembiochem* 15, 1627–1637. doi:10.1002/cbic.201402067

Weigand, J. E., Sanchez, M., Gunnesch, E. B., Zeiher, S., Schroeder, R., and Suess, B. (2008). Screening for engineered neomycin riboswitches that control translation initiation. *RNA* 14, 89–97. doi:10.1261/rna.772408

Weigand, J. E., Schmidtke, S. R., Will, T. J., Duchardt-Ferner, E., Hammann, C., Wöhnert, J., et al. (2011). Mechanistic insights into an engineered riboswitch: a switching element which confers riboswitch activity. *Nucl. Acids Res.* 39, 3363–3372. doi:10.1093/nar/gkq946

Zgarbova, M., Otyepka, M., Šponer, J., Mladek, A., Banáš, P., Cheatham, T., III, et al. (2011). Refinement of the nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theor. Comput.* 7, 2886–2902. doi:10.1021/ct200162x

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 © 2021 Chyży, Kulik, Re, Sugita and Trylska. 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 Molecular Biosciences

BRIEF RESEARCH REPORT
published: 10 March 2021
doi: 10.3389/fmolb.2021.636940

Case Report: Bayesian Statistical Inference of Experimental Parameters via Biomolecular Simulations: Atomic Force Microscopy

Sotaro Fuchigami, Toru Niina and Shoji Takada**

Department of Biophysics, Graduate School of Science, Kyoto University, Kyoto, Japan

The atomic force microscopy (AFM) is a powerful tool for imaging structures of molecules bound on surfaces. To gain high-resolution structural information, one often superimposes structure models on the measured images. Motivated by high flexibility of biomolecules, we previously developed a flexible-fitting molecular dynamics (MD) method that allows protein structural changes upon superimposing. Since the AFM image largely depends on the AFM probe tip geometry, the fitting process requires accurate estimation of the parameters related to the tip geometry. Here, we performed a Bayesian statistical inference to estimate a tip radius of the AFM probe from a given AFM image via flexible-fitting molecular dynamics (MD) simulations. We first sampled conformations of the nucleosome that fit well the reference AFM image by the flexible-fitting with various tip radii. We then estimated an optimal tip parameter by maximizing the conditional probability density of the AFM image produced from the fitted structure.

Keywords: atomic force microscopy, probe tip, flexible-fitting, coarse-grained molecular simulation, CafeMol

OPEN ACCESS

Edited by:
Fabio Trovato,
Freie Universität Berlin, Germany

Reviewed by:
Holger Flechsig,
Kanazawa University, Japan
Yi He,
University of New Mexico, Albuquerque, NY, United States
Abhishek Singharoy,
Arizona State University, Tempe, AZ, United States

****Correspondence:***
Sotaro Fuchigami
sotaro@biophys.kyoto-u.ac.jp
Shoji Takada
takada@biophys.kyoto-u.ac.jp

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *02 December 2020*
Accepted: *25 January 2021*
Published: *10 March 2021*

Citation:
Fuchigami S, Niina T and Takada S (2021) Case Report: Bayesian Statistical Inference of Experimental Parameters via Biomolecular Simulations: Atomic Force Microscopy. Front. Mol. Biosci. 8:636940. doi: 10.3389/fmolb.2021.636940

INTRODUCTION

The atomic force microscopy (AFM) is a powerful tool for imaging the structures of molecules bound on surface at atomic resolution (Ando et al., 2001; Kodera et al., 2006; Kodera et al., 2010; Uchihashi et al., 2011; Casuso et al., 2012; Ando et al., 2013; Ando et al., 2014; Dufrêne et al., 2017). However, because the AFM measurement gives only height information and because biomolecular AFM measurements often provide medium-resolution images, one often seeks molecular structures that fit to the AFM image. The fitting of rigid molecules can be achieved simply by translating and rotating a given structure model to find the best match to the AFM image. On the other hand, when the target molecules are flexible, as are often the case for biomolecules, one needs to allow a structural change of the model upon superimposing. This so-called flexible-fitting has been successfully applied in the modeling based on the cryo-electron microscopy data by various methods, for example, the molecular dynamics flexible fitting (MDFF) method and its extension (Trabuco et al., 2008; McGreevy et al., 2016; Singharoy et al., 2016), the correlation-coefficient-based method (Orzechowski and Tama, 2008), and CryoFold (Shekhar et al., 2020). The flexible-fitting can be realized by using molecular dynamics (MD) simulations, where a fitting score is integrated with the standard molecular mechanics force field. Recently, we developed a flexible-fitting MD method for finding molecular structures that fit the AFM image (Niina et al., 2020).

Generally, these fitting processes require knowledge on parameters that characterize the measurement, of which values are often unknown a priori. Therefore, one challenge is to infer

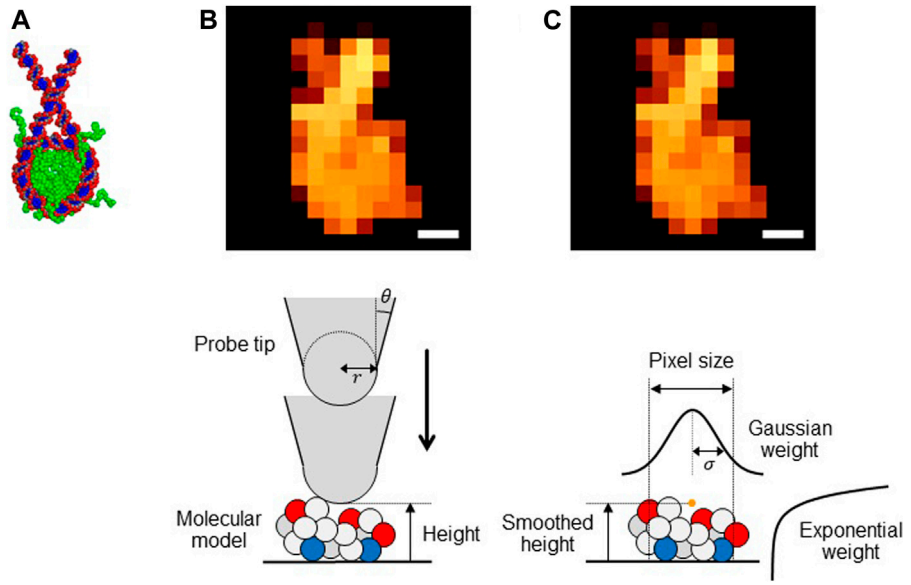


FIGURE 1 | (A) Structure of a nucleosome used in this study. **(B and C)** Pseudo-AFM images of nucleosome (Scale bar: 5 nm) generated **(B)** by the collision detection method and **(C)** by the smoothed method **(upper)** and schematic views of these methods **(lower)**.

these parameters, simultaneously finding the target molecular structures. In the case of the AFM measurement, such parameters include the size and the shape of the probe tip, which are usually unknown, but strongly affect the resulting images.

In this brief report, we focus on the inference of the radius of the AFM probe tip, assuming that it has the spherical shape. Using a Bayesian statistical inference approach, we examine feasibility of the inference of the AFM probe tip radius via flexible-fitting process. We chose a nucleosome as the test molecule (**Figure 1A**). The nucleosome consists of a histone octamer and a duplex DNA of 223 base pairs, and the initial structure was modeled using crystal structure (PDB ID: 3LZ0), as described in our previous work (Niina et al., 2017; Fuchigami et al., 2020).

METHODS

Pseudo-AFM Image Generating Method

The collision detection method: The collision-detection method is to generate a pseudo-AFM image for a given molecular structure. It assumes a simple geometry of the probe tip; the cone-shape with its terminus being a sphere (**Figure 1B**, lower). Thus, the tip can be characterized by the two parameters, the tip radius R and the apex angle θ. In this study, we fix the latter as $\theta = 10$ degree. For each pixel, we calculate the height of the bottom of the sphere at which a steric collision between the tip end and atoms in the molecular models occurs (**Figure 1B**).

The collision-detection method is available as a tool "afmize" (http://doi.org/10.5281/zenode.3362044). The BioAFMviewer, which provides the pseudo-AFM image via a similar method, is also available (Amyot and Flechsig, 2020).

The smoothed method: While the collision detection method offers perhaps the most straightforward way to generate a pseudo-AFM image from a given structure model, the method does not give a form differentiable with respect to atomic coordinates of target molecules. This precludes its usage in the flexible-fitting MD since MD simulations require the force calculations that include the differentiation of the fitting score. To this end, we previously proposed a differentiable proxy of the collision detection method, which we call the smoothed method. For a target biomolecule represented by its coordinate (x_j, y_j, z_j) and radius r_j for the j-th particle $(j = 1, \ldots, N)$, we define the smoothed height $H_p^{(\mathrm{sim})}$ at the p-th pixel (x_p, y_p) as

$$H_p^{(\mathrm{sim})} = \gamma \log\left\{1 + \sum_{j=1}^{N} \exp\left[\frac{-\left(x_j - x_p\right)^2 - \left(y_j - y_p\right)^2}{2\sigma^2}\right] \exp\left(\frac{z_j + r_j}{\gamma}\right)\right\}. \quad (1)$$

We assumed xy-plane as the AFM stage and thus that the z-coordinate represents the height of the target molecule. This function contains smoothing in xy-directions (characterized by σ), as well as the smoothing in z-direction (characterized by γ) (**Figure 1C**).

Importantly, in the smoothed method expression, the tip radius R does not appear directly. Instead, we previously showed that the optimal value of σ well correlates with the tip radius R. Thus, by changing σ, we can effectively generate pseudo-AFM images of different R's.

We note that a smaller γ corresponds to a closer approximation to the collision-detection method, but a too

small γ leads to a sharp change in the force and thus to the instability of MD simulations. In this study, we fix $\gamma = 0.1$ nm.

The smoothed method is also available in "afmize" (http://doi.org/10.5281/zenode.3362044).

Flexible-Fitting Molecular Dynamics Simulation

Flexible-fitting MD simulations utilize the total potential energy function, $V_{\mathrm{total}} = V_{\mathrm{phys}} + V_{\mathrm{AFM}}$. Here, V_{phys} represents the physical interaction of the target molecules as well as their interaction with the surface, the latter of which is modeled as a simple Lennard-Jones potential along the z axis. The second term is defined as $V_{\mathrm{AFM}}(R) = \kappa N k_{\mathrm{B}} T[1 - \mathrm{c.c.}(R)]$, where k_{B} is the Boltzmann constant, T is the temperature (300 K in this study), κ is a dimensionless parameter that controls the strength of the bias (unity in this study), and R is the coordinate of a simulated molecule. (Niina et al., 2020). The c.c.(R) is the modified correlation coefficients between the pseudo-AFM image $H_p^{(\mathrm{sim})}(R)$ of the simulating structure R and the real experimental AFM image $H_p^{(\mathrm{exp})}$, defined as

$$\mathrm{c.c.}(R) = \frac{\sum_{p \in \mathrm{pixels}} H_p^{(\mathrm{exp})} H_p^{(\mathrm{sim})}(R)}{\sqrt{\sum_{p \in \mathrm{pixels}} \left(H_p^{(\mathrm{exp})}\right)^2} \sqrt{\sum_{p \in \mathrm{pixels}} \left(H_p^{(\mathrm{sim})}(R)\right)^2}}. \quad (2)$$

See the reference (Niina et al., 2020) for more details.

Coarse-Grained Molecular Dynamics Simulation

In this study, as the physical interaction part of V_{phys}, we used a well-tested coarse-grained model for proteins and DNAs, which have been extensively used in many recent studies, including simulations for nucleosomes (Niina et al., 2017; Brandani et al., 2018; Tan and Takada, 2020). Briefly, proteins are represented by chains of beads, each of which represents an amino acid. For DNAs, each nucleotide is approximated by three particles, each representing sugar, phosphate, and base groups. We used the energy function AICG2+ for proteins (Li et al., 2014), 3SPN.2C for DNA (Freeman et al., 2014). Protein-DNA interactions are modeled by the excluded volume term and the electrostatic interaction (Fuchigami et al., 2020). See the reference (Fuchigami et al., 2020) for more details.

All the simulations were performed by CafeMol (Kenzaki et al., 2011). Specifically, the current simulation setup is provided as an example of the flexible-fitting MD simulation in the CafeMol package.

RESULTS

Relationship Between the Tip Radius and the Parameter σ

Before performing the flexible-fitting MD simulations, we quantify the relation between the AFM probe tip radius in the collision detection method and the σ parameter used in the smoothed method. For ten distinct structures of nucleosomes obtained by coarse-grained MD simulations, we first generated pseudo-AFM images by the collision detection method with various tip radii R's, ranging between 0.1 and 3.0 nm (with 0.1 nm increment). For each R value, we searched the optimal σ value with which the smoothed method gives the maximum correlation coefficient (c.c.) with the image generated by the collision detection method (**Figure 2**). We found that, on average, the optimal σ value for each nucleosome structure linearly increases with the tip radius R. In addition, the optimal σ value depends, albeit modestly, on the nucleosome structure. Especially, for larger tip radius, the optimal σ value varies structure by structure more significantly (**Figure 2**). For each tip radius R, we quantified the optimal σ value by its mean $\langle \sigma(R) \rangle$ and the standard deviation $\Delta\sigma(R)$ of the ten samples (**Supplementary Table S1**).

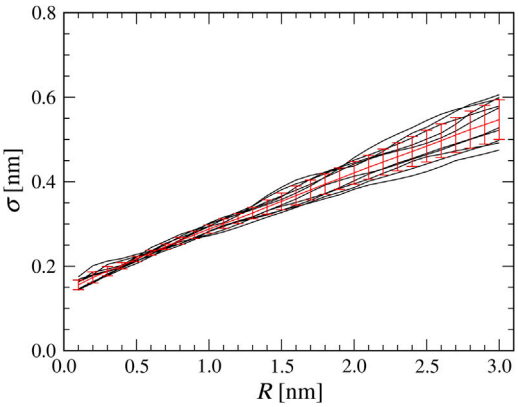


FIGURE 2 | Relationship between a tip radius R in the collision detection method and a parameter σ in the smoothed method. The results from ten structures of nucleosome are shown in black lines, and the mean with the standard deviation for each tip radius R are shown in red.

Using the Bayes' theorem, we can estimate the probability density to have the tip radius R for a given σ as,

$$P(R|\sigma) = \frac{P(\sigma|R)P(R)}{\sum_R P(\sigma|R)P(R)} \quad (3)$$

where we assumed the discrete representation of the tip radius (0.1 nm increment). Without any prior knowledge, we assume $P(R)$ is constant over the range between 0.1 and 3.0 nm. We also assume that $P(\sigma|R)$ is well approximated by the normal distribution with the mean $\langle \sigma(R) \rangle$ and the standard deviation $\Delta\sigma(R)$, i.e., $P(R|\sigma) = N(\langle \sigma(R) \rangle, \Delta\sigma(R)^2) / \sum_R N(\langle \sigma(R) \rangle, \Delta\sigma(R)^2)$. The resulting conditional probability density is given in **Supplementary Table S2**.

Bayesian Statistical Inference of the Tip Radius Using Flexible-Fitting MD Simulation

To examine the Bayesian statistical inference of the tip radius unambiguously, we need a test system for which we know the "correct" tip radius. To this end, in this study, we performed the

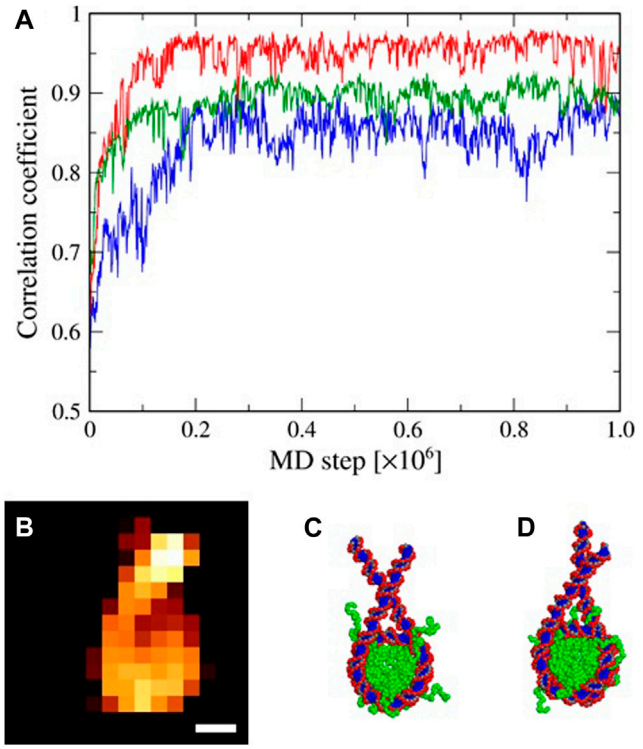


FIGURE 3 | Flexible-fitting MD simulations. **(A)** Representative time series of the correlation coefficient (c.c.) of pseudo-AFM image of the simulated structure with the "experimental" AFM image. Shown here are results from three different σ values, 0.1, 0.3, and 0.5 nm in blue, red, and green, respectively. **(B)** The synthetic-AFM image used as an "experimental" AFM image in the flexible-fitting. Scale bar: 5 nm. **(C)** The nucleosome initial structure. **(D)** The simulated structure with the highest c.c..

so-called twin experiment. We first prepared a reference AFM image of the nucleosome using the collision detection method with the tip radius of 1.0 nm. This AFM image serves as the "experimental AFM image" in this study. Thus, the "correct" tip radius in this study is 1.0 nm.

For the experimental AFM image (**Figure 3B**), starting from a different structure (**Figure 3C**), we performed 10^6-step flexible-fitting MD simulations of the nucleosome system with various values of σ ranging from 0.10 to 0.50 nm (with the 0.01 nm increment) (exemplified in **Figure 3A**). We repeated the same simulations ten times with different random numbers for each value of σ. The time series of the c.c. between the experimental and the simulated AFM images clearly show that the c.c. stochastically increases as a function of time, reaching to a plateau in all the simulations, successfully obtaining well-fitted structures (**Figure 3D**). The averaged plateau value, however, changes depending on the value of σ (**Figure 3A**).

To quantify the difference in the reached correlation coefficient (c.c.), we then estimated the probability density distributions of the c.c. for different σ values using the latter half of the time series data, i.e., totally 5,000 snapshots (representative cases shown in **Figure 4**). For example, a very small σ, such as 0.10 nm, resulted in the distribution with low c.c.'s, suggesting a clearly poor fitting. A rather large σ, such as 0.5 nm, also gave a poor fitting. In between the two limiting cases, there exists an optimal value of σ. Our previous studies indicated that the mean, or the median, of the c.c. distribution is not a good indicator (Niina et al., 2020). Instead, we need to focus on the upper tail of the distribution of c.c.. Among the nine cases in **Figure 4A** (between 0.1 and 0.5 nm, with the 0.05 nm increment), we found σ = 0.3 nm the optimal value.

Looking into a finer scale (0.01 nm increment) shown in **Figure 4B**, we found the highest c.c. value 0.989 at σ = 0.31 nm. From the values of conditional probability density shown in **Supplementary Table S2**, it is found that the most probable value of tip radius R is 1.2 nm, slightly larger than the "correct" tip radius 1.0 nm. This difference between the correct and the estimated tip radii is inevitable and acceptable because AFM image is insensitive to sub-nanometer-order difference in tip radius. Indeed, pseudo-AFM images generated using 1.0 nm-

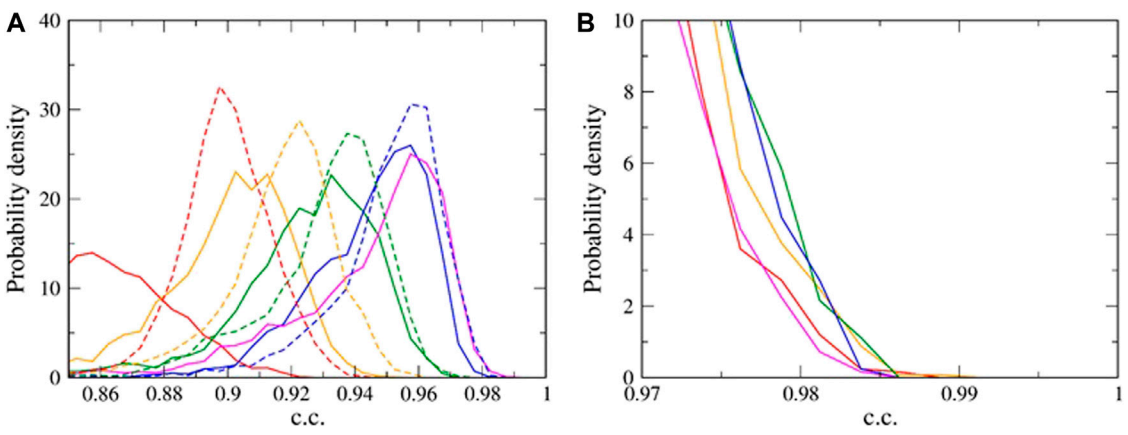


FIGURE 4 | The probability density distributions of the correlation coefficient (c.c.) for representative σ values. **(A)** The overall distribution. The σ in nm unit are 0.1 (red solid), 0.15 (orange solid), 0.20 (green solid), 0.25 (blue solid), 0.30 (magenta solid), 0.35 (blue broken), 0.40 (green broken), 0.45 (orange broken), and 0.50 (red broken). **(B)** A close-up view at a high c.c. range. The σ in nm unit are 0.30 (red), 0.31 (orange), 0.32 (green), 0.33 (blue), and 0.34 (magenta).

radius tip shows a high c.c. of 0.99 or higher with the images generated using tip with radii of 0.7, 0.8, 0.9, 1.1, 1.2, and 1.3 nm.

To test the statistical significance of the estimation of σ value, we performed a simple bootstrap test. We first prepared all patterns of combinations of five samples taken from ten highest c.c. values obtained from ten flexible-fitting MD simulations with the same σ value. The number of combination patterns is ${}_{10}C_5 = 252$. In each combination pattern, the highest c.c. value in the five values was used as the estimated value. In this way, 252 estimated highest c.c. values were obtained for each of seven σ values, $\sigma = 0.28$, 0. 29, 0.30, 0.31, 0.32, 0.33, and 0.34 nm. Then, we performed all-to-all pairwise comparison of 252 estimated highest c.c. values for $\sigma = 0.31$ nm with 252 estimates for the other six σ values. The $\sigma = 0.31$ nm exhibited higher values against for $\sigma = 0.28$, 0. 29, 0.30, 0.32, 0.33, and 0.34 nm with the probabilities 0.956, 0.998, 0.723, 0.677, 0.986, and 0.957, respectively. Thus, the σ value can be estimated as 0.31±0.01 nm. Note that the bootstrap test was performed with five c.c. values, whereas the real estimate used ten c.c. values, making the latter somewhat more reliable.

DISCUSSION

In this brief report, we investigated the statistical inference of the AFM probe tip radius via flexible-fitting MD simulations. First, we statistically characterized the relationship between the tip radius appeared in the collision detection method and the σ value used in the smoothed method. Then, with various σ values, we repeated the flexible-fitting MD simulations to the pre-defined AFM image of a nucleosome structure. Based on the upper tail part of the probability density distribution of the correlation coefficient, we found an optimal value of σ, as well as a well-fitted structure, for the given AFM image. Combining it with the above-mentioned relationship between the tip radius and the σ value, we could infer the probe tip radius, with reasonable accuracy.

As shown in **Figure 2**, a good linear correlation is observed between a tip radius R and a parameter σ. In this case, therefore, the simple linear regression-based method could also work well to infer a tip radius from a σ value.

One possible limitation in the current approach is in the treatment of flexible regions of proteins. In the current test system, the nucleosome contains the histone octamer, which contain rather long disordered tails. Usually, the AFM measurements do not clearly detect the configuration of the histone tails. On the other hand, our collision detection method and the smoothed method detect such flexible regions in the same way as the well-folded regions. While fitting of the well-folded regions is realized in the current simulations, disordered tail configurations were not well fitted, which could cause some discrepancy in the tip radius inference.

While the current approach can infer the tip radius with reasonable accuracy, there can be room for improvement especially in terms of the estimate efficiency. The current approach needs to repeat the flexible-fitting MD with various σ values. This is feasible for the inference of one parameter, and can be accelerated by using an efficient search method, for example, a two-stage searching method with coarse and fine increments. But it is nearly impossible for the case of many unknown parameters. In the context of AFM measurement, there can be other relevant parameters, such as the apex angle of the probe cone. For these cases, some improvement may be necessary. Since inference of the experimental parameters in the flexible-fitting MD is quite a computationally demanding process, we may better perform similar inference within the rigid-body fitting process. In addition, relationship between the tip radius and the σ value might be somewhat system-dependent. It would be confirmed by performing the same analysis using other molecules.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

SF and ST designed the work. SF and TN developed the methodology. SF performed the calculation. SF and ST wrote the original manuscript. All the authors discussed the result and checked the manuscript.

FUNDING

This work was supported mainly by the Japan Science and Technology Agency (JST) grant (JPMJCR1762) (ST). The work is also partly supported by Grant-in-Aid for Scientific Research (C) to SF (Grant Number 19K06598) from Japan Society for the Promotion of Science (JSPS).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.636940/full#supplementary-material.

REFERENCES

Amyot, R., and Flechsig, H. (2020). BioAFMviewer: an interactive interface for simulated AFM scanning of biomolecular structures and dynamics. *PLoS Comput. Biol.* 16, e1008444. doi:10.1371/journal.pcbi.1008444

Ando, T., Kodera, N., Takai, E., Maruyama, D., Saito, K., and Toda, A. (2001). A high-speed atomic force microscope for studying biological macromolecules. *Proc. Natl. Acad. Sci. U.S.A.* 98, 12468–12472. doi:10.1073/pnas.211400898

Ando, T., Uchihashi, T., and Kodera, N. (2013). High-Speed AFM and applications to biomolecular systems. *Annu. Rev. Biophys.* 42, 393–414. doi:10.1146/annurev-biophys-083012-130324

Ando, T., Uchihashi, T., and Scheuring, S. (2014). Filming biomolecular processes by high-speed atomic force microscopy. *Chem. Rev.* 114, 3120–3188. doi:10.1021/cr4003837

Brandani, G. B., Niina, T., Tan, C., and Takada, S. (2018). DNA sliding in nucleosomes via twist defect propagation revealed by molecular simulations. *Nucleic Acids Res.* 46, 2788–2801. doi:10.1093/nar/gky158

Casuso, I., Khao, J., Chami, M., Paul-Gilloteaux, P., Husain, M., Duneau, J. P., et al. (2012). Characterization of the motion of membrane proteins using high-speed atomic force microscopy. *Nat. Nanotechnol.* 7, 525–529. doi:10.1038/nnano.2012.109

Dufrêne, Y. F., Ando, T., Garcia, R., Alsteens, D., Martinez-Martin, D., Engel, A., et al. (2017). Imaging modes of atomic force microscopy for application in molecular and cell biology. *Nat. Nanotechnol.* 12, 295–307. doi:10.1038/nnano.2017.45

Freeman, G. S., Hinckley, D. M., Lequieu, J. P., Whitmer, J. K., and De Pablo, J. J. (2014). Coarse-grained modeling of DNA curvature. *J. Chem. Phys.* 141, 165103. doi:10.1063/1.4897649

Fuchigami, S., Niina, T., and Takada, S. (2020). Particle filter method to integrate high-speed atomic force microscopy measurements with biomolecular simulations. *J. Chem. Theory Comput.* 16, 6609–6619. doi:10.1021/acs.jctc.0c00234

Kenzaki, H., Koga, N., Hori, N., Kanada, R., Li, W., Okazaki, K., et al. (2011). CafeMol: a coarse-grained biomolecular simulator for simulating proteins at work. *J. Chem. Theory Comput.* 7, 1979–1989. doi:10.1021/ct2001045

Kodera, N., Yamamoto, D., Ishikawa, R., and Ando, T. (2010). Video imaging of walking myosin v by high-speed atomic force microscopy. *Nature* 468, 72–76. doi:10.1038/nature09450

Kodera, N., Sakashita, M., and Ando, T. (2006). Dynamic proportional-integral-differential controller for high-speed atomic force microscopy. *Rev. Sci. Instrum.* 77 (2006), 083704. doi:10.1063/1.2336113

Li, W., Wang, Y., and Takada, S. (2014). Energy landscape views for interplays among folding, binding, and allostery of calmodulin domains. *Proc. Natl. Acad. Sci. U.S.A.* 111 (2014), 10550–10555. doi:10.1073/pnas.1402768111

McGreevy, R., Teo, I., Singharoy, A., and Schulten, K. (2016). Advances in the molecular dynamics flexible fitting method for cryo-EM modeling. *Methods* 100, 50–60. doi:10.1016/j.ymeth.2016.01.009

Niina, T., Brandani, G. B., Tan, C., and Takada, S. (2017). Sequence-dependent nucleosome sliding in rotation-coupled and uncoupled modes revealed by molecular simulations. *PLoS Comput. Biol.* 13, e1005880. doi:10.1371/journal.pcbi.1005880

Niina, T., Fuchigami, S., and Takada, S. (2020). Flexible fitting of biomolecular structures to atomic force microscopy images via biased molecular simulations. *J. Chem. Theory Comput.* 16, 1349–1358. doi:10.1021/acs.jctc.9b00991

Orzechowski, M., and Tama, F. (2008). Flexible fitting of high-resolution x-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations. *Biophys. J.* 95, 5692–5705. doi:10.1529/biophysj.108.139451

Shekhar, M., Terashi, G., Gupta, C., Sarkar, D., Debussche, G., Sisco, N. J., et al. (2020). CryoFold: determining protein structures and ensembles from cryo-EM data. *bioRxiv* 687087. doi:10.1101/687087

Singharoy, A., Teo, I., McGreevy, R., Stone, J. E., Zhao, J., and Schulten, K. (2016). Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. *eLife* 5, e16105. doi:10.7554/eLife.16105

Tan, C., and Takada, S. (2020). Nucleosome allostery in pioneer transcription factor binding. *Proc. Natl. Acad. Sci. U.S.A.* 117, 20586–20596. doi:10.1073/pnas.2005500117

Trabuco, L. G., Villa, E., Mitra, K., Frank, J., and Schulten, K. (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* 16, 673–683. doi:10.1016/j.str.2008.03.005

Uchihashi, T., Iino, R., Ando, T., and Noji, H. (2011). High-speed atomic force microscopy reveals rotary catalysis of rotorless F_1-ATPase. *Science* 333, 755–758. doi:10.1126/science.1205510

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 © 2021 Fuchigami, Niina and Takada. 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 Molecular Biosciences

BRIEF RESEARCH REPORT
published: 11 March 2021
doi: 10.3389/fmolb.2021.629773

CUTie2: The Attack of the Cyclic Nucleotide Sensor Clones

Florencia Klein[1,2], *Florencia Sardi*[3], *Matías R. Machado*[1], *Claudia Ortega*[4], *Marcelo A. Comini*[3] *and Sergio Pantano*[1,5]*

[1]BioMolecular Simulation Group, Institut Pasteur de Montevideo, Montevideo, Uruguay, [2]Graduate Program in Chemistry, Facultad de Química, Universidad de La República, Montevideo, Uruguay, [3]Laboratory Redox Biology of Trypanosomes, Institut Pasteur de Montevideo, Montevideo, Uruguay, [4]Recombinant Protein Unit, Institut Pasteur de Montevideo, Montevideo, Uruguay, [5]Shanghai Institute for Advanced Immunochemical Studies, ShanghaiTech University, Shanghai, China

The detection of small molecules in living cells using genetically encoded FRET sensors has revolutionized our understanding of signaling pathways at the sub-cellular level. However, engineering fluorescent proteins and specific binding domains to create new sensors remains challenging because of the difficulties associated with the large size of the polypeptides involved, and their intrinsically huge conformational variability. Indeed, FRET sensors' design still relies on vague structural notions, and trial and error combinations of linkers and protein modules. We recently designed a FRET sensor for the second messenger cAMP named CUTie (**C**yclic nucleotide Universal **T**ag for **i**maging **e**xperiments), which granted sub-micrometer resolution in living cells. Here we apply a combination of sequence/structure analysis to produce a new-generation FRET sensor for the second messenger cGMP based on Protein kinase G I (PKGI), which we named CUTie2. Coarse-grained molecular dynamics simulations achieved an exhaustive sampling of the relevant spatio-temporal coordinates providing a quasi-quantitative prediction of the FRET efficiency, as confirmed by *in vitro* experiments. Moreover, biochemical characterization showed that the cGMP binding module maintains virtually the same affinity and selectivity for its ligand thant the full-length protein. The computational approach proposed here is easily generalizable to other allosteric protein modules, providing a cost effective-strategy for the custom design of FRET sensors.

Keywords: CUTie, biosensor, FRET, molecular dynamics, coarse-grained, Sirah, cyclic nucleotide, signaling

OPEN ACCESS

Edited by:
Fabio Trovato,
Freie Universität Berlin, Germany

Reviewed by:
Kjetil Wessel Andressen,
University of Oslo, Norway
Francesco Tavanti,
University of Modena and Reggio Emilia, Italy
Laura Orellana,
Karolinska Institutet (KI), Sweden

***Correspondence:**
Sergio Pantano
spantano@pasteur.edu.uy

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *15 November 2020*
Accepted: *20 January 2021*
Published: *11 March 2021*

Citation:
Klein F, Sardi F, Machado MR, Ortega C, Comini MA and Pantano S (2021) CUTie2: The Attack of the Cyclic Nucleotide Sensor Clones. Front. Mol. Biosci. 8:629773. doi: 10.3389/fmolb.2021.629773

INTRODUCTION

The development of genetically encoded sensors based on Föster resonance energy transfer (FRET) effect is a well-established methodology for the non-invasive and real-time study of a plethora of cellular events (Nikolaev and Lohse, 2006; Berrera et al., 2008; Meng and Sachs, 2011; Calamera et al., 2019).

In general, FRET sensors are composed of a molecular detector, which undergoes a conformational change upon a given signal. A couple of fluorophores suitable for FRET and linked to convenient domains of the detector module complete the sensor architecture. Two fluorophores are a suited FRET pair if there is a substantial overlap between the spectra of emission and absorption of the donor and acceptor, respectively (Lindenburg and Merkx, 2014). The use of spectral variants of the Green Fluorescent Protein (GFP) as FRET pairs opened the possibility to generate genetically encoded sensors for monitoring intracellular signaling (Zaccolo

et al., 2000). In particular, FRET sensors applied to the study of cyclic nucleotides (CNs) have been instrumental for dissecting molecular details of the corresponding signaling pathways, and several protein architectures have been reported to work as FRET sensors for CNs (Sprenger and Nikolaev, 2013; Pendin et al., 2016).

In the case of CNs, most often the detector is a cyclic nucleotide-binding domain (CNBD) genetically fused to suitable spectral variants of GFP. When the intracellular concentration of cyclic nucleotides rises, they bind to the CNBD, triggering an allosteric conformational change that modifies both fluorophores' relative distance and/or orientation, translating into a change of the FRET signal (Lindenburg and Merkx, 2014). Despite the proven usefulness of genetically encoded sensors, very little work has been done to advance on methods to systematically engineer FRET sensors achieving quantitative predictions of FRET efficiency (Stangherlin et al., 2014). Indeed, the design of novel sensors is guided by a limited structural knowledge of the CNBDs. Such a procedure often results in an undetermined and frustrating number of trial and error attempts.

We were among the firsts to apply modeling techniques to improve cAMP FRET sensors (Lissandron et al., 2005) and provide simple structure-based rules to guide the design of CN sensors (Machado and Pantano, 2015). More recently, we showed that the combination of modeling techniques with coarse-grained (CG) simulations effectively led to a quasi-quantitative prediction of the FRET change upon raising cAMP concentrations. Moreover, the rational design of the so-called CUTie sensor produced for the first time a FRET sensor targetable to different subcellular compartments that unveiled the presence of unexpectedly small cellular (nano)domains regulated by cAMP (Chao et al., 2019). Quantitative predictions of the FRET efficiency of the CUTie sensor were achieved thanks to the comprehensive sampling provided by CG simulations. As confirmed experimentally, our cost-effective approach achieved an exhaustive conformational sampling of the space accessible to the fluorescent proteins linked to the cAMP-binding domain of the regulatory subunit of the RIIβ isoform of Protein kinase A (Surdo et al., 2017). Here, we aimed to prove the general validity of our approach by designing and testing a FRET sensor for cyclic guanosine 3′-5′-monophosphate (cGMP). This latest generation sensor will be referred as CUTie2 hereafter.

The second messenger cGMP is involved in several downstream intracellular signaling pathways. Among others, cGMP plays a central role in regulating the cardiomyocyte function (Perera et al., 2015; Bork et al., 2019; Zhao et al., 2020) and is involved in neurological diseases (Knott et al., 2017; Schepers et al., 2019). In mammalian cells, the production of cGMP is regulated by the binding of Nitric oxide (NO) to soluble guanylyl cyclases and natriuretic peptides, that activate particulate guanylyl cyclases GC-A and GC-B; while, its degradation is tightly controlled by eight out of the eleven isoforms of Phosphodiesterases (Friebe et al., 2020).

The development of cGMP sensors poses the challenge of achieving simultaneously high affinity for cGMP vs. cAMP, which is normally present at higher concentrations in overlapping subcellular compartments (Stangherlin and Zaccolo, 2012). Nevertheless, a number of cGMP sensors have been reported (Sprenger and Nikolaev, 2013). The most frequently used sensors are based on mammalian proteins, which, due to their relatively high EC_{50} (Cygnet2.1 and cGI-500 EC_{50} = 0.5 μM, cGES-DE5 EC_{50} = 1.9 μM) (Honda et al., 2001; Nikolaev et al., 2006; Russwurm et al., 2007) they are useful

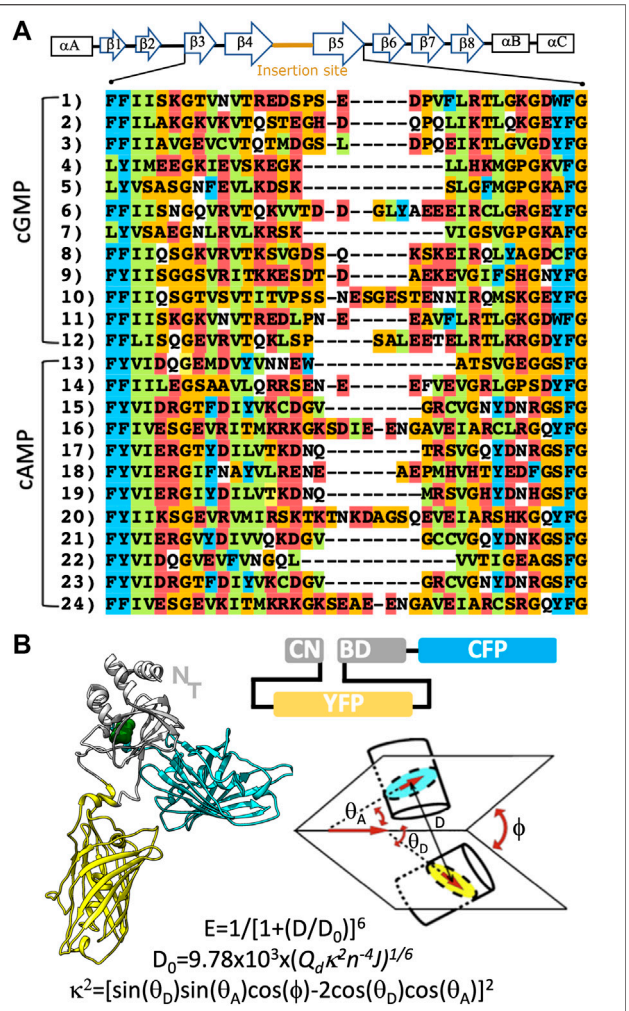


FIGURE 1 | Design of CUTie2. **(A)** Global secondary structure of a CNBD and multiple sequence alignment of related domains. Rectangles and arrows are used to indicate α-helices and β-strands. Sequences 1 to 12, and 13 to 24 correspond to cGMP and cAMP binding proteins, respectively. These correspond to the UNIPROT codes, for the cGMP binding domains: 1) Q13976, 2) Q13237, 3) A0A444U9Q8, 4) A0A443SJC3, 5) B0X970, 6) A0A4Y2N1F1, 7) A0A4Y2BIM3, 8) A0A419PY19, 9) A0A2P8Y1R0, 10) A0A3R7JPE0, 11) A0A444UAH2, 12) A0A0M4ENX8 and the cAMP binding modules: 13 and 14) P00514, 15 and 16) P12369, 17) A0A2J8TG41, 18) E0VRT6, 19 and 20) A0A091CJC2, 21) A0A556UFG6, 22) A9QQ52, 23 and 24) A0A384A0R8. Aminoacids are colored by physicochemical character according to CLUSTAL. **(B)** Cartoon representation and domain organization of the CUTie2 sensor. Notice that the N-terminal of the protein remains free to be eventually fused to arbitrary targeting domains. The schematic architecture of the sensor and definition of the geometric factors and equations used to calculate the FRET signal are provided on the right bottom.

in environments with high concentrations of cGMP. So far, the sensor with higher affinity for cGMP has an EC50. ~ 40 nM (Niino et al., 2009; Calamera et al., 2019).

Therefore, it is highly desirable to count with a targetable sensor with similar characteristics to those available for cAMP.

The design of allosteric sensors requires the knowledge of the structure and conformational landscape accessible to the protein. Schematically, the binding site of CNBDs is constituted by an 8-stranded β-barrel, which is flanked by two helices commonly designated as helix A and B (**Figure 1A**). In most CNBDs, a C-terminal helix (helix C) acts as a lid that closes the binding pocket in the presence of CNs (Berman et al., 2005). In absence of the ligand, the C helix is highly mobile to the point it can't be solved by X-ray crystallography. This conformational behavior makes this family of protein structures good candidates for FRET sensors, and their C-terminal an obvious place to insert a fluorescent protein. During the design of the original CUTie sensor, we found that the loop connecting β-strands 4-5 (named 4-5 loop) is poorly conserved, with frequent insertions in different cAMP binding domains (see **Figure 1A**). This suggested that the 4-5 loop is structurally permissive to insert a fluorescent molecule without altering CNBD´s characteristics (Surdo et al., 2017). Therefore, we further explore the possibility of engineering another CNBD to generate a FRET sensor for cGMP. With this aim, selected on the human Protein kinase G (PKG), which has two isoforms (I and II) (Butt et al., 1993; Hofmann et al., 2009), sharing the same overall architecture and CNBD organization (Vaandrager et al., 2005; Campbell et al., 2016).

In-vitro experiments showed that the CUTie2 sensor developed in-silico maintains a nearly native affinity (EC_{50} = 277 nM) and over two orders of magnitude selectivity for cGMP vs. cAMP, reaching a maximal FRET change of 26%. Since CUTie2 maintains the same architecture of its cognate CUTie, it might offer the possibility to dissecting simultaneously the submicrometric coexistence of cAMP and cGMP signals. Moreover, because of the global similarity of CNBDs, the procedure illustrated here should be generalizable to arbitrary proteins of this class.

METHODS

In Silico Design

To find a suitable insertion point for the first fluorescent module, we perform a multiple sequence alignment of protein containing CNBDs. To ensure the generality of our results, we consider sequences from mammals, insect, fish, arachnid, crustacean and annelid (**Figure 1A**). CNBD protein sequences are retrieved from the UniProt database (Bateman, 2019, https://www.uniprot.org) and aligned using ClustalW in the MPI Bioinformatics Toolkit (Bioinformatics Toolkit, 2020).

3D starting conformations are built on the base of X-ray structures of PKGIβ protein (PDBid: 4KU7, residues K223 to Y351 (Huang et al., 2014)). Because of practical reasons, both fluorescent modules (Yellow and Cyan Fluorescent Proteins) were modeled using the X-ray structure PDBid: 1QYO (Barondeau et al., 2003). This corresponds to a triple mutation of GFP, without the chromophore moiety. This variant presents the typical GFP folding and obviates the need to parameterize CG chromophores. This shortcut is well tolerated within the coarseness of our approach. Initially, the CNBD, linkers and the fluorescent module corresponding to the YFP were manually positioned at distances compatible with covalent bonds and ligated with Chimera (Pettersen et al., 2004). Subsequently, the CFP module is attached to Y351 in the CNBD. Once a model is complete, its structure is energy optimized by CG simulations (see below). Different initial conformers are generated by arbitrarily rotating torsional angles between the CNBD and the CFP module. For simulations in the cGMP-free protein, a new set of conformers are generated in which the ligand is removed before starting the molecular dynamics (MD) simulations.

CG Simulations

CG MD simulations are performed using the SIRAH 2.0 force field under the conditions reported in (Machado et al., 2019). The interaction parameters for CG cGMP and the simulation protocol are reported in supplementary material (**Supplementary Figure S1**). Briefly, atomistic models are mapped to CG using SIRAH Tools (Machado and Pantano, 2016) and solvated with CG water up to 2 nm beyond the last bead in each direction. Explicit CG ions are added to set an ionic strength of 15 mM. Although simulation boxes are different for each protomer, systems typically contained 50,000 beads, roughly representing half a million atoms. The simulation protocol consists of 5,000 steps of unrestrained energy minimization, and five MD equilibration steps in NPT ensemble at 1 bar using the v-rescale thermostat and Parrinello-Rahman barostat:

(1) Solvent equilibration by 5 ns at 10 K and time step of 2.0 fs.
(2) 40 ps at 300 K and time step of 2.0 fs.
(3) 400 ps at 300 K with a time step of 20 fs.
(4) 400 ps at 300 K with a time step 20 fs and positional restrains of 1,000 kJ mol-1nm-2 on the whole protein.
(5) Same as the previous one but for 5 ns.
(6) Production simulations for 10 μs.

Each equilibration step starts from 0 K temperature, (i.e. velocities are zeroed at each restart). We generated 16 different initial conformers in cGMP-bound and eight in cGMP-free conformation, totalizing a cumulative simulation time of 0.24 ms. Snapshots are recorded every 100 ps for analysis.

Non-bonded interactions are calculated with a 1.2 nm cutoff and PME method (Darden et al., 1993; Essmann et al., 1995) for long-range electrostatics. Simulations are performed using GROMACS 4.6.7 (http://www.gromacs.org).

Analysis of the Trajectories

The FRET efficiency (E) depends only on the inter-fluorophore distance (D) and the relative orientation of the chromophores (commonly named orientational factor, or simply κ^2). The change in FRET is calculated by subtracting the average FRET values in cGMP-bound and cGMP-free conformations. D is estimated as the distance between the geometric centers of

CFP and YFP, while κ^2 factor is calculated from the dipoles defined between the geometric center of the fluorescent protein and the Cα of Ser147 in each fluorescent module (**Figure 1B**). This agrees with the orientation of the dipole moment estimated from DFT calculations for a variety of chromophores in different fluorescent proteins (Rosell and Boxer, 2003). Root Mean Square Deviations (RMSD) are calculated on the Cα beads.

The FRET efficiencies in cGMP-bound and -free states are calculated as the running average with sliding windows of 100 ns over the concatenated trajectories. Obviously, the final averages are independent of the order in which the trajectories are concatenated. The standard deviations in the simulated FRET are calculated after convergence, (i.e. after a cumulative time = 0.04 ms).

EXPERIMENTAL

All chemicals and reagents used are of analytical grade and purchased from SIGMA. The DNA plasmid is synthetized by Genscript and sequenced by the Molecular Biology Unit, Institut Pasteur of Montevideo.

The recombinant form of CUTie2 is expressed as an N-terminally six His-tagged protein using the vector pET28a (+) and *Escherichia coli* Rosetta (DE3) as heterologous expression host. The transformed bacteria are grown at 37°C in 2 YT medium containing 50 μg/ml Kanamycin and 34 μg/ml Chloramphenicol until an optical density of 0.8. Thereafter, 0.5 mM IPTG (Euromedex) is added to the culture and incubation extended for 16 h at 20°C. After centrifugation at 6,000 *g* for 30 min at 4°C, the cell pellet is resuspended in 50 mM Tris-HCl pH 8.0, 300 mM NaCl, 20 mM imidazole (Buffer A), containing a EDTA-free protease inhibitors cocktail (ROCHE) and Lysozyme (end concentration 1 mg/ml). The cell suspension is subjected to sonication and the debris removed by centrifugation at 16,000 *g* for 1 h at 4°C. The clarified extract is applied to a 1 ml HisTrap column (GE Healthcare) and upon washing in Buffer A, the His-tagged protein is eluted with buffer A containing 500 mM imidazole. The fractions containing the cGMP sensor protein are pooled and concentrated by centrifugation (5,000 *g* at 4°C) in a Vivaspin-20 filter (30 kDa cutoff), and further polished by a Superdex G-200 size exclusion chromatography (10/30 column) run in 50 mM Tris-HCl (pH 8.0), 500 mM NaCl.

Protein concentration is measured at 280 nm, where ε280 for CUTie2 60.990 M^{-1} cm^{-1}. In a Hellma® fluorescence cuvette (total reaction volume of 120 μL), the sensor (100 nM) is treated with different concentrations of cGMP or cAMP (0 nM–10 mM) in PBS buffer containing 1 mM EDTA pH 7.0. Upon addition of cGMP or cAMP, the fluorescence emission spectrum (λexc = 435 nm, λem = 460–600 nm) of CUTie2 is recorded in a Cary Eclipse fluorimeter every 1 min during 10 min and using slit widths of 5 nm and a PMT of 780. The fluorescence intensity (FI) values corresponding to CFP (485 nm) and YFP (527 nm) in the absence (C0) or presence of different concentrations of the corresponding cyclic nucleotides (Cx) are used to calculate the percentage FRET as follows:

$$\text{FRET} = \left[\left(\frac{FI_{Cx}^{527nm}}{FI_{Cx}^{485nm}}\right) - \left(\frac{FI_{C0}^{527nm}}{FI_{C0}^{485nm}}\right)\right] \times 100 \quad . \qquad (1)$$

The % FRET *vs.* Log [cNMP] is plotted, fitted to non-linear regression equations and statistically analyzed using the GraphPad Prism six Software (San Diego, CA, United States). Three independent titration experiments are performed with three different batches of the recombinant sensor.

RESULTS AND DISCUSSION

In Silico Design

In order to define the structure of the cGMP sensor, we aligned the protein sequences reported in **Figure 1A**. We identify the 4-5 loop as a suitable insertion site in cGMP binding domains. The second fluorescent module is added to the C-terminal of the protein, after the end of the C-helix. The scheme of the construct and a molecular representation of CUTie2 are shown in **Figure 1B**.

The module corresponding to YFP is inserted within the cGMP-CNBD. It substitutes the original amino acids Asp_{286}-Ser-Pro-Ser-Glu_{282} in the 4-5 loop by the linkers used in the CUTie sensor, as previous MD simulations suggested they avoid spurious distortions in the binding domain (Surdo et al., 2017). The substitution of the original amino acids in the sequence is not expected to alter significantly the conformation of the CNBD as they are not determined in the X-ray structure. This assumption could be verified *a posteriori* (see below). The CFP is linked directly to the C-terminal of the CNBD. As a proof of concept of our methodology, we choose the domain β of human PKG I, for which the 3D structure and cGMP binding characteristics are well defined (Huang et al., 2014). A schematic representation of the CUTie2 sensor and its primary sequence are shown in **Figure 1B**, and **Supplementary Figure S2**, respectively.

Having the primary sequence and structural knowledge of the protein modules, it is possible to construct 3D representations of the sensor. However, a quantitative prediction of the FRET efficiency requires a complete sampling of all relevant collective variables. In this regard, it may be important to recall that the FRET effect is a dipole-dipole interaction, which rapidly decays as the $1/D^6$ (**Figure 1B**), where D is the distance between chromophores. In practical terms, 10 nm is considered as an upper limit for the distance of interaction between two fluorescent proteins. On the other hand, the numerical value of κ^2 ranges from 0 to 4 (see definition in **Figure 1B**). Therefore, a proper sampling of the relevant coordinates to quantify the FRET effect must achieve a complete sampling of D and κ^2 within the above range. In principle, MD simulations can be used to this aim. However, the computational cost of such simulations is prohibitive. Hence, we sorted out to use the accurate and topologically unbiased SIRAH force field for CG simulations (Machado et al., 2019). Different starting conformers are

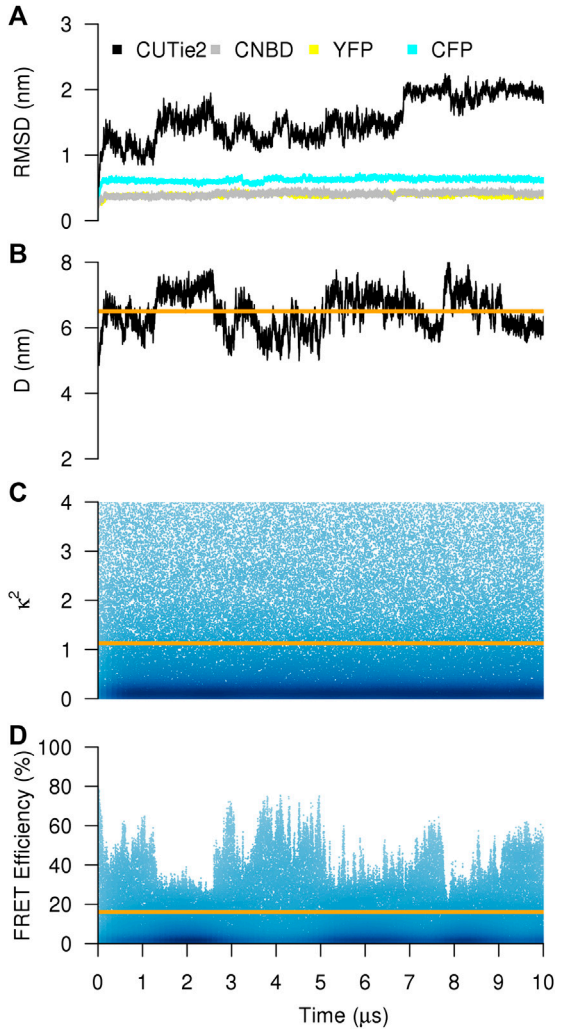


FIGURE 2 | CG simulation of the CUTie2 sensor. **(A)** Instantaneous RMSD of one randomly chosen conformer in cGMP-bound conformation. Black, gray, yellow, and cyan correspond to the whole molecule, the CNBD, the YFP, and the CFP modules, respectively. **(B–D)** Distance between chromophores, κ^2, and FRET efficiency. The orange line in each panel corresponds to the average over the entire trajectory.

prepared in the cGMP-bound and -free states and simulated for 10 μs each.

Figure 2 shows the main determinants of a simulation of cGMP-bound to CUTie2. The relative mobility of each protein module (CNBD and the two fluorescent variants), and the whole construct is characterized in terms of their RMSD from their initial positions. Individual domains show quite stable and relatively low RMSD values. However, the CUTie2 sensor as a whole shows excursions up to 2 nm with stability regions in time-windows of about 2 μs (**Figure 2A**). Those flat RMSD regions show a rough correspondence to the distance between chromophores (**Figure 2B**), indicating that the stable regions in the global RMSD can be identified with large-scale movements of the fluorescent modules. Indeed, the distance between chromophores vary from 5 nm up to 8 nm in this particular staring conformer. Calculation of κ^2 shows that the simulation samples the complete spectrum of values with a much faster dynamics (**Figure 2C**). The reason for this difference is rather obvious, while significant variations in D imply the motion of entire protein modules, the dynamics of κ^2 is related to rotations of torsional dihedrals associated to the protein backbone, which happen spontaneously in the nanosecond timescale at room temperature. The combination of both collective variables as detailed in **Figure 1B** allows for a theoretical estimation of the FRET illustrated in **Figure 2D**. Because of the marked difference in the dynamics of D and κ^2, the FRET efficiency is mostly correlated to the inverse of D.

Experimentally, the FRET efficiency averages on a wide number of conformations and within time windows of minutes. Intuitively, a proper comparison between simulated and experimental FRET efficiencies requires performing simulations of the sensor in cGMP-bound and–free states, calculating the average FRET efficiencies in both states and, then subtract both values to obtain the variation in FRET efficiency upon cGMP binding. For such comparison to be meaningful, one should assume a broad sampling of the relevant variables, namely, D and κ^2. While **Figure 2B** suggests that this is true for κ^2, D is still limited to a restricted range (**Figure 2C**).

Therefore, we generate two sets of simulations with different starting conformers in cGMP-bound and–free states. After each 10 μs simulation, we concatenate the trajectories and calculate the running average of the FRET efficiency (**Figure 3A**). New conformers are generated until convergence is reached. In the cGMP-bound case we generate a total of 16 different conformers for a cumulative time of 0.16 ms (see **Supplementary Figure S3**). For the cGMP-free case, only eight conformers are deemed necessary since in the absence of cGMP the two fluorophores sample preferentially higher D distances (see **Supplementary Figure S4**).

The average change in FRET efficiency resulted in 22% (s.d. 2, **Figure 3A**).

To further illustrate the conformational sampling we plot the FRET efficiencies of the different starting conformers as a function of D and κ^2 to provide a breakdown of the sampling achieved (**Figure 3B**, only the cGMP-bound case is shown for brevity). The whole set of points covers the range $D < 10$ nm and $0 > \kappa^2 > 4$. The spreading of the distribution clearly shows that some conformers remain trapped in relatively narrow conformational regions, while others can widely explore the D variable. Using the backmapping capability of the SIRAH force field (Machado and Pantano, 2016), we produce pseudo atomistic representative snapshots from three different starting conformations. The violet molecule (indicated with the letter "a" in **Figure 3B**) remains blocked in a high FRET configuration, while the red molecule (indicated with "b" in **Figure 3B**) explores a much wider range. In contrast, the cyan molecule (indicated with "c" in **Figure 3B**) spends most of the time with the two fluorescent modules far apart from each other, making little contributions to the global FRET. It is also interesting to notice that, because of the topology of the sensor, there is a region of values of D between 3 and 4 nm, which is structurally forbidden.

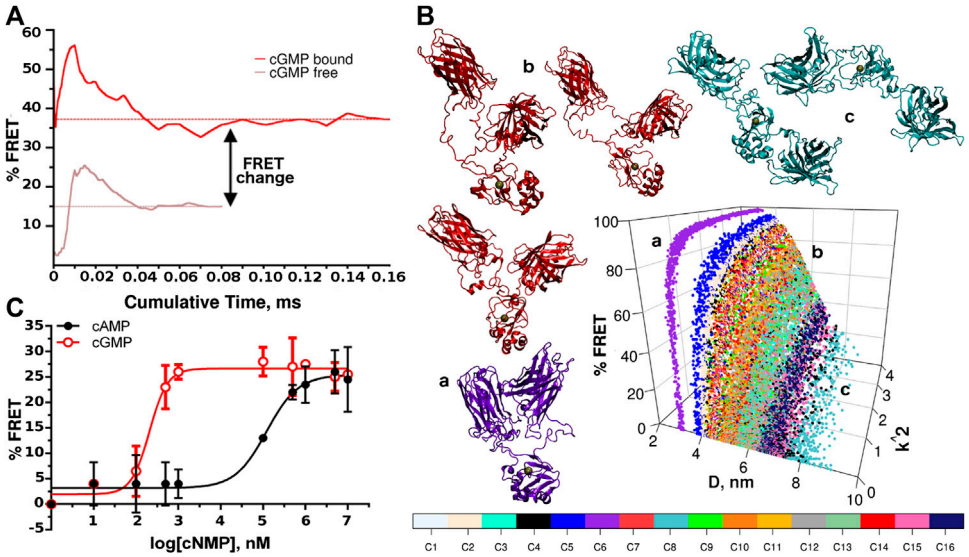


FIGURE 3 | FRET efficiency and experimental validation. **(A)** Averaged FRET values calculated from different conformers vs. simulated time for cGMP-bound and -free sets of simulations. **(B)** Assessment of the completeness of the conformational sampling for the cGMP-bound simulations. The FRET efficiency is reported as a function of D and κ^2 for each of the conformers (indicated by different colors). Six particular conformers taken from three trajectories are shown in cartoon representation to illustrate the conformational dispersion achieved during the simulations. Gray spheres indicate the cGMP binding sites. The letters in lower case indicate the position of the corresponding conformer in the graph. **(C)** Representative concentration-response plot for recombinant CUTie2 titrated with different concentrations of cGMP and cAMP (0–10 mM). The data were fitted to the Boltzmann equation ($R^2 \geq 0.98$).

Experimental Validation

In order to provide experimental confirmation of the *in silico* predictions, the recombinant form of the CUTie2 sensor was expressed and the FRET signal measured *in vitro* for cGMP and cAMP in three independent experiments (see **Supplementary Figure S5**).

As shown in **Figure 3C**, the maximal FRET efficiency obtained at saturating concentrations of cGMP (FRET ≥26%) compares well with the computational predictions. Moreover, the effective concentration of cGMP that triggers a 50% FRET response of CUTie2 (EC_{50}) to cGMP was 277 ± 60 nM, which nicely matches the value reported for this CNBD by using competition fluorescence polarization (Huang et al., 2014). Interestingly, the Hill´s slope estimated from the three independent cGMP titration assays with CUTie2 is 1.5 ± 0.4, which is within the range of values reported for PKG I (Huang et al., 2014) and a related FRET sensor (Honda et al., 2001), and may suggest a certain cooperative effect upon ligand binding. As mentioned before, the selectivity of the sensor for cGMP is essential for *in vivo* applications since in many subcellular compartments the cAMP concentration is higher than that of cGMP (Stangherlin and Zaccolo, 2012). In this regard, CUTie2 titration with cAMP revealed an EC_{50} of 121 ± 11 μM, which shows that the sensor has a >400-fold higher selectivity for cGMP.

DISCUSSION AND CONCLUSION

This manuscript presents a computational methodology for the design and quasi-quantitative prediction of FRET efficiency of genetically encodable sensors for cGMP. The agreement with experimental data obtained here, and with the previously published CUTie sensor for cAMP (Surdo et al., 2017) supports the general validity and transferability of our approach. Within the standard deviations of FRET changes obtained by the experimental and computational determinations, our approach could be considered quasiquantitatively accurate. The capacity of both fluorescent modules to produce FRET at cGMP concentrations comparable to those measured for the intact protein strongly suggests that both the fluorescent proteins and the CNBD adopt their expected folding. This fact is not entirely obvious if we consider that YFP is inserted within the folding of the CNBD. In this regard, it may be worth mentioning that peptide linkers are present between the CNBD and YFP sequence (**Supplementary Figure S2**). These linkers are the same used in the CUTie sensor. The marked differences in sequence conservation in the region of the 4-5 loop shown in **Figure 1A** and the correct functioning of CUTie and CUTie2 strongly suggest that these linkers are well suited for inserting a fluorescent module within the 4-5 loop of any CNBD. The equivalence in cGMP affinity and cAMP selectivity in relation to the intact CNBD further highlight this fact. This opens the possibility to generate a portfolio of sensors with different ligand affinities based on other biochemically well-characterized CNBDs or other ligand-specific domains that undergo a similar allosteric-dependent conformational change. CUTie and CUTie2 share the common feature of having a free N-terminal that can be fused to arbitrary targeting proteins/sequences. Namely, both can be regarded as FRET tags that can be targeted to different subcellular localizations determined by the accompanying protein or targeting signal. In summary, this new sensor holds

great potential to unravel simultaneous signaling pathways with submicrometric resolution and high cyclic nucleotide selectivity in living cells.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

FK and MRM performed and analyzed simulations and prepared figures. FS and CO performed and analyzed experiments. MAC and SP analyzed the data. All authors participated in the writing of the manuscript and approved it.

FUNDING

This work was not funded by any specific grant. It was supported by general intramural financial resources and Ph.D. fellowships to some of the authors. This work was partially supported by FOCEM (MERCOSUR Structural Convergence Fund), COF 03/11.

ACKNOWLEDGMENTS

FK and FS acknowledge the support of Agencia Nacional de Investigación e Innovación (ANII), Uruguay (PhD fellowships POS_NAC_2018_1_151183 and POS_FMV_2015_1_1005325, respectively). FK acknowledge the Graduate Program in Chemistry (PEDECIBA), Facultad de Química, Universidad de la República, Uruguay. SP, MRM, CO, and MAC belong to the Uruguayan National System of Researchers (SNI) from ANII. The Tesla K40 used for this research was 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/fmolb.2021.629773/full#supplementary-material.

REFERENCES

Barondeau, D. P., Putnam, C. D., Kassmann, C. J., Tainer, J. A., and Getzoff, E. D. (2003). Mechanism and energetics of green fluorescent protein chromophore synthesis revealed by trapped intermediate structures. *Proc. Natl. Acad. Sci. United States* 100, 12111–12116. doi:10.1073/pnas.2133463100

Bateman, A. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi:10.1093/nar/gky1049

Berman, H. M., Ten Eyck, L. F., Goodsell, D. S., Haste, N. M., Kornev, A., and Taylor, S. S. (2005). The cAMP binding domain: an ancient signaling module. *Proc. Natl. Acad. Sci. United States* 102, 45–50. doi:10.1073/pnas.0408579102

Berrera, M., Dodoni, G., Monterisi, S., Pertegato, V., Zamparo, I., and Zaccolo, M. (2008). A Toolkit for real-time detection of cAMP: insights into compartmentalized signaling. *Handb. Exp. Pharmacol.* 2008, 285–298. doi:10.1007/978-3-540-72843-6_12

Bioinformatics Toolkit (2020). *Bioinformatics Toolkit*. Available at: Accessed: https://toolkit.tuebingen.mpg.de/. (August 23, 2020).

Bork, N. I., Molina, C. E., and Nikolaev, V. O. (2019). CGMP signalling in cardiomyocyte microdomains. *Biochem. Soc. Trans.* 47, 1327–1339. doi:10.1042/BST20190225

Butt, E., Geiger, J., Jarchau, T., Lohmann, S. M., and Walter, U. (1993). The cGMP-dependent protein kinase--gene, protein, and function. *Neurochem. Res.* 18, 27–42. doi:10.1007/BF00966920

Calamera, G., Li, D., Ulsund, A. H., Kim, J. J., Neely, O. C., Moltzau, L. R., et al. (2019). FRET-based cyclic GMP biosensors measure low cGMP concentrations in cardiomyocytes and neurons. *Commun. Biol.* 2, 1–12. doi:10.1038/s42003-019-0641-x

Campbell, J. C., Kim, J. J., Li, K. Y., Huang, G. Y., Reger, A. S., Matsuda, S., et al. (2016). Structural basis of cyclic nucleotide selectivity in cGMP-dependent protein kinase II. *J. Biol. Chem.* 291, 5623–5633. doi:10.1074/jbc.M115.691303

Chao, Y. C., Surdo, N. C., Pantano, S., and Zaccolo, M. (2019). Imaging cAMP nanodomains in the heart. *Biochem. Soc. Trans.* 47, 1383–1392. doi:10.1042/BST20190245

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

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

Friebe, A., Sandner, P., and Schmidtko, A. (2020). cGMP: a unique 2nd messenger molecule - recent developments in cGMP research and development. *Naunyn. Schmiedebergs. Arch. Pharmacol.* 393, 287–302. doi:10.1007/s00210-019-01779-z

Hofmann, F., Bernhard, D., Lukowski, R., and Weinmeister, P. (2009). CGMP regulated protein Kinases (cGK). *Handb. Exp. Pharmacol.* 191, 137–162. doi:10.1007/978-3-540-68964-5_8

Honda, A., Adams, S. R., Sawyer, C. L., Lev-Ram, V., Tsien, R. Y., and Dostmann, W. R. (2001). Spatiotemporal dynamics of guanosine 3′,5′-cyclic monophosphate revealed by a genetically encoded, fluorescent indicator. *Proc. Natl. Acad. Sci. United States* 98, 2437–2442. doi:10.1073/pnas.051631298

Huang, G. Y., Kim, J. J., Reger, A. S., Lorenz, R., Moon, E. W., Zhao, C., et al. (2014). Structural basis for cyclic-nucleotide selectivity and cGMP-selective activation of PKG i. *Structure* 22, 116–124. doi:10.1016/j.str.2013.09.021

Knott, E. P., Assi, M., Rao, S. N., Ghosh, M., and Pearse, D. D. (2017). Phosphodiesterase inhibitors as a therapeutic approach to neuroprotection and repair. *Int. J. Mol. Sci.* 18, 696. doi:10.3390/ijms18040696

Lindenburg, L., and Merkx, M. (2014). Engineering genetically encoded FRET sensors. *Sensors (Basel)* 14, 11691–11713. doi:10.3390/s140711691

Lissandron, V., Terrin, A., Collini, M., D'alfonso, L., Chirico, G., Pantano, S., et al. (2005). Improvement of a FRET-based indicator for cAMP by linker design and stabilization of donor-acceptor interaction. *J. Mol. Biol.* 354, 546–555. doi:10.1016/j.jmb.2005.09.089

Machado, M., and Pantano, S. (2015). Structure-based, in silico approaches for the development of novel cAMP FRET reporters. *Methods Mol. Biol.* 1294, 41–58. doi:10.1007/978-1-4939-2537-7_4

Machado, M. R., and Pantano, S. (2016). SIRAH tools: mapping, backmapping and visualization of coarse-grained models. *Bioinformatics* 32, 1568–1570. doi:10.1093/bioinformatics/btw020

Machado, M. R., Barrera, E. E., Klein, F., Sónora, M., Silva, S., and Pantano, S. (2019). The SIRAH 2.0 force field: altius, fortius, citius. *J. Chem. Theor. Comput.* 15, 2719–2733. doi:10.1021/acs.jctc.9b00006

Meng, F., and Sachs, F. (2011). Visualizing dynamic cytoplasmic forces with a compliance-matched FRET sensor. *J. Cel Sci.* 124, 261–269. doi:10.1242/jcs.071928

Niino, Y., Hotta, K., and Oka, K. (2009). Simultaneous live cell imaging using dual FRET sensors with a single excitation light. *PLoS One* 4, e6036. doi:10.1371/journal.pone.0006036

Nikolaev, V. O., Gambaryan, S., and Lohse, M. J. (2006). Fluorescent sensors for rapid monitoring of intracellular cGMP. *Nat. Methods* 3, 23–25. doi:10.1038/nmeth816

Nikolaev, V. O., and Lohse, M. J. (2006). Monitoring of cAMP synthesis and degradation in living cells. *Physiology (Bethesda)* 21, 86–92. doi:10.1152/physiol.00057.2005

Pendin, D., Greotti, E., Lefkimmiatis, K., and Pozzan, T. (2016). Exploring cells with targeted biosensors. *J. Gen. Physiol.* 149, 1–36. doi:10.1085/jgp.201611654

Perera, R. K., Sprenger, J. U., Steinbrecher, J. H., Hübscher, D., Lehnart, S. E., Abesser, M., et al. (2015). Microdomain switch of cGMP-regulated phosphodiesterases leads to ANP-induced augmentation of β-adrenoceptor-stimulated contractility in early cardiac hypertrophy. *Circ. Res.* 116, 1304–1311. doi:10.1161/CIRCRESAHA.116.306082

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

Rosell, F. I., and Boxer, S. G. (2003). Polarized absorption spectra of green fluorescent protein single crystals: transition dipole moment directions. *Biochemistry* 42, 177–183. doi:10.1021/bi0266535

Russwurm, M., Mullershausen, F., Friebe, A., Jäger, R., Russwurm, C., and Koesling, D. (2007). Design of fluorescence resonance energy transfer (FRET)-based cGMP indicators: a systematic approach. *Biochem. J.* 407, 69–77. doi:10.1042/BJ20070348

Schepers, M., Tiane, A., Paes, D., Sanchez, S., Rombaut, B., Piccart, E., et al. (2019). Targeting phosphodiesterases-towards a tailor-made approach in multiple sclerosis treatment. *Front. Immunol.* 10, 1727. doi:10.3389/fimmu.2019.01727

Sprenger, J., and Nikolaev, V. (2013). Biophysical techniques for detection of cAMP and cGMP in living cells. *Int. J. Mol. Sci.* 14, 8025–8046. doi:10.3390/ijms14048025

Stangherlin, A., and Zaccolo, M. (2012). cGMP-cAMP interplay in cardiac myocytes: a local affair with far-reaching consequences for heart function. *Biochem. Soc. Trans.* 40, 11–14. doi:10.1042/BST20110655

Stangherlin, A., Koschinski, A., Terrin, A., Zoccarato, A., Jiang, H., Fields, L. A., et al. (2014). Fluorescent protein-based biosensors: methods and protocols. in *Fluorescent protein-based biosensors*. Editors J. Zhang, Q. Ni, and R. H. Newman (Totowa, NJ, United States: Humana Press). doi:10.1007/978-1-62703-622-1

Surdo, N. C., Berrera, M., Koschinski, A., Brescia, M., MacHado, M. R., Carr, C., et al. (2017). FRET biosensor uncovers cAMP nano-domains at β-adrenergic targets that dictate precise tuning of cardiac contractility. *Nat. Commun.* 8, 15031–15114. doi:10.1038/ncomms15031

Vaandrager, A. B., Hogema, B. M., and De Jonge, H. R. (2005). Molecular properties and biological functions of cGMP-dependent protein kinase II. *Front. Biosci.* 10, 2150–2164. doi:10.2741/1687

Zaccolo, M., De Giorgi, F., Cho, C. Y., Feng, L., Knapp, T., Negulescu, P. A., et al. (2000). A genetically encoded, fluorescent indicator for cyclic AMP in living cells. *Nat. Cel Biol.* 2, 25–29. doi:10.1038/71345

Zhao, D., Guallar, E., Vaidya, D., Ndumele, C. E., Ouyang, P., Post, W. S., et al. (2020). Cyclic guanosine monophosphate and risk of incident heart failure and other cardiovascular events: the ARIC study. *J. Am. Heart Assoc.* 9. doi:10.1161/JAHA.119.013966

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 © 2021 Klein, Sardi, Machado, Ortega, Comini and Pantano. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 16 March 2021
doi: 10.3389/fmolb.2021.618068

Check for updates

OPEN ACCESS

Edited by:
Fabio Trovato,
Freie Universität Berlin, Germany

Reviewed by:
Bin Sun,
Loyola University Chicago, United States
Dinler Amaral Antunes,
Rice University, United States
Brandon L. Garcia,
The Brody School of Medicine at East Carolina University, United States

***Correspondence:**
Arvind Sahu
arvindsahu@nccs.res.in
Giulia Palermo
giulia.palermo@ucr.edu

†These authors have contributed equally to this work and share first authorship
‡Dedicated to our late beloved mentor Dimitrios Morikis

§**Present address:**
Avneesh K. Gautam, Department of Medicine, Immunology and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: 16 October 2020
Accepted: 08 January 2021
Published: 16 March 2021

Citation:
Narkhede YB, Gautam AK, Hsu RV, Rodriguez W, Zewde NT, Harrison RES, Arantes PR, Gaieb Z, Gorham RD, Kieslich C, Morikis D, Sahu A and Palermo G (2021) Role of Electrostatic Hotspots in the Selectivity of Complement Control Proteins Toward Human and Bovine Complement Inhibition. Front. Mol. Biosci. 8:618068. doi: 10.3389/fmolb.2021.618068

Role of Electrostatic Hotspots in the Selectivity of Complement Control Proteins Toward Human and Bovine Complement Inhibition

Yogesh B. Narkhede[1†], Avneesh K. Gautam[2†§], Rohaine V. Hsu[1†], Wilson Rodriguez[1†], Nehemiah T. Zewde[1], Reed E. S. Harrison[1], Pablo R. Arantes[1], Zied Gaieb[1], Ronald D. Gorham Jr.[1], Chris Kieslich[1,3], Dimitrios Morikis[1‡], Arvind Sahu[2] and Giulia Palermo[1,4*]*

[1]Department of Bioengineering, University of California, Riverside, CA, USA, [2]National Centre for Cell Science, Pune University Campus, Ganeshkhind, India, [3]Department of Chemical Engineering, Auburn University, Auburn, AL, USA, [4]Department of Chemistry, University of California, Riverside, CA, USA

Poxviruses are dangerous pathogens, which can cause fatal infection in unvaccinated individuals. The causative agent of smallpox in humans, *variola virus*, is closely related to the bovine *vaccinia virus*, yet the molecular basis of their selectivity is currently incompletely understood. Here, we examine the role of the electrostatics in the selectivity of the smallpox protein SPICE and vaccinia protein VCP toward the human and bovine complement protein C3b, a key component of the complement immune response. Electrostatic calculations, in-silico alanine-scan and electrostatic hotspot analysis, as introduced by Kieslich and Morikis (*PLoS Comput. Biol*. 2012), are used to assess the electrostatic complementarity and to identify sites resistant to local perturbation where the electrostatic potential is likely to be evolutionary conserved. The calculations suggest that the bovine C3b is electrostatically prone to selectively bind its VCP ligand. On the other hand, the human isoform of C3b exhibits a lower electrostatic complementarity toward its SPICE ligand. Yet, the human C3b displays a highly preserved electrostatic core, which suggests that this isoform could be less selective in binding different ligands like SPICE and the human Factor H. This is supported by experimental cofactor activity assays revealing that the human C3b is prone to bind both SPICE and Factor H, which exhibit diverse electrostatic properties. Additional investigations considering mutants of SPICE and VCP that revert their selectivity reveal an "electrostatic switch" into the central modules of the ligands, supporting the critical role of the electrostatics in the selectivity. Taken together, these evidences provide insights into the selectivity mechanism of the complement regulator proteins encoded by the *variola* and *vaccinia viruses* to circumvent the complement immunity and exert their pathogenic action. These fundamental aspects are valuable for the development of novel vaccines and therapeutic strategies.

Keywords: Complement cofactor activity, Computational modeling, SPICE, VCP, Variola (smallpox) virus

INTRODUCTION

Poxviruses are highly dangerous pathogens, whose most notable member is *variola virus*, the causative agent of smallpox, which killed millions of people before the massive vaccination campaign by the World Health Organization (World Health Organization, 1980). Closely related to *variola virus*, the vaccine-derived agent, called *vaccinia virus*, is causing outbreaks in dairy cattle in India and Brazil (Trindade et al., 2007; Gurav et al., 2011). Intriguingly, while these poxviruses are closely related, *variola virus* is strictly human-specific, whereas *vaccinia virus* also infects bovines. However, the molecular basis of this selectivity is currently incompletely understood. Its knowledge is fundamental to understand the onset of smallpox outbreaks dairy cattle, as arising from *vaccinia virus*, as well as to develop novel vaccines and therapeutic strategies.

Both viruses actively evade the complement system, which is an effector arm of the immune system and one of the major innate immune responses to viruses. Thanks to an intricate network of proteins, the complement system regulates important pathophysiological functions, eliminates pathogens, ensures homeostasis and forms a bridge between the innate and adaptive immunity (Bennett et al., 2017; Reis et al., 2019). When the complement system is activated, the convertase enzyme converts the complement 3 (C3) protein into its C3b component, which can associate to cell surfaces (**Figure 1A**). Further cleavage events yield C3d (the terminal cleavage product of C3), which is a natural biomarker of complement activation. Both *variola* and *vaccinia* viruses encode complement regulator proteins, named SPICE (smallpox inhibitor of complement enzymes) and VCP (vaccinia virus complement control protein) respectively, which are structural homologues of the proteins that regulate complement activation. Specifically, both SPICE and VCP mimic the interactions of the human Factor H (FH, which is structurally similar to SPICE and VCP), such as competing with FH in binding C3b (**Figure 1B**). At the domain level, human FH mainly interacts with the following domains: CUB, TED, MG1, MG2, MG6, MG7, and MG8 (**Figures 1C,D**; **Supplementary Figure S1**). There are two mechanisms of complement regulation in humans—cofactor activity (CA) and decay accelerating activity (DAA). The cofactor activity involves a cofactor (e.g., FH/VCP/SPICE) and a plasma serine protease factor I (FI) that proteolytically degrades the complement proteins C3b/C4b, while decay-accelerating activity involves

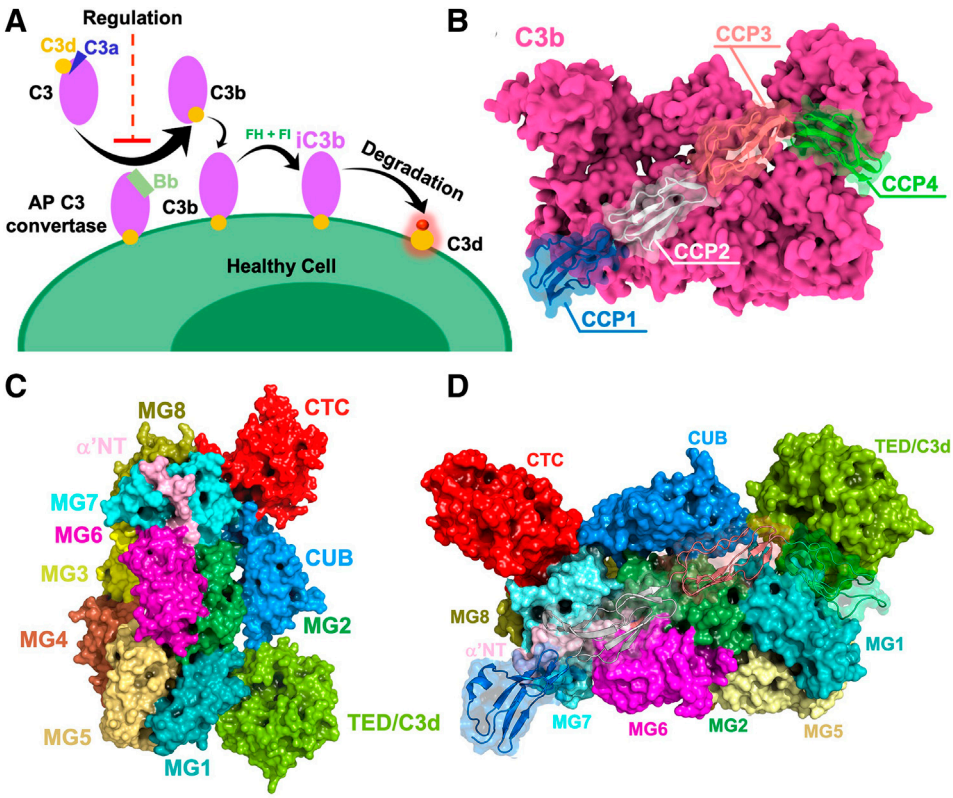


FIGURE 1 | (A) General overview of complement activation and regulation on healthy cells. Once complement is activated, the active fragment C3b is generated by the C3 convertase. Further cleavage events yield inactivated iC3b, which gets converted to C3d (yellow). **(B)** Human complement C3b protein bound to the human Factor H (FH), as captured via x-ray crystallography (PDB ID: 2WII) (Wu et al., 2009). The C3b protein is depicted as a molecular surface (magenta), binding the first four complement control protein (CCP) modules of FH (CCP1-4, shown using different colors). **(C)** Surface representations of complement C3b domains shown in different colors. **(D)** Complement C3b in complex with CCP1-4 of human FH. Domains CUB, TED, MG1, MG2, MG6, MG7, and MG8 of C3b provide a binding interface for human FH.

the decay or dissociation of the catalytic domain of the convertases. With these mechanisms, the smallpox and vaccinia viruses circumvent the complement immunity, exerting their pathogenic function (Lambris et al., 2008; Agrawal et al., 2017; Rosbjerg et al., 2017). At the molecular level, the complement regulator proteins SPICE, VCP, and the human FH consist of bead-like repeating subunits, known as Complement Control Protein (CCP) modules, which are covalently linked forming "beads-on-a-string" structures (Ojha et al., 2014). SPICE, VCP, and bovine FH display a sequence identity with the human FH of 27, 33, and 65%, respectively (**Supplementary Figure S2**) (Altschul et al., 1990; Gish and States, 1993). However, the first four CCP modules that bind C3b display a similar architecture and a sequence identity >90% with the human FH (**Figure 1B**; **Supplementary Figure S2**) (Wu et al., 2009). Indeed, only the K108 and K120 of SPICE correspond to E108 and E120 in VCP, which suggests a change in the electrostatic properties (Yadav et al., 2012). Considering the similarity of SPICE and VCP, it is intriguing how these two viral proteins could achieve selectivity toward the human and bovine isoforms of C3b (Yadav et al., 2012). This has been further highlighted by experimental studies that show the SPICE-HuC3b interaction has a K_D of 1.3 μM (Forneris et al., 2016), while that of Hu FH(1-4)-HuC3b interaction is 11.0 μM (Wu et al., 2009). When relative binding was measured using SPR, the relative binding strength of SPICE for HuC3b was ~130-fold higher than for VCP (Yadav et al., 2008), and relative binding strength of VCP for BoC3b was ~2.6 fold higher than for SPICE (Yadav et al., 2012). Several computational studies have revealed a key role of the electrostatics in the binding of CCP modules to the complement receptors, (Sfyroera et al., 2005; Zhang and Morikis, 2006; Yadav et al., 2008; Liszewski et al., 2009; El-Assaad et al., 2011; Kieslich et al., 2011c; Kieslich and Morikis, 2012; Yadav et al., 2012; Harrison et al., 2015; Mohan et al., 2015; Harrison et al., 2020), demonstrating that the diverse electrostatic properties of CCP modules mediate different specificities toward the complement proteins. Kieslich and Morikis (Kieslich and Morikis, 2012) have examined the effect of electrostatic interactions at the level of the complement C3 fragment C3d and its complement receptor 2 (CR2), whose formation constitutes the link between innate and adaptive immunity, proposing that the C3d-CR2 electrostatic specificity is critical in the onset of adaptive immunity. That study demonstrated that functional "electrostatic hotspots" importantly contribute to the protein-protein association and to the selectivity of C3d and its receptor. In this respect, electrostatic hotspots comprising clusters of charged residues have been found to be crucial regulators also across several other biomolecular systems, functioning also as effective drug binding sites (Foloppe et al., 2006; Adhireksan et al., 2017).

In this work, we examine the role of the electrostatics in the selectivity of SPICE and VCP toward the human and bovine isoforms of C3b. Electrostatic calculations, and in-silico alanine-scanning of SPICE, VCP, and human factor H bound to human and bovine C3b have been performed to characterize electrostatic hotspots, as introduced by Kieslich and Morikis (Kieslich and Morikis, 2012), identifying sites resistant to local perturbation where the electrostatic potential is likely to be evolutionary conserved (Kieslich and Morikis, 2012). The calculations suggest that the bovine C3b is electrostatically prone to selectively bind its VCP exogenous ligand. On the other hand, the human isoform of C3b exhibits a lower electrostatic complementarity toward its SPICE ligand. Yet, the human C3b displays a highly preserved electrostatic core, which suggests that this isoform could be less selective and bind different ligands, such as both SPICE and the human FH. This is supported by experimental Cofactor Activity Assays revealing that the human C3b is prone to bind both SPICE and FH, which exhibit a diverse ESI map and different electrostatic properties. Further investigations of the SPICE and VCP mutants reverting their selectivity (Yadav et al., 2012) reveal that an "electrostatic switch" into their central modules plays a key role in the process. Overall, these evidences highlight the critical role of the electrostatics in the specificity of SPICE and VCP toward the human and bovine isoforms of the C3b protein, providing fundamental insights that can help the development of novel vaccines and therapeutic strategies.

COMPUTATIONAL MATERIALS AND METHODS

Structure Preparation

Structural models have been based on the crystallographic coordinates of the structure of the human C3b-FH complex (PDB ID: 2WII, **Figure 1B**) (Wu et al., 2009) solved at 2.7 Å resolution. SPICE and VCP display a sequence identity (Madeira et al., 2019) with the endogenous human FH of 27 and 33% respectively (Altschul et al., 1990; Gish and States, 1993). However, our investigations focused on the first four CCP modules that bind C3b, which display a similar architecture and a sequence identity >90% with the human FH (Wu et al., 2009). The Bovine C3b, SPICE, and VCP structures were modeled using PDB 2WII as a template through homology modeling using the software Modeller (Eswar et al., 2006). The human C3b portion of 2WII served as a reference for bovine C3b, while factor H was used to model the structures of VCP and SPICE. Following the homology modeling, an iterative qualitative assessment (Williams et al., 2018) and refinement of the models was performed using UCLA SAVES server (SAVES, 2020) and Modeller. The final models were aligned to PDB 2WII yielding the following complexes: Human C3b-SPICE, Human C3b-VCP, Bovine C3b-SPICE, and Bovine C3b-VCP.

Computational Alanine-Scan, Electrostatic Calculations, and Hotspot Analysis

Electrostatic hotspot analysis has been performed using the AESOP (Analysis of Electrostatic Similarities Of Proteins) code, developed by Morikis (Kieslich et al., 2011b; Harrison et al., 2017) and extensively employed for electrostatic analysis of proteins (Kieslich et al., 2011a, Kieslich et al., 2011b; Gorham et al., 2011a, Gorham et al., 2011b; Harrison et al., 2015; Mohan et al., 2015, Mohan et al., 2016; Zewde et al., 2018; Harrison and

Morikis, 2019). AESOP enables the analysis of electrostatic hotspots through the calculation of an Electrostatic Similarity Indices (ESI), (Kieslich and Morikis, 2012), which assesses the electrostatic potential upon perturbation (i.e., mutagenesis), identifying regions of high electrostatic similarity, or those regions least affected by perturbation. High ESI values identify regions with resistance to perturbation, where the electrostatic potential is likely to be evolutionarily conserved, and constitute functional sites or electrostatic hotspots (Kieslich and Morikis, 2012). The calculation of the ESI involves the following steps: 1) generation of alanine scan mutations for every ionizable residue in a protein complex, one at a time, 2) calculation of Poisson-Boltzmann electrostatic potentials for each mutant protein and parent, 3) calculation of ESI values and projection on the protein surface for visualization. The resulting surface maps can be referred to as ESI perturbation maps, displaying regions of the protein resistant to perturbation (i.e., displaying high ESI values) and identifying electrostatic hotspots. Upon alanine-scan and calculation of the electrostatic potentials, cumulative distribution of Electrostatic Similarity Index (ESI) is computed as:

$$\mathrm{ESI}(i,j,k) = \frac{1}{N}\sum_{n=1}^{N} 1 - \frac{\left|\varphi_A(i,j,k) - \varphi_{B,n}(i,j,k)\right|}{\max\left(\left|\varphi_A(i,j,k)\right|, \left|\varphi_{B,n}(i,j,k)\right|\right)},$$

where φ_A is the electrostatic potential of the original protein and $\varphi_{B,n}$ is the electrostatic potential of the N mutants at the grid point (i,j,k). The ESI is calculated at each grid point (i,j,k) and normalized by the number of electrostatic potential comparisons N. This measure of electrostatic similarity describes the similarity of the electrostatic potential of a set of mutants to the original parent protein at a given grid point, for a family of N mutants. For the ESI calculations, an ionic strength corresponding to 0 mM monovalent ion concentration was used, as described earlier (Kieslich and Morikis, 2012). As described above, electrostatic hotspots can be evaluated as the projections of the ESI values on the protein surface.

In this work, the number of protein mutants generated by alanine scan are as follows: 66 mutants for FH, 42 mutants for SPICE, 46 mutants for VCP, 394 mutants for the human C3b, and 419 mutants for the bovine C3b. The calculations were performed at pH 7, mutating the following ionizable amino acids: Asp, Glu, Arg, Lys, and His. The apparent pK_a of the His residues was predetermined using PROPKA (Søndergaard et al., 2011). Electrostatic potentials were calculated by solving the linearized Poisson-Boltzmann equation using APBS (Baker et al., 2001). Positional coordinates in the PDB files were converted to PQR files in preparation for the Poisson-Boltzmann calculations, by adding hydrogen atoms, atomic radii, and partial charges, using PDB2PQR (Dolinsky et al., 2007) and the PARSE force field (Sitkoff et al., 1994). Electrostatic calculations were performed for the families of mutant and parent original proteins, using the separated protein and ligand components from the protein-ligand complex. The protein-ligand complex structure was centered in a grid with 161 × 161 × 257 grid points and grid lengths of 144 × 156 × 228 Å, resulting in a resolution of less than 1 Å/grid point. The protein or ligand was then removed to performed electrostatic potential calculations of each component. All mutants of each protein were centered in an identical manner to assure proper comparison of electrostatic potentials. Ionic strength corresponding to concentration of monovalent solvent ions of 150 mM was used in the calculations. Protein and solvent dielectric coefficients were set to 20 and 78.54, respectively, as determined in previous studies (Gorham et al., 2011b). Calculations were performed at 298.15 K temperature. Additional electrostatic potential analysis was performed for the VCP mutants E108K, E120K, and E108K/E120K, and the SPICE mutants K108E, K120E, K108E/K120E/N144E. Electrostatic calculations have been performed as described above, using ionic strengths corresponding to 150 mM. The software Chimera (Pettersen et al., 2004) has been employed to project ESI values on the 3D structures, thereby visualizing the electrostatic hotspots.

Cofactor Activity Assay

The cofactor activity of factor H in conjunction with factor I toward human and bovine C3b was performed by employing a fluid phase cofactor assay described earlier (Gautam et al., 2015). In brief, 2 μg of C3b (human or bovine) was mixed with 1.1 μg of factor H in presence or absence of 15 ng of factor I (human or bovine) in a total volume of 15 μl, and reaction mix was incubated at 37 °C for different time intervals. After the indicated time points, 5 μl of the sample buffer containing DTT was added to the reaction mix to stop the reaction. The C3b cleavage products were then resolved on 9% SDS-PAGE gel. The percentage of C3b cleaved with time was quantitated by densitometric analysis of α′-chain of C3b. The time required for 50% cleavage of the α′-chain of C3b was calculated from the plots generated by plotting the percentage of α′-chain of C3b against time. Human and bovine C3b, and bovine factor I were purified from the respective plasma as described (Yadav et al., 2012). Human factor H and factor I were purchased from Complement Technology, Inc. (TX, United States). Although it would be more meaningful to test bovine endogenous ligand (i.e., bovine FH) to test our hypothesis, however, our efforts to purify bovine FH from the sera failed, and the same is not available commercially.

RESULTS AND DISCUSSION

The results of homology modeling for the bovine complement C3b, SPICE, and VCP have been summarized in **Supplementary Table S1** and **Supplementary Figures S3–S6**. The analyses revealed that the models are of sufficient quality for the ensuing computational electrostatic calculations.

In order to understand the role of the electrostatics in the complementarity between C3b and its ligands, we performed electrostatic calculations, thereby analyzing the spatial distributions of the electrostatic potential. **Figure 2A** shows the surface projections of the electrostatic potentials for the bovine and human C3b, as well as for its ligands FH, VCP, and SPICE. An "open book" view highlights the surfaces of contact at the interface of the protein-ligand complex. The surface projections of the electrostatic potential of both bovine

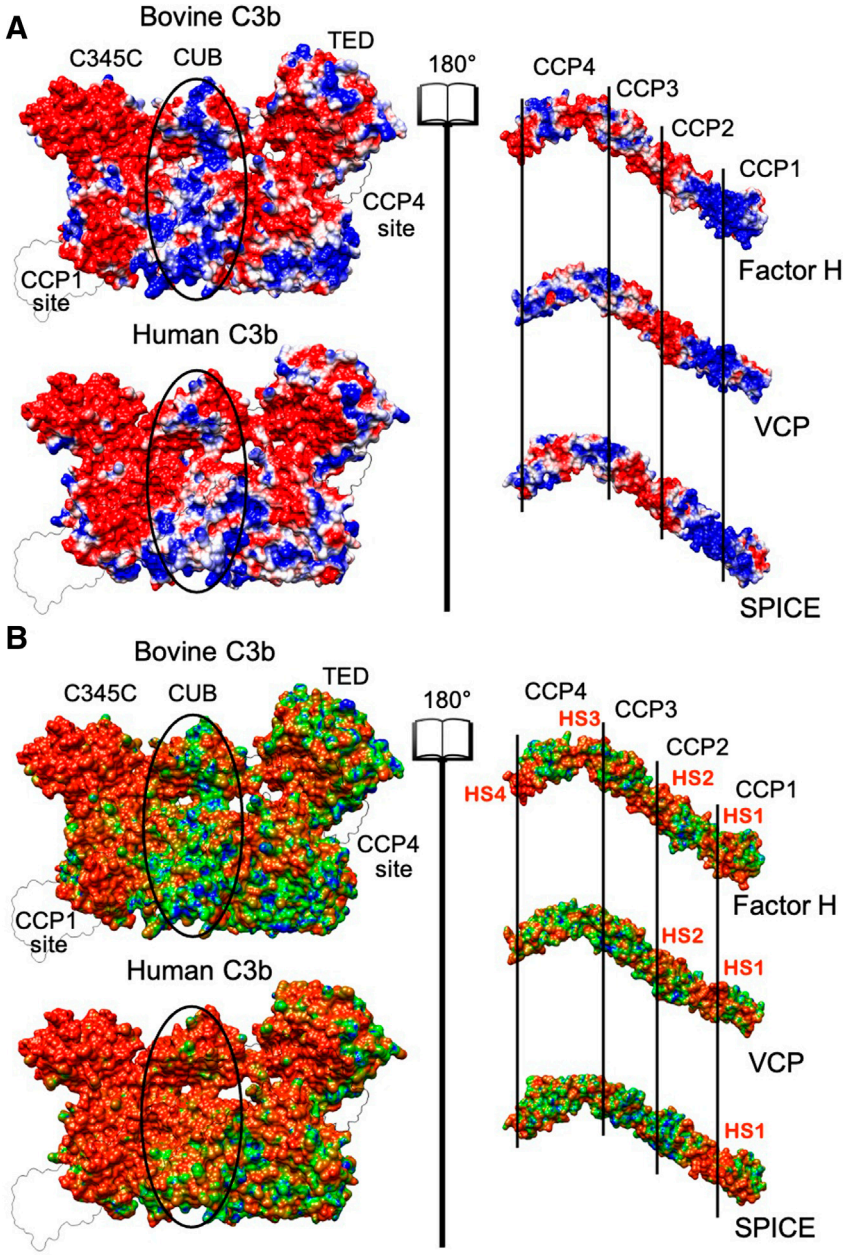


FIGURE 2 | (A) Electrostatic potential surface projection for the bovine and human C3b proteins (left panel) and for the Factor H, VCP and SPICE ligands (right panel). The left and right panels are separated by an "open book" view, which reveals the surfaces of contact at the interface of the protein-ligand complex. The protein domains of C3b (C345C, CUB, and TED) and the complement control protein (CCP) modules of Factor H, VCP and SPICE (CCP1-4) are also shown. On the 3D structure of C3b, the location of the CCP1 and CCP4 binding sites is also shown. The ellipses depict differences in the central regions of bovine and human C3b. The electrostatic potential is colored from red (negative) to blue (positive). **(B)** Electrostatic Similarity Index (ESI), projected on the 3D structure of the bovine and human C3b protein (left panel) and of the Factor H, VCP and SPICE ligands (right panel). Observed electrostatic hotspots on the ligands are marked as HS1–HS4. The color scheme blue–green–orange represents low–to–high similarity, corresponding to ESI values of 0.6–0.8–1.0 for the bovine and human C3b, and ESI values of 0.4–0.7–1.0 for the ligands.

and human C3b are overall similar. However, a more negative character in the CUB domain of the human C3b was observed, compared to the bovine protein (net charges are $-30e^-$ for human C3b and $-16e^-$ for bovine C3b). The surface projections of the electrostatic potential of the ligands, FH, VCP, and SPICE, display both similarities and pronounced differences. Electrostatic complementarity between both the bovine and human C3b proteins and their ligands is evident at the level of the positively charged ligand module CCP1. Indeed, the CCP1 module of FH, VCP, and SPICE exhibits electrostatic complementarity with the negatively charged MG7 domain, where the CCP1 module binds (**Figure 2A**) (Wu et al., 2009). Additionally, our hotspot analyses reveal that the CUB and MG6 domains of bovine and human C3b harbor most of the variations

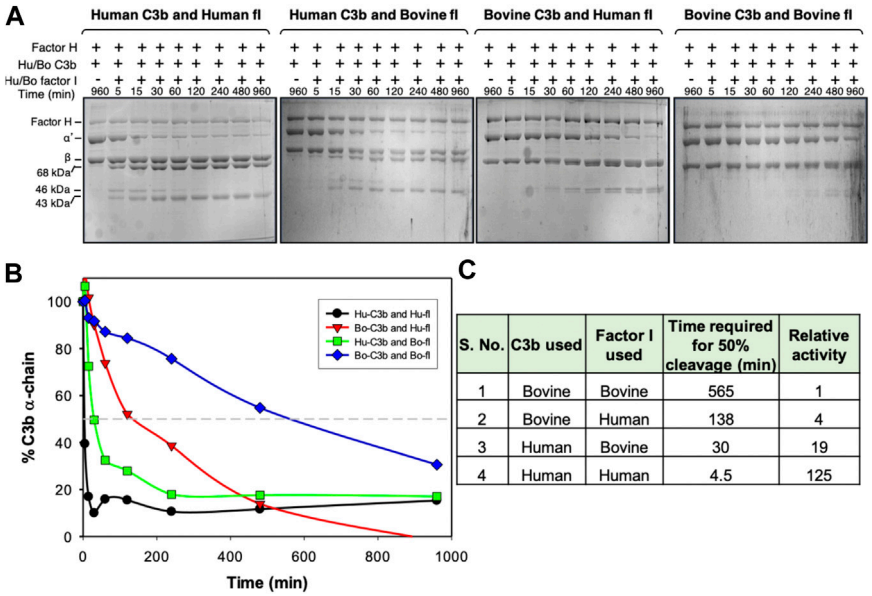
that play a key role in CCP2 binding for FH, SPICE, and VCP (**Figure 2B**). Partial electrostatic complementarity is observed between module CCP4 of FH, VCP, and SPICE and the TED domain of both bovine and human C3b, although the distribution of positive and negative patches in CCP4 differs in FH, VCP, and SPICE. In the case of the middle modules CCP2 and CCP3 of VCP and SPICE (which display a more negative electrostatic surface), the electrostatic complementarity appears to be favored with the bovine C3b, with respect to the human C3b, which shows increased negatively charged surface. This observation agrees with the increased specificity observed for VCP toward the bovine C3b (Yadav et al., 2012). However, it does not explain the lower specificity of binding for SPICE, which is instead human complement-specific.

As discussed before, (Kieslich and Morikis, 2012) visual inspection of the electrostatic potential may not adequately highlight the electrostatic contributions to protein-protein association. Hence, to gain further insights into the electrostatic complementarity, we employed an Electrostatic Similarity Index (ESI) (Carbó et al., 1980; Hodgkin and Richards, 1987). This analysis enables to assess the electrostatic potential upon perturbation (i.e., mutagenesis, details are reported in the Methods section), identifying regions of high "electrostatic similarity" or those regions least affected by perturbation. High ESI values indicate regions with resistance to perturbation with evolutionarily conserved electrostatic potential (Kieslich and Morikis, 2012), and pinpoints the presence of electrostatic hotspots. Regions of C3b and its ligands characterized by high ESI values are shown in **Figure 2B**, as identified by the surface projection of ESI values on the 3D structures. The human C3b displays a larger electrostatic hotspot (orange surface) in the binding site of all ligand modules (CCP1-CCP4), extending to the C345C and CUB domains as well, compared to bovine C3b. A larger electrostatic hotspot is also observed in the TED domain of the human C3b, which is the binding site of the CCP4 module. The FH, VCP and SPICE ligands display however electrostatic hotspot diversity. An electrostatic hotspot in between modules CCP1-CCP2 is observed in all three ligands, but is more pronounced in SPICE (indicated in **Figure 2B** as Hot Spot 1, HS1). VCP and FH display a second hotspot at the CCP2-CCP3 interface (HS2), while FH also displays other two regions characterized by electrostatic hotspots at the level of CCP4 (HS3, HS4). Interpretation of the ESI maps suggests that HS1, which is conserved in the three ligands, mediates binding affinities for all three ligands at the level of the CCP1 binding site on both the human and bovine isoforms of C3b. A comparison with **Figure 2A** also shows that HS1 corresponds to positively charged areas, which are prone to bind the negatively charged binding site on C3b.

The bovine C3b, which preferentially binds VCP, shows that several electrostatic hotspots, where the electrostatics is resistant to perturbation, are distributed on the protein surface (**Figure 2B**). This is also observed in the VCP and FH ligands, where the HS1-3 also display a high electrostatic complementarity with the bovine C3b (**Figure 2A**). This is in line with the higher affinity of VCP for the bovine C3b, compared to human C3b (Yadav et al., 2012). Indeed, the negative charge of HS2 in VCP participates in attractive interactions with the positive charges in the CUB region of the bovine C3b, while establishing repulsive interactions with the negative charges in the CUB region of the human C3b (**Figure 2A**), thus contributing to the higher selectivity of VCP for bovine C3b.

The human C3b (which is SPICE-specific) shows an extended hotspot area, while SPICE shows the absence of hotspots in its complementary central modules CCP2 and CCP3 (**Figure 2B**). This observation indicates that while the human C3b strongly preserves its core electrostatics, its SPICE ligand is highly susceptible to electrostatic perturbation. Moreover, the central modules of SPICE display low electrostatic complementarity with the CUB region of the human C3b (**Figure 2A**) and the positively charged HS1 hotspot is the sole electrostatic anchor to the human C3b. The highly preserved electrostatic core indicated by the extended hotspot area in the human C3b suggests that this protein could be less selective in binding to different ligands like SPICE and FH, which exhibit diverse ESI map and different electrostatic properties (Yadav et al., 2012).

Overall, analysis of the ESI maps suggests that changes in the electrostatic properties of the central modules of VCP and SPICE (i.e., in between CCP2-CCP3) could modulate the selectivity for the bovine vs. the human C3b (Yadav et al., 2012). Most notably, the bovine C3b and its VCP ligand display a common heterogeneity of the ESI maps (**Figure 2B**) and a high electrostatic complementarity (**Figure 2A**). This suggests that the bovine C3b is electrostatically prone to bind its VCP and FH ligands, which exhibit similar electrostatic properties. The human isoform of C3b displays a large electrostatic hotspot, which indicates a highly preserved electrostatic core. On the other hand, its SPICE ligand shows the absence of hotspots in the central modules and low electrostatic complementarity, such that the positively charged HS1 hotspot is the sole electrostatic anchor to the human C3b. These evidences suggest that the human C3b, which exhibits a highly preserved electrostatic core, could be less selective and bind different ligands like SPICE and FH, which exhibit diverse ESI map and different electrostatic properties. To support this hypothesis, we performed cofactor activity assays, where the activity of the human FH at the level of the human C3b was measured (FH + HuC3b), in comparison with that of the human FH at the level of the bovine C3b (FH + BoC3b). These assays were performed in the presence of a critical serine protease enabling the reaction, *viz.* factor I, both from bovine and human species (**Figure 3**). Compared to our controls (FH + HuC3b + HuFI), a 31-fold decrease in cofactor activity was observed when the bovine C3b is bound to the human FH (FH + BoC3b + HuFI), revealing that the bovine C3b is less prone to bind the human FH. Additionally, when comparing Cofactor activity Assay of FH + HuC3b + BoFI, and FH + BoC3b + BoFI, we observe a 19-fold decrease in relative activity, further supporting the evidence that the bovine C3b is less specific than the human C3b for the human FH. These cofactor activity assays were performed analogously to our previous experimental investigation describing the specificity of VCP and SPICE toward the bovine and human C3b, respectively (Yadav et al., 2012). Taken together, these experimental data indicate that the human C3b is prone to bind to the human FH (**Figure 3**), as well as its SPICE ligand (Yadav et al., 2012). Considering the remarkable differences in terms of electrostatic properties of the human FH and SPICE ligands (**Figure 2**), as well as the highly preserved electrostatic core of the human C3b, the above reported experimental evidence



S. No.	C3b used	Factor I used	Time required for 50% cleavage (min)	Relative activity
1	Bovine	Bovine	565	1
2	Bovine	Human	138	4
3	Human	Bovine	30	19
4	Human	Human	4.5	125

FIGURE 3 | Cofactor activity of human factor H for human and bovine C3b. **(A)** Analysis of cofactor activity by SDS-PAGE gel analysis. C3b (human or bovine) was mixed with human factor H and factor I (human or bovine), incubated at 37 °C for the indicated time, and examined for C3b cleavage (α′-chain cleavage) by running the reaction mix on SDS-PAGE. **(B)** The amount of α′-chain at different time points was quantitated densitometrically and plotted against time. **(C)** Relative cofactor activity of factor H against human and bovine C3b in the presence of human or bovine factor I. The molar concentration of the complement components used in the cofactor assay was: 758 nM Hu/Bo C3b, 11 nM Hu/Bo factor I and 473 nM Hu factor H.

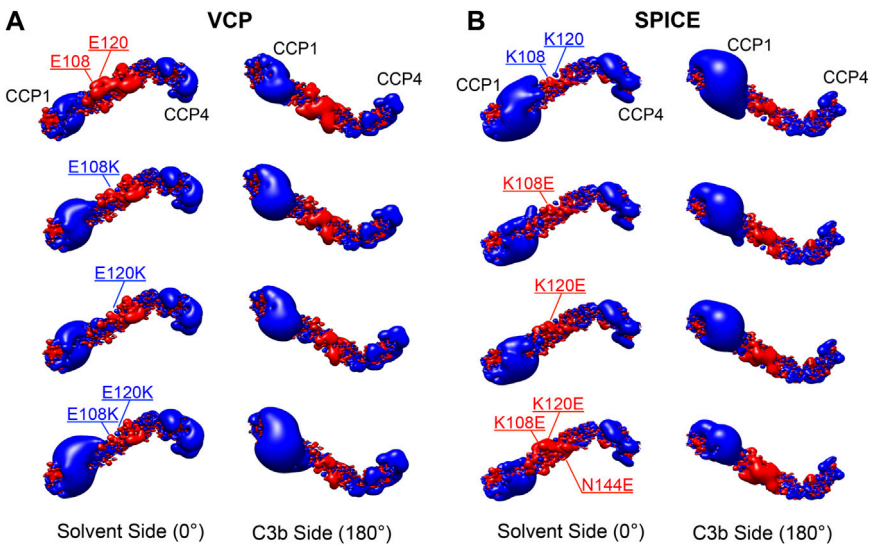


FIGURE 4 | Electrostatic potential maps for the **(A)** Vaccinia virus complement control protein (VCP) and **(B)** smallpox inhibitor of complement enzymes (SPICE), upon point mutations of the E108K, E120K, and E108K/E120K residues in VCP, and upon the reverse mutations of K108E, K120E, and K108E/K120E/N144E in SPICE. The electrostatic potential is represented as isopotential contour maps, with blue and red contours having isovalues of +2 and −2 k_BT/e^-, respectively. Two orientations of the ligands are shown, displaying their solvent side and the side binding C3b (upon a 180° rotation around the horizontal axis). The electrostatic potential calculations were performed with 150 mM ionic strength.

support that the human C3b is intrinsically less selective than the bovine C3b. Finally, it is notable that the experimental results reported in this paper, as well as the data from our previous study, have been performed by including the enzyme Factor I in the experimental assay. The binding of the FI enzyme occurs at a distal site with respect to the FH binding site on C3b, thereby not interfering with the FH-C3b interaction (Xue et al., 2017).

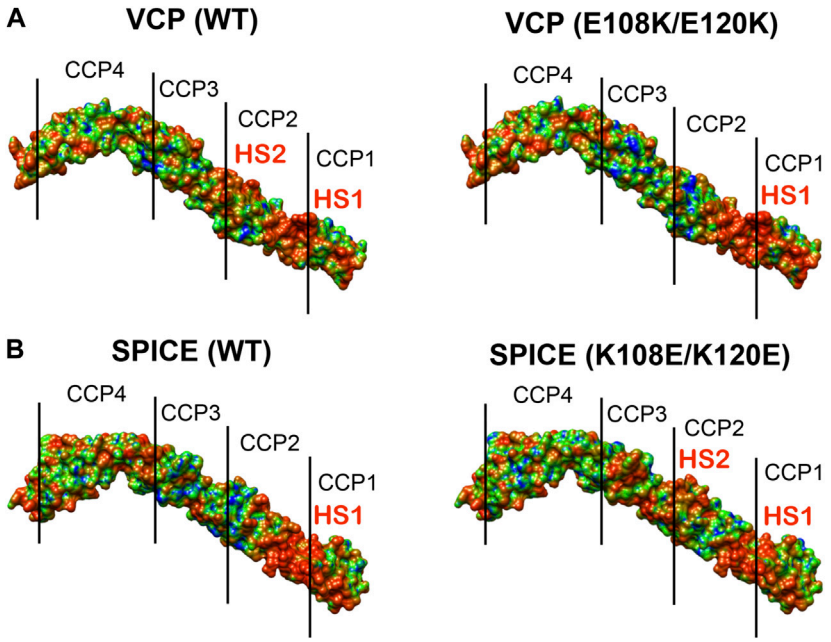


FIGURE 5 | Electrostatic Similarity Index (ESI), projected on the 3D structure of **(A)** the vaccinia virus complement control protein (VCP) and of **(B)** the smallpox inhibitor of complement enzymes (SPICE). ESI is computed and projected on the proteins as wild-type (WT, left panel) and upon mutation of the residues 108 and 120 (right panel). Observed electrostatic hotspots on the ligands are marked as HS1–HS2. The color scheme blue–green–orange represents low–to–high similarity, corresponding to ESI values of 0.4–0.7–1.0.

To further investigate the electrostatic properties of SPICE and VCP, we extended our electrostatic analysis to two mutants of SPICE and VCP. We introduced mutations that have been experimentally shown to revert the selectivity toward the bovine and human C3b (Yadav et al., 2012). In SPICE, we introduced the K108E, K120E, and N144E mutations that confer bovine selectivity, whereas in VCP, we introduced the E108K, and E120K mutations that confer SPICE-like activity. **Figure 4** shows the spatial distributions of electrostatic potentials of VCP (panel A), including the E108K, E120K, and E108K/E120K mutations, and of SPICE (panel B) including the reverse mutations (K108E ad K120E) and the N144E mutation, at an ionic strength of 150 mM. In VCP, the E108K and E120K mutations augment the positive electrostatic potential of modules CCP1 and CCP2. On the other hand, in SPICE, the K108E, K120E, and N144E mutations reduce the corresponding positive electrostatic potential. The increase of the positive electrostatic potential in the VCP modules CCP1 and CCP2, upon switching of the 108 and 120 residues from E–to–K, agrees well with a switch in selectivity toward the human C3b, which displays a highly negative electrostatic surface (**Figure 2A**) (Yadav et al., 2012). As opposite, the decrease of the positive electrostatic potential upon the reverse K–to–E mutations of residues 108 and 120 in SPICE, confers reduced affinity for the highly negatively charged electrostatic surface of the human C3b. Considering also the experimental evidence that these mutations revert the selectivity of the SPICE and VCP viral proteins, (Yadav et al., 2012) these outcomes suggest that an "electrostatic switch" into their central modules is critical for the opposite selectivity toward the human and bovine C3b. To gain more insights on this "electrostatic switch," we performed the inspection of the electrostatic hotspots in the VCP and SPICE mutants through the calculation of the ESI maps. We analyzed the charge reversal for residues 108 and 120 in module CCP2 from E–to–K in VCP and from K–to–E in SPICE. This analysis has been carried out on the VCP E108K/E120K and SPICE K108E/K120E mutants, and compared with the wild-type ligands (**Figure 5**). As a result, the E108K/E120K mutation produces a VCP protein with a SPICE-like electrostatic hotspot map. Indeed, the E108K/E120K mutation augments the HS1 in VCP (to become similar to that of SPICE) and abolishes the HS2 hotspot (not present in SPICE). This is in good agreement with the experimental evidence that the E108K, E120K mutations in VCP confer to the latter SPICE-like activity (Yadav et al., 2012). On the other hand, the SPICE double mutation K108E/K120E generates a VCP-like electrostatic hotspot map, with reduction of the HS1 area and appearance of HS2. This confirms the evidence of an "electrostatic switch" in the central modules of SPICE and VCP.

CONCLUSION

Here, we performed a computational study aimed at shedding light into the molecular mimicry of the viral proteins VCP and SPICE of the human FH binding to the complement protein C3b. These viral proteins exploit their structural similarity with the human FH to evade the immune response of the complement system, which is one of the important factors in their pathogenicity. Although SPICE and VCP are structural homologues and differ for a few amino acid mutations, they exhibit different selectivity toward the human and bovine

isoforms of C3b (Yadav et al., 2012). Inspired by an early study by Kieslich and Morikis (Kieslich and Morikis, 2012), demonstrating that the diverse electrostatic properties of the Complement Control Proteins (CCP) mediate different specificities toward the complement proteins, we performed an electrostatic analysis of the SPICE and VCP ligands and of their human and bovine C3b counterparts. Electrostatic calculations, in-silico alanine-scan and electrostatic hotspot analysis were used to characterize the electrostatic complementarity of the protein-protein interaction and the onset of electrostatic hotspots, which are crucial regions where the electrostatic potential is resistant to permutation and is likely to be evolutionarily conserved (Kieslich and Morikis, 2012).

The calculations reveal that the bovine C3b, which preferentially binds VCP (Yadav et al., 2012), shows high electrostatic complementarity with its VCP ligand, displaying also several electrostatic hotspots that are also observed in VCP and FH. This suggests that the bovine C3b is electrostatically prone to selectively bind its VCP ligand. The human isoform of C3b displays a highly preserved electrostatic core, as arising from the observation of an extended electrostatic hotspot. On the other hand, its SPICE ligand shows the absence of hotspots in the central modules and low electrostatic complementarity, such that the positively charged HS1 hotspot is the sole electrostatic anchor to the human C3b. Considering this low electrostatic complementarity, the highly preserved electrostatic core in the human C3b suggests that this protein could be less selective in binding ligands like SPICE and FH. Additional investigations considering mutants of SPICE and VCP that revert their selectivity toward the bovine and human C3b (Yadav et al., 2012), reveal that an "electrostatic switch" into the central modules of the ligands is critical for their opposite selectivity. Taken together, these evidences highlight the critical role of the electrostatics in the specificity of SPICE and VCP toward the human and bovine isoforms of the C3b protein. Considering the pathogenesis of the *vaccinia virus*, causing outbreaks in dairy cattle in India and Brazil (Trindade et al., 2007; Gurav et al., 2011), the mechanistic aspects arising from this electrostatic analysis could provide critical insights to develop novel vaccines and therapeutic strategies.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation, upon request.

AUTHOR CONTRIBUTIONS

YN, RH, and WR performed computational analysis and wrote the manuscript. AG performed experiments. NZ, RH, PA, and ZG performed computational analysis. RG Jr, CK, and DM originally developed the AESOP tool. DM conceived the initial idea. AS directed experimental research. GP directed computational research and wrote the manuscript. YN, RH, WR, and AG contributed equally.

FUNDING

This material is based upon work supported by the National Institutes of Health (Grant R01 EY027440, to GP). This material is also based upon work supported by the National Science Foundation under Grant No. CHE-1905374 (to GP). AS is a J. C. Bose National Fellow. His laboratory is supported by project grant BT/PR28506/MED/29/1307/2018 from the Department of Biotechnology, New Delhi, India. The AESOP code has been originally developed by the Morikis research group and is currently maintained and improved by the Palermo laboratory. AESOP is freely available at: https://aesop.readthedocs.io/en/latest/index.html and on Github at: https://github.com/biomodel/aesop/blob/master/docs/index.rst.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.618068/full#supplementary-material.

REFERENCES

Adhireksan, Z., Palermo, G., Riedel, T., Ma, Z., Muhammad, R., Rothlisberger, U., et al. (2017). Allosteric cross-talk in chromatin can mediate drug-drug synergy. *Nat. Commun.* 8, 14860. doi:10.1038/ncomms14860

Agrawal, P., Nawadkar, R., Ojha, H., Kumar, J., and Sahu, A. (2017). Complement evasion strategies of viruses: an overview. *Front. Microbiol.* 8, 1117. doi:10.3389/fmicb.2017.01117

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

Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* 98, 10037–10041. doi:10.1073/pnas.181342398

Bennett, K. M., Rooijakkers, S. H., and Gorham, R. D. (2017). Let's tie the knot: marriage of complement and adaptive immunity in pathogen evasion, for better or worse. *Front. Microbiol.* 8, 89. doi:10.3389/fmicb.2017.00089

Carbó, R., Leyda, L., and Arnau, M. (1980). How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quant. Chem.* 17, 1185–1189. doi:10.1002/qua.560170612

Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G., et al. (2007). PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* 35, W522–W525. doi:10.1093/nar/gkm276

El-Assaad, A. M., Kieslich, C. A., Gorham, R. D., and Morikis, D. (2011). Electrostatic exploration of the C3d-FH4 interaction using a computational alanine scan. *Mol. Immunol.* 48, 1844–1850. doi:10.1016/j.molimm.2011.05.007

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M., et al. (2006). Comparative protein structure modeling using modeller. *Curr Protoc Bioinformatics* 54, 5.6.1–5.6.37. doi:10.1002/0471250953.bi0506s15

Foloppe, N., Matassova, N., and Aboul-ela, F. (2006). Towards the discovery of drug-like RNA ligands? *Drug Discov. Today* 11, 1019–1027. doi:10.1016/j.drudis.2006.09.001

Forneris, F., Wu, J., Xue, X., Ricklin, D., Lin, Z., Sfyroera, G., et al. (2016). Regulators of complement activity mediate inhibitory mechanisms through a common C3b-binding mode. *EMBO J.* 35, 1133–1149. doi:10.15252/embj.201593673

Gautam, A. K., Panse, Y., Ghosh, P., Reza, M. J., Mullick, J., and Sahu, A. (2015). Mutational analysis of Kaposica reveals that bridging of MG2 and CUB domains of target protein is crucial for the cofactor activity of RCA proteins. *Proc. Natl. Acad. Sci. U.S.A.* 112, 12794–12799. doi:10.1073/pnas.1506449112

Gish, W., and States, D. J. (1993). Identification of protein coding regions by database similarity search. *Nat. Genet.* 3, 266–272. doi:10.1038/ng0393-266

Gorham, R. D., Kieslich, C. A., and Morikis, D. (2011a). Electrostatic clustering and free energy calculations provide a foundation for protein design and optimization. *Ann. Biomed. Eng.* 39, 1252–1263. doi:10.1007/s10439-010-0226-9

Gorham, R. D., Kieslich, C. A., Nichols, A., Sausman, N. U., Foronda, M., and Morikis, D. (2011b). An evaluation of Poisson-Boltzmann electrostatic free energy calculations through comparison with experimental mutagenesis data. *Biopolymers* 95, 746–754. doi:10.1002/bip.21644

Gurav, Y. K., Raut, C. G., Yadav, P. D., Tandale, B. V., Sivaram, A., Pore, M. D., et al. (2011). Buffalopox outbreak in humans and animals in Western Maharashtra, India. *Prev. Vet. Med.* 100, 242–247. doi:10.1016/j.prevetmed.2011.03.008

Harrison, R. E. S., Mohan, R. R., Gorham, R. D., Kieslich, C. A., and Morikis, D. (2017). AESOP: a Python library for investigating electrostatics in protein interactions. *Biophys. J.* 112, 1761–1766. doi:10.1016/j.bpj.2017.04.005

Harrison, R. E. S., and Morikis, D. (2019). Molecular mechanisms of macular degeneration associated with the complement factor H Y402H mutation. *Biophys. J.* 116, 215–226. doi:10.1016/j.bpj.2018.12.007

Harrison, R. E. S., Zewde, N. T., Narkhede, Y. B., Hsu, R. V., Morikis, D., Vullev, V. I., et al. (2020). Factor H-inspired design of peptide biomarkers of the complement C3d protein. *ACS Med. Chem. Lett.* 11, 1054–1059. doi:10.1021/acsmedchemlett.9b00663

Harrison, R. E. S., Gorham, R. D., and Morikis, D. (2015). Energetic evaluation of binding modes in the C3d and Factor H (CCP 19-20) complex. *Protein Sci.* 24, 789–802. doi:10.1002/pro.2650

Hodgkin, E. E., and Richards, W. G. (1987). Molecular similarity based on electrostatic potential and electric field. *Int. J. Quant. Chem.* 32, 105–110. doi:10.1002/qua.560320814

Kieslich, C. A., Gorham, R. D., and Morikis, D. (2011a). Is the rigid-body assumption reasonable? *J. Non-Cryst. Solids* 357, 707–716. doi:10.1016/j.jnoncrysol.2010.05.087

Kieslich, C. A., and Morikis, D. (2012). The two sides of complement C3d: evolution of electrostatics in a link between innate and adaptive immunity. *PLoS Comput. Biol.* 8, e1002840. doi:10.1371/journal.pcbi.1002840

Kieslich, C. A., Morikis, D., Yang, J., and Gunopulos, D. (2011b). Automated computational framework for the analysis of electrostatic similarities of proteins. *Biotechnol. Prog.* 27, 316–325. doi:10.1002/btpr.541

Kieslich, C. A., Vazquez, H., Goodman, G. N., de Victoria, A. L., and Morikis, D. (2011c). The effect of electrostatics on factor H function and related pathologies. *J. Mol. Graph. Model.* 29, 1047–1055. doi:10.1016/j.jmgm.2011.04.010

Lambris, J. D., Ricklin, D., and Geisbrecht, B. V. (2008). Complement evasion by human pathogens. *Nat. Rev. Microbiol.* 6, 132. doi:10.1038/nrmicro1824

Liszewski, M. K., Leung, M. K., Hauhart, R., Fang, C. J., Bertram, P., and Atkinson, J. P. (2009). Smallpox inhibitor of complement enzymes (SPICE): dissecting functional sites and abrogating activity. *J. Immunol.* 183, 3150–3159. doi:10.4049/jimmunol.0901366

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 47, W636–W641. doi:10.1093/nar/gkz268

Mohan, R. R., Gorham, R. D., and Morikis, D. (2015). A theoretical view of the C3d:CR2 binding controversy. *Mol. Immunol.* 64, 112–122. doi:10.1016/j.molimm.2014.11.006

Mohan, R. R., Huber, G. A., and Morikis, D. (2016). Electrostatic steering accelerates C3d:CR2 association. *J. Phys. Chem. B* 120, 8416–8423. doi:10.1021/acs.jpcb.6b02095

Ojha, H., Panwar, H. S., Gorham, R. D., Morikis, D., and Sahu, A. (2014). Viral regulators of complement activation: structure, function and evolution. *Mol. Immunol.* 61, 89–99. doi:10.1016/j.molimm.2014.06.004

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

Reis, E. S., Mastellos, D. C., Hajishengallis, G., and Lambris, J. D. (2019). New insights into the immune functions of complement. *Nat. Rev. Immunol.* 19, 503–516. doi:10.1038/s41577-019-0168-x

Resolution of the World Health Assembly. Global Smallpox Eradication. WHA resolution 33. 4, 1980: https://www.who.int/publications/i/item/WHA33-4

Rosbjerg, A., Genster, N., Pilely, K., and Garred, P. (2017). Evasion mechanisms used by pathogens to escape the lectin complement pathway. *Front. Microbiol.* 8, 868. doi:10.3389/fmicb.2017.00868

SAVES (2020). SAVES v5.0—DOE-MBI structure lab UCLA. Available at: https://servicesn.mbi.ucla.edu/SAVES/ (Accessed December 22, 2020).

Sfyroera, G., Katragadda, M., Morikis, D., Isaacs, S. N., and Lambris, J. D. (2005). Electrostatic modeling predicts the activities of orthopoxvirus complement control proteins. *J. Immunol.* 174, 2143–2151. doi:10.4049/jimmunol.174.4.2143

Sitkoff, D., Sharp, K. A., and Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* 98, 1978–1988. doi:10.1021/j100058a043

Søndergaard, C. R., Olsson, M. H., Rostkowski, M., and Jensen, J. H. (2011). Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pKa values. *J. Chem. Theor. Comput.* 7, 2284–2295. doi:10.1021/ct200133y

Trindade, G. S., Emerson, G. L., Carroll, D. S., Kroon, E. G., and Damon, I. K. (2007). Brazilian vaccinia viruses and their origins. *Emerg. Infect. Dis.* 13, 965–972. doi:10.3201/eid1307.061404

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., et al. (2018). MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci.* 27, 293–315. doi:10.1002/pro.3330

Wu, J., Wu, Y. Q., Ricklin, D., Janssen, B. J., Lambris, J. D., and Gros, P. (2009). Structure of complement fragment C3b-factor H and implications for host protection by complement regulators. *Nat. Immunol.* 10, 728–733. doi:10.1038/ni.1755

Xue, X., Wu, J., Ricklin, D., Forneris, F., Di Crescenzio, P., Schmidt, C. Q., et al. (2017). Regulator-dependent mechanisms of C3b processing by factor i allow differentiation of immune responses. *Nat. Struct. Mol. Biol.* 24, 643–651. doi:10.1038/nsmb.3427

Yadav, V. N., Pyaram, K., Ahmad, M., and Sahu, A. (2012). Species selectivity in poxviral complement regulators is dictated by the charge reversal in the central complement control protein modules. *J. Immunol.* 189, 1431–1439. doi:10.4049/jimmunol.1200946

Yadav, V. N., Pyaram, K., Mullick, J., and Sahu, A. (2008). Identification of Hot spots in the Variola virus complement inhibitor (SPICE) for human complement regulation. *J. Virol.* 82, 3283–3294. doi:10.1128/JVI.01935-07

Zewde, N., Mohan, R. R., and Morikis, D. (2018). Immunophysical evaluation of the initiating step in the formation of the membrane attack complex. *Front. Physiol.* 6, 130. doi:10.3389/fphy.2018.00130

Zhang, L., and Morikis, D. (2006). Immunophysical properties and prediction of activities for vaccinia virus complement control protein and smallpox inhibitor of complement enzymes using molecular dynamics and electrostatics. *Biophys. J.* 90, 3106–3119. doi:10.1529/biophysj.105.068130

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 © 2021 Narkhede, Gautam, Hsu, Rodriguez, Zewde, Harrison, Arantes, Gaieb, Gorham, Kieslich, Morikis, Sahu and Palermo. 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 Molecular Biosciences

REVIEW
published: 17 March 2021
doi: 10.3389/fmolb.2021.645527

Bottom-Up Coarse-Grained Modeling of DNA

Tiedong Sun[1], *Vishal Minhas*[1], *Nikolay Korolev*[1], *Alexander Mirzoev*[1], *Alexander P. Lyubartsev*[2]* *and Lars Nordenskiöld*[1]*

[1] School of Biological Sciences, Nanyang Technological University, Singapore, Singapore, [2] Department of Materials and Environmental Chemistry, Stockholm University, Stockholm, Sweden

Recent advances in methodology enable effective coarse-grained modeling of deoxyribonucleic acid (DNA) based on underlying atomistic force field simulations. The so-called bottom-up coarse-graining practice separates fast and slow dynamic processes in molecular systems by averaging out fast degrees of freedom represented by the underlying fine-grained model. The resulting effective potential of interaction includes the contribution from fast degrees of freedom effectively in the form of potential of mean force. The pair-wise additive potential is usually adopted to construct the coarse-grained Hamiltonian for its efficiency in a computer simulation. In this review, we present a few well-developed bottom-up coarse-graining methods, discussing their application in modeling DNA properties such as DNA flexibility (persistence length), conformation, "melting," and DNA condensation.

Keywords: DNA condensation, coarse-grained model, molecular renormalization group, inverse Monte Carlo, multi-scale coarse-graining, force matching, relative entropy, persistence length

OPEN ACCESS

Edited by:
Fabio Trovato,
Freie Universität Berlin, Germany

Reviewed by:
Yong Wang,
University of Copenhagen, Denmark
Valentina Tozzini,
Istituto Nanoscienze, Consiglio Nazionale delle Ricerche, Italy

***Correspondence:**
Lars Nordenskiöld
LarsNor@ntu.edu.sg
Alexander P. Lyubartsev
alexander.lyubartsev@mmk.su.se

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *23 December 2020*
Accepted: *22 February 2021*
Published: *17 March 2021*

Citation:
Sun T, Minhas V, Korolev N, Mirzoev A, Lyubartsev AP and Nordenskiöld L (2021) Bottom-Up Coarse-Grained Modeling of DNA. Front. Mol. Biosci. 8:645527. doi: 10.3389/fmolb.2021.645527

1. INTRODUCTION

Deoxyribonucleic acid (DNA) is the genetic information carrier of higher living organisms. To pass the encoded information from generation to generation, it has to allow efficient duplication and compaction. We still lack a full understanding of many fundamental aspects of DNA physics that determine DNA properties and function. To name a few, double-helical DNA encounters topological difficulties during replication (Postow et al., 2001) and compaction (Schiessel, 2003). Physically, curvature and elasticity of the DNA double helix are crucial for DNA compaction into chromatin and chromosomes, which is essential to cell division and gene expression regulation. Additionally, the conformational dynamics of DNA is crucial to its interaction with other macromolecules in the cell (Stelzl et al., 2017).

Furthermore, DNA-based nanoscale materials have attracted a large amount of attention in recent years. Particularly, the programmable design of DNA origami has found promising application in many fields, such as cancer therapy (Rajagopalan and Yakhmi, 2017). Hence, the knowledge of how DNA mechanics operates from atomistic level (nanometer) to macromolecule level (micrometer) will undoubtedly advance our understanding of the machinery of living organisms and grant us more control in designing nanomaterials and nanomachines based on DNA.

As an indispensable tool in DNA study, molecular modeling has advanced considerably in the past two decades. Developers continuously improve the extensively used all-atom (AA) force field models. Many AA force field deficiencies were exposed and subsequently corrected after microsecond-long AA simulations become available (Hart et al., 2012; Galindo-Murillo et al., 2016).

With extensive simulations, we have gained atomic-level insights on DNA dynamics (Lavery et al., 2009), DNA flexibility (Minhas et al., 2020), and other physical properties of DNA.

Despite the success of AA DNA models in studying DNA dynamics, flexibility, and binding properties, many important questions, such as nucleosome organization within chromatin that package DNA in the eukaryotic cell nucleus, are still out of their reach. With recent advances in multi-microsecond AA simulations, partial DNA unwrapping from the histone core of the nucleosome can be characterized in detail (Shaytan et al., 2016). However, other important relative motion between DNA and the histone proteins of the histone octamer (HO) that wraps DNA in nucleosomes (Kono and Ishida, 2020), such as sliding, is still beyond the capability of AA-MD. Lowering the model's resolution by merging groups of highly correlated atoms to individual interaction sites is an effective way to study these features. These low-resolution models are usually called coarse-grained (CG) models. In such models, the fluctuations of unimportant degrees of freedom are implicit. Furthermore, with a low number of degrees of freedom, we effectively deal with a smoother free energy surface of the molecular system. Consequently, CG models speed up the simulation in two ways–there are fewer interacting particles in the simulation system–and faster dynamics due to the smoother free energy surface. Within CG models, we can observe molecular dynamics at a large temporal and spatial scale using widely available computational resources, such as workstations and freely accessible simulation software packages.

The development of CG models is not always straightforward, and it remained something of an art rather than rigorous science for a long time. For simple molecular systems, such as liquids of small molecules, the CG model is usually straightforward with a small number of parameters. It is convenient to tune these parameters to reproduce correct macroscopic quantities, such as density and/or surface tension. It is widely known as "top-down" modeling. As the need for a complex model arises, it is usually challenging, if not impossible, to complete the modeling with only top-down approaches as the number of parameters in the model could be exceedingly large. There is not enough information for all parameters to be determined with high confidence.

Another CG modeling approach relies on the lower level, higher resolution models, commonly referred to as fine-grained (FG) models. These so-called "bottom-up" methods achieve the desired modeling by averaging out the unimportant degrees of freedom. For instance, water molecules are usually implicit in CG modeling of proteins, while their effects on, for example, protein conformation is implicitly included in the effective Hamiltonian of the CG model. Therefore, the central task of bottom-up CG modeling is to derive all parameters of the target CG model with the information presented by an underlying fine-grained model, usually an AA model.

Mathematically, bottom-up coarse-graining strives to project a higher dimensional Hamiltonian to a lower-dimensional space, with the requirement that a certain set of properties is maintained. This set of properties is defined by carefully choosing an objective function that determines the properties of interest and their relative weights. Practitioners have a number of algorithms based on statistical mechanics tackling this projection with various objective functions. Some approaches try to preserve interaction forces by averaging out the fast degrees of freedom by integration (Noid et al., 2008). Some minimize the information loss in the coarse-graining process, characterized by a relative entropy (Shell, 2008, 2016; Chaimovich and Shell, 2011). Others aim to reproduce structural features of the molecular system represented by the radial distribution functions (RDF) (Lyubartsev and Laaksonen, 1995; Soper, 1996). It is worth noting that the objective function does not necessarily consist of elementary physical quantities such as force or particle correlation. It can be a combination of various quantities with predetermined weights, such as distance distributions (Leonarski et al., 2013), force variations (Zhang et al., 2018; Wang et al., 2019), and entropy loss (Wang and Gomez-Bombarelli, 2019). Even macroscopic quantities can be included in such designed objective function to accommodate a wide range of modeling goals. These modeling efforts with designed objective functions gained more popularity in the past decade as machine learning approaches became common. Nevertheless, choosing or designing a suitable objective function is the most critical part in such cases, as the optimization methods are readily available using machine learning algorithms. This process is something of an art, requiring experience in solving such modeling problems. Generally, all CG modeling efforts can be described by a common framework as depicted in **Figure 1**. The key elements in the framework are the objective function, the optimization algorithm, and the simulation engine. The simulation engine produces simulation results that are evaluated by the objective function with respect to the targeting reference data, before the trial CG model being subjected to optimization. The modeling is successful when the quality of the CG model is good enough judged by the objective function. In this review, we will focus on bottom-up coarse-graining algorithms originating from statistical mechanics. Readers interested in novel objective functions and modeling with machine learning are referred to the review by Gkeka et al. (2020).

Though mathematically well-defined, the bottom-up coarse-graining process is not straightforward in practice, especially for complex molecular models, such as DNA and proteins. Many aspects of interactions and configurations of DNA could collectively determine one specific property, e.g., bending flexibility. It is challenging to model all these kinds of interactions accurately at the same time. Nevertheless, many coarse-grained DNA models, based on atomistic force fields have been developed, following the bottom-up philosophy. In this review, we review the application of bottom-up coarse-graining methods for studying and understanding DNA properties. We will first introduce bottom-up modeling methods in section 2. Some selected representative bottom-up DNA models will be summarized to give an overview of its recent development. The pros and cons of bottom-up modeling different properties of DNA will be presented. Lastly, we will summarize the current stage of bottom-up DNA models and discuss the future development of bottom-up coarse-graining of DNA. DNA is a highly charged polyelectrolyte. The long-range electrostatic

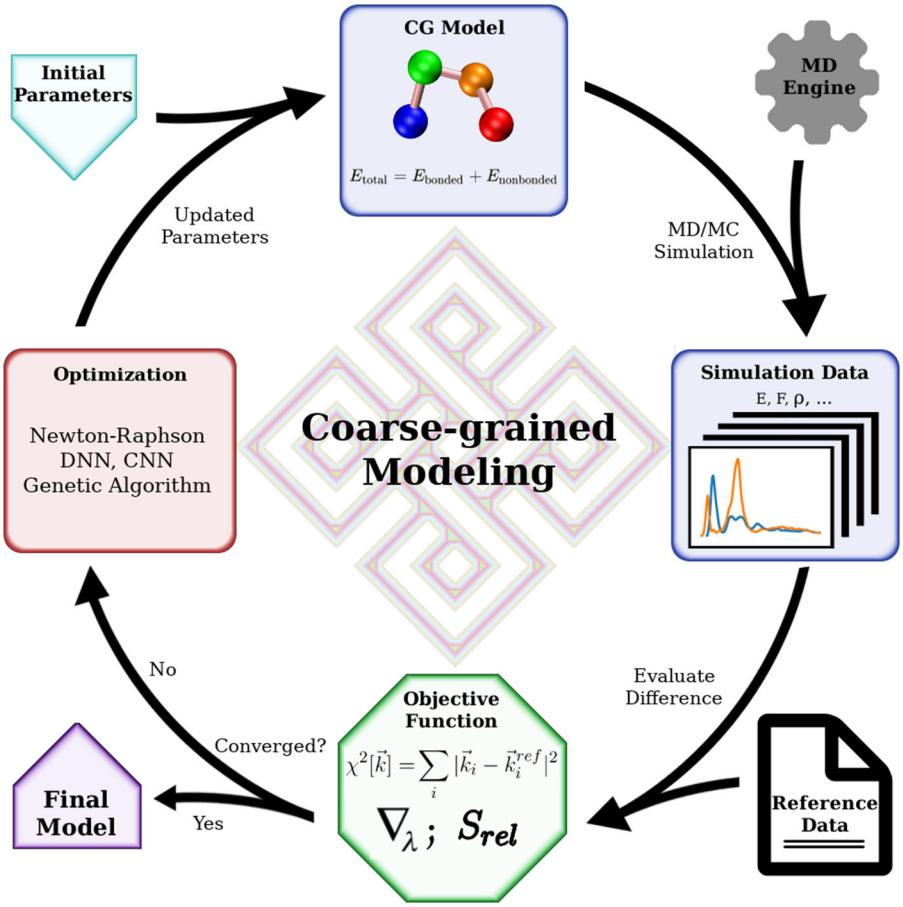


FIGURE 1 | General workflow of coarse-grained modeling. Essential tools for coarse-grained modeling include a software that performs simulation, an objective function to evaluate simulation data and an optimization algorithm to improve the CG model. The final model, which will reproduce the reference data, is obtained when the objective function is satisfied. When the reference data is extracted from fine-grained model simulations, this flow chart represents a general bottom-up coarse-graining algorithm. When the reference data is macroscopic quantities obtained in experiments, this flow chart indicates a top-down coarse-graining procedure. Hybrid coarse-grained modeling is achieved with reference data from both lower-level models and experiments, with a corresponding objective function.

interactions between DNA, small mobile ions (due to salt), and other charged small molecules, as well as biomacromolecules, dominates most, if not all, their physicochemical properties (Bloomfield et al., 2000). Dramatic salt effects and the strong influence of the valence of counterions on DNA physical properties are observed. In some cases, it leads to counter-intuitive behavior such as like-charged attraction between the DNA polyions (Guldbrand et al., 1986; Nordenskiold et al., 2008; Korolev et al., 2010, 2016). Bottom-up approaches represents arguably the most rigorous way of extracting the effective electrostatic potentials between the charged CG sites.

2. BOTTOM-UP COARSE-GRAINING

2.1. Theoretical Background

In physical terms, bottom-up coarse-graining is the process of removing unimportant degrees of freedom (DOF) from a detailed high-resolution model and formulation of a simpler model, which contains only essential DOFs. Assume that at the high-resolution level, the system is described by a Hamiltonian (potential energy) $H(\mathbf{q_1}, \cdots, \mathbf{q_n})$, where $\{\mathbf{q_i}; i = 1, \cdots, n\}$ are coordinates of particles. The potential energy function typically represents the atomistic force field. However, it can be the potential energy of an already existing CG model, or in the case of *ab initio* modeling, can represent the energy surface obtained from quantum chemical computations. Coarse-graining is described in terms of mapping of FG coordinates (degrees of freedom) $\{\mathbf{q_i}; i = 1, \cdots, n\}$ to CG coordinates $\{\mathbf{Q_j}; j = 1, \cdots, N\}$ with $N \ll n$, which is mathematically expressed as mapping functions M:

$$\mathbf{Q_j} = M_j(\mathbf{q_1}, \cdots, \mathbf{q_n}) \tag{1}$$

Generally important DOF, represented by coarse-grained sites, can be chosen in different ways, often based on experience grounded in chemical and physical intuition. Typically one chooses CG sites according to the center-of-masses (COM) of the atom groups forming the CG units, while other choices

(such as taking coordinates of specific atoms) are also possible. For instance, one can aim to minimize the information loss due to the mapping operation (Giulini et al., 2020), or choose beads representing collective motions (Zhang et al., 2008). The Hamiltonian $H(\mathbf{q_1}, \cdots, \mathbf{q_n})$ defines all properties of the high-resolution system and, through the mapping functions (Equation 1), all properties of the CG system. The task of the bottom-up approach is to define the effective interaction potentials for the CG sites, denoted $H_{CG}(\mathbf{Q_1}, \cdots, \mathbf{Q_N})$, which provides the same properties for the CG system as the properties defined by the FG Hamiltonian $H(\mathbf{q_1}, \cdots, \mathbf{q_n})$ through the CG mapping. In other words, bottom-up modeling is the practice of solving the inverse problem: to determine the CG interaction potential that reproduces known properties obtained from simulations of the FG system.

The CG Hamiltonian, which satisfies the consistency condition, can be deduced from the FG Hamiltonian by integrating over non-interesting degrees of freedom: (Noid et al., 2008; Lyubartsev et al., 2015)

$$H_{CG}(\mathbf{Q_1}, \cdots, \mathbf{Q_N}) = -\frac{1}{\beta} \ln \int \prod_{i=1}^{n} d\mathbf{q_i} \prod_{j=1}^{N} \delta(\mathbf{Q_j} - M_j(\mathbf{q_1}, \cdots, \mathbf{q_n})) \\ exp(-\beta H(\mathbf{q_1}, \cdots, \mathbf{q_n})) + C \quad (2)$$

The above coarse-grained Hamiltonian, H_{CG}, is also known as an N-body potential of mean force. It provides precisely the same structural properties for the CG model as the FG system mapped by Equation (1) to the CG representation. Thermodynamic properties (average energy, free energies, pressure) can also be reproduced, given the fact that the original and CG Hamiltonians have the same partition function. However, a caveat is that the CG Hamiltonian depends on the thermodynamic conditions (temperature, volume, or density). These dependencies need to be considered while obtaining thermodynamic properties by taking the derivative of the partition function with respect to thermodynamic parameters (Lyubartsev, 2018). Reconstruction of correct dynamics in the CG system is a more challenging task, which leads to a generalized Langevin equation with a memory function (Romiszowski and Yaris, 1991). Approximately, dynamics can be reconstructed within a dissipative particle dynamics approach implementing the Mori-Zvanzig formalism (Eriksson et al., 2008; Hijon et al., 2010) or by normal Langevin dynamics with an appropriately chosen friction constant.

However, modeling a CG system using an N-body potential (Equation 2) is practically impossible. In all implementation of bottom-up coarse-graining, one resorts to simpler functions such as additive pair potentials. Hence, the task is reformulated into finding the best possible approximation to the exact Hamiltonian in Equation (2), in the form of additive pair interactions:

$$H_{CG}(\mathbf{Q_1}, \cdots, \mathbf{Q_N}) \approx \Sigma_{i,j} U_{ij}(R_{ij}) \quad (3)$$

where R_{ij} is the distance between CG sites i and j.

We note that the use of only pair potentials is not a restriction. Other types of interactions can be included in Equation (3). For example, angle (3-body) or torsion (4-body) potential terms are commonly used for macromolecular CG models. Other forms expressing multi-body interactions can also be included as long as they can be handled efficiently by the simulation software. The task of building a CG force field can be reformulated into finding "as good as possible" an approximation to the exact many-body potential according to Equation (3). For instance, finding the best fit of forces coming from both sides of Equation (3) gives rise to the force-matching method (Ercolessi and Adams, 1994; Izvekov et al., 2004), also known as the Multi-Scale Coarse-Graining (MS-CG) method (Izvekov and Voth, 2005). Another approach that rests on minimizing the entropy difference between FG and CG models corresponds to the relative entropy minimization method (Shell, 2008).

All bottom-up approaches can be approximately divided into two categories: thermodynamics-based coarse-graining aiming at a reproduction of thermodynamic properties (free energies, average forces), and structure-based coarse-graining, aiming at the reproduction of structural properties of the FG system. The theoretical justification of structure-based coarse-graining is the Henderson theorem (Henderson, 1974) that defines a one-to-one relationship between a set of radial distribution functions (RDF) and a set of pair potentials for CG sites. Rudzinski and Noid (2011) later generalized the Henderson theorem to include multiple RDFs between different types of CG sites and intramolecular structural properties such as bond lengths and angles distributions. Below we present several well-developed bottom-up CG modeling methods and discuss the connections among them.

2.2. Iterative Boltzmann Inversion

We first discuss the simplest algorithm of bottom-up coarse-graining, which is Iterative Boltzmann Inversion (IBI). This method is usually categorized as structure-based coarse-graining, where the pair potential approximation (Equation 3) is fitted to reproduce various distribution functions obtained in atomistic simulations. Common target distributions include radial distribution functions (RDF), bond length distributions, angle value distributions, etc. The IBI approach is implemented through an iterative algorithm (Soper, 1996; Reith et al., 2003). The pair interaction potential is in each iteration updated with a correction term originating from the mean-field approximation:

$$U_{\alpha\beta}^{i+1} = U_{\alpha\beta}^{i} - k_B T \ln \left(\frac{g_{\alpha\beta}^{i}(r)}{g_{\alpha\beta}^{ref}(r)} \right) \quad (4)$$

The pair potential between site types α and β used in iteration $i+1$ is obtained by applying the correction term (second term on the right-hand side) to the pair potential in iteration i. To determine the correction, the pair RDF, $g_{\alpha\beta}^{i}(r)$, is obtained through proper sampling with the current interaction potential, $U_{\alpha\beta}^{i}$, and subsequently compared with the reference RDF, $g_{\alpha\beta}^{ref}(r)$, from an FG simulation. The iterative algorithm is started with a bootstrapping potential, usually, a simple potential of mean force: (Soper, 1996; Reith et al., 2003)

$$U_{\alpha\beta}^{0}(r) = -k_B T \ln g_{\alpha\beta}^{ref}(r) \quad (5)$$

The convergence of the effective potential is expected when the correction term approaches zero, which also means (Equation 4) that the CG RDF is nearly equal to the RDF of the reference FG simulations. Thus, the inverse problem is solved, and the effective potential is obtained.

Correction of the potential according to Equation (4) is straightforward to implement. This correction is determined only by the value of the same distribution function, while correlations between different interaction terms are completely neglected. As a result, the IBI approach often faces convergence problem even for relatively simple systems such as ion solutions (Hess et al., 2006), where RDFs between different pairs of anions and cations are strongly correlated with each other. In the practical calculation of CG potentials by RDF inversion, it might be instructive to start the iterative process using the IBI approach. This brings the system RDFs close to the reference values, after which one may switch to other algorithms, which take into account correlations between different interaction terms and provide better convergence when the RDFs are close to the reference functions.

2.3. Inverse Monte Carlo

The Inverse Monte Carlo (IMC) method (Lyubartsev and Laaksonen, 1995; Lyubartsev, 2018) (also known as Newton Inversion; Lyubartsev et al., 2010) is a general method to invert ensemble averages, and particularly RDFs, to effective pair potentials. For any multi-component system, it goes through an inverse process and produces as output the effective pair potentials between CG sites, which in direct CG simulations reproduce the same RDFs as those obtained in the detailed FG simulations.

Within the IMC approach, both RDFs and interaction potentials are discretized into two sets of values: histogram of particle-particle distances $\{G_\alpha; \alpha = 1, \cdots, m\}$, which after ensemble averaging and normalization to bulk particle density yields RDF, and tabulated pair potentials $\{U_\alpha; \alpha = 1, \cdots, m\}$, which are determined for the same set of distances as the RDFs. In standard simulations (direct problem), we have the interaction potential U as input, and by running MC or MD simulation, we can evaluate averages $\{\langle G_\alpha \rangle\}$ and thus obtain the RDF as output. In coarse-graining by IMC, we solve the inverse task: from averages determined in FG simulations ($\{\langle G_\alpha^{ref} \rangle\}$) we determine the CG effective potential U.

The solution to this non-linear inverse problem can be obtained iteratively by the Newton-Raphson method (hence named the Newton Inversion). Let us determine the Jacobian of the $G(U)$ dependence by defining its elements:

$$J_{\alpha\gamma} = \frac{\partial \langle G_\alpha \rangle}{\partial U_\gamma} \tag{6}$$

This Jacobian (an $m \times m$ matrix) determines how changes of the potential are related to the changes of the RDF:

$$\Delta \vec{G} = J \Delta \vec{U} \tag{7}$$

where we use vector notations for the sets of values of the potential and RDF. Respectively, corrections to the potential, which produce the desired changes of the RDF are determined by the inverse matrix:

$$\Delta \vec{U} = J^{-1}(\vec{G} - \vec{G}^{ref}) \tag{8}$$

The Jacobian itself can be computed in Monte Carlo or MD simulations by (Lyubartsev and Laaksonen, 1995):

$$\frac{\partial \langle G_\alpha \rangle}{\partial U_\gamma} = -\frac{1}{k_B T}(\langle G_\alpha G_\gamma \rangle - \langle G_\alpha \rangle \langle G_\gamma \rangle) \tag{9}$$

Recall that G is discretized RDF and U is discretized potential energy, as formulated in IMC and implemented in the MagiC software (Mirzoev et al., 2019). More generally, G can be any set of observables, and U_γ can represent an arbitrary parameter of potential energy. The $G(U)$ dependence in this general sense can be computed in similar ways as in Equation (9), which has been used in the molecular renormalization group (Savelyev and Papoian, 2009a,b, 2010) and ForceBalance (Wang et al., 2014) methods.

Equations (8) and (9) allow us to solve the inverse problem by an iterative procedure. We start from a trial set of potential. In practical simulations, one can start either from zero or from the pair potentials of mean force and then run an MC simulation, followed by computation of RDFs expressed in terms of $\langle G_\alpha \rangle$, as well as the cross-correlation terms according to Equation (9). It is followed by inverting the Jacobian defined by Equation (6) (which solves the corresponding system of linear equations) to obtain corrections to the interaction potential U. The procedure is repeated until convergence.

In practical computations, the direct use of Equations (8) and (9) may lead to divergence since the method is based on linear extrapolation (Equation 7) of a generally non-linear relationship. A simple way to regularize the procedure is to iterate by "small steps" to ensure staying in the linear regime, i.e., to multiply the difference of RDF, $\vec{G} - \vec{G}^{ref}$, by a scaling factor $0 < \lambda < 1$.

Dealing with long-range electrostatic interactions, which are important in the coarse-graining of highly charged systems such as DNA, requires special considerations. In the first application of IMC to the ionic solution (Lyubartsev and Laaksonen, 1995), the electrostatic part of the interaction was separately treated in the simulations. The non-electrostatic part of the interactions is expected to be of short range. While the electrostatic interaction is considered invariable, the short-range interaction is optimized by the IMC procedure to reproduce the RDF within the specified cut-off distance.

2.4. Molecular Renormalization Group

Savelyev and Papoian (2009a,b) have proposed an approach similar to IMC bottom-up coarse-graining inspired by the renormalization group theory. The so-called Molecular Renormalization Group Coarse-Graining (MRG-CG) relies on iteratively updating the trial interaction potential, using a Jacobian identical to Equation (6). The novelty is that the Jacobian is obtained by

$$\frac{\partial \langle G_\alpha \rangle}{\partial K_\gamma} = -\frac{1}{k_B T}\left[\left\langle G_\alpha \frac{\partial \mathcal{U}}{\partial K_\gamma} \right\rangle - \langle G_\alpha \rangle \left\langle \frac{\partial \mathcal{U}}{\partial K_\gamma} \right\rangle\right] \tag{10}$$

where K_γ represents an arbitrary parameter of the potential $\mathcal{U}$. It is apparent that when the Hamiltonian takes the form $U = \sum_\alpha K_\alpha G_\alpha$, Equation (10) becomes Equation (9). Therefore, MRG-CG is a generalization of IMC.

In the original implementation of IMC (Lyubartsev and Laaksonen, 1995), tabulated potentials were used as the default functional form of the coarse-grained potential, as can be expressed as a collection of delta-functions:

$$U_{CG} = \sum_i^m K_i \delta_i(\mathbf{r} - \mathbf{r}_i) \tag{11}$$

where i goes through all m entries of the tabulated potential. In other words, the CG Hamiltonian is expanded on a delta-function basis set, usually with a large number of parameters $\{K_i\}$. With the generalization introduced by MRG-CG, one can choose a convenient basis set with much fewer parameters for expanding the Hamiltonian. High efficiency is expected if the chosen basis set is optimal.

2.5. Relative Entropy Minimization

In this approach, a quantity called the relative entropy based on information theory was proposed to be used as a minimization objective in bottom-up coarse-graining. The relative entropy is defined as (Chaimovich and Shell, 2011)

$$\begin{aligned} S_{rel} &= \sum_i \mathbb{P}_{FG}(i) \ln \frac{\mathbb{P}_{FG}(i)}{\mathbb{P}_{CG}(M(i))} + \langle S_{map} \rangle_{FG} \\ S_{map}(I) &= \ln \sum_i \delta_{I,M(i)} \end{aligned} \tag{12}$$

where the summation goes over all configurations $\{i\}$ in FG ensemble, $\mathbb{P}(i)$ is the probability of configuration i in FG ensemble, *FG* and *CG* denote FG reference quantities and CG quantities respectively. Note that the mapping entropy S_{map} is not dependent on the CG effective potential. It represents the entropy loss of mapping a set of FG microstates $\{i\}$ into one single CG microstate I. Under the canonical ensemble, we have the following expression for relative entropy after substituting configurational probabilities:

$$S_{rel} = \beta \langle U_{CG} - U_{FG} \rangle_{FG} - \beta (A_{CG} - A_{FG}) + \langle S_{map} \rangle_{FG} \tag{13}$$

Here, $A = -k_B T \ln Z$ is the Helmholtz free energy, and all averaging is performed under the FG ensemble. Note that the free energy term A_{CG} is evaluated over the FG ensemble, with the operation of the mapping function and the CG potential function. Hence, the term $-\beta(A_{CG} - A_{FG})$ characterizes the free energy difference over the same FG ensemble, but in representations of respective CG and FG models.

From an information theory point of view, relative entropy quantifies the information loss in the coarse-graining process. A good CG model is expected to have minimal information loss when compared to its FG reference model. Therefore, we minimize the relative entropy S_{rel} to generate a CG model with respect to the adjustable parameters in CG Hamiltonian.

Minimizing relative entropy with respect to an arbitrary parameter λ of CG Hamiltonian requires

$$\begin{aligned} \frac{\partial S_{rel}}{\partial \lambda} &= \beta \left\langle \frac{\partial U_{CG}}{\partial \lambda} \right\rangle_{FG} - \beta \left\langle \frac{\partial U_{CG}}{\partial \lambda} \right\rangle_{CG} = 0 \\ \frac{\partial^2 S_{rel}}{\partial \lambda^2} &= \beta \left\langle \frac{\partial^2 U_{CG}}{\partial \lambda^2} \right\rangle_{FG} - \beta \left\langle \frac{\partial^2 U_{CG}}{\partial \lambda^2} \right\rangle_{CG} + \beta^2 \left\langle \left(\frac{\partial U_{CG}}{\partial \lambda} \right)^2 \right\rangle_{CG} \\ &\quad - \beta^2 \left\langle \frac{\partial U_{CG}}{\partial \lambda} \right\rangle_{CG}^2 > 0 \end{aligned} \tag{14}$$

Similar to other coarse-graining algorithms, the minimization of relative entropy is achieved through iterations. With a simple update rule (Shell, 2008) as:

$$\lambda^{i+1} = \lambda^i - \frac{\partial S_{rel}/\partial \lambda}{\partial^2 S_{rel}/\partial \lambda^2} \tag{15}$$

the parameter of Hamiltonian, λ, is updated between iteration i and $i+1$ by the negative ratio of the first derivative to the second derivative of S_{rel} with respect to λ. Here, the minimization of S_{rel} is implemented as a Newton-Raphson iterative algorithm. In the case of U_{CG} being linear in λ, i.e., $U_{CG}(\mathbf{Q}^N) = \lambda f(\mathbf{Q}^N) + \cdots$, where $f(\mathbf{Q}^N)$ is a function depending on CG coordinates, Equation (15) becomes

$$\lambda^{i+1} = \lambda^i - k_B T \frac{\langle f \rangle_{FG} - \langle f \rangle_{CG}}{\langle f^2 \rangle_{CG} - \langle f \rangle_{CG}^2} \tag{16}$$

We may note here that when expanding the CG Hamiltonian with a linear basis set, the correction term in Equation (16) is calculated through the correlation of f. It indicates connections to structure-based coarse-graining. We will discuss this further in section 2.8.

2.6. Force Matching

Initially proposed by Ercolessi and Adams (1994), the force matching method was later given a solid theoretical basis by Izvekov et al. (Izvekov et al., 2004; Izvekov and Voth, 2005; Noid et al., 2008). The objective in force matching is to minimize the difference in the force as quantified by a functional of the force residual:

$$\chi^2[\mathbf{F}] = \frac{1}{3N} \left\langle \sum_{I=1}^N |\mathbf{f}_I(\mathbf{q}^n) - \mathbf{F}_I(M_{\mathbf{Q}}^N(\mathbf{q}^n))|^2 \right\rangle \tag{17}$$

In this way, given a CG mapping $M_{\mathbf{Q}}^N(\mathbf{q}^n)$, the forces given by the FG force field, $\mathbf{f}_I(\mathbf{q}^n)$, as exerted on atoms forming the CG site I, are replaced by the force given by the CG force field, $\mathbf{F}_I(\mathbf{Q}^N)$. As in other coarse-graining methods, a pairwise basis set, $\{\phi_2(R_{IJ})\}$, is usually adopted to expand the CG force field for efficiency:

$$\mathbf{F}_I(\mathbf{Q}^N) = \sum_{I \neq J} \phi_2(R_{IJ}) \hat{e}_{IJ} = \sum_k c_k \sum_{I \neq J} u_k(R_{IJ}) \hat{e}_{IJ} \tag{18}$$

where $\{u_k\}$ is a set of B-spline functions, $\hat{e}_{IJ}$ is the unit vector pointing from bead I to bead J. The minimization problem of the

force residual functional, with respect to a parameter set $\{c_k\}$, can be solved in a variational manner. In the further development of the force matching method, regularization (Lu et al., 2010) and an iterative algorithm (Lu et al., 2013) were introduced to improve accuracy in reproducing structure correlations.

When working with pair-wise additive potential, the force matching method can also be accomplished by solving a linear system. Early implementation of the force matching algorithm, employed discrete delta functions to represent forces in the form of (Noid et al., 2007)

$$f(r) = \sum_{d}^{N_d} f_d \delta_D(r - r_d) \tag{19}$$

with $\delta_D(r) = 1$ when $-\Delta r/2 \leq r < \Delta r/2$, and $\delta_D(r) = 0$ otherwise. By minimizing the force residual (Equation 17) with respect to the force table elements f_d, a linear equation system is obtained (Noid et al., 2007)

$$\sum_{d'} f_{d'} G_{dd'} = b_d \tag{20}$$

where

$$b_d = \left\langle \sum_i \sum_{j \neq i} (\vec{F}_i^{I,AA} \cdot \vec{u}_{ij}^I) \delta_D(r_{i,j}^I - r_d) \right\rangle_I \tag{21}$$

and

$$G_{dd'} = \left\langle \sum_i \sum_{j \neq i} \sum_{k \neq i,j} (\vec{u}_{ij}^I \cdot \vec{u}_{ik}^I) \delta_D(r_{ij}^I - r_d) \delta_D(r_{ik}^I - r_{d'}) \right\rangle_I \tag{22}$$

The symmetric matrix G contains all the information, up to three-body correlation, to connect the table elements of CG forces, f_d, and forces in the fine-grained ensemble, b_d.

2.7. Other Coarse-Graining Approaches

While bottom-up coarse-graining, as described, is a rigorous and self-consistent approach, the accuracy of the bottom-up CG model relies on the quality of the underlying fine-grained model. Deficiency in the all-atom force field could result in incorrect behavior of the derived CG model. For instance, in Maffeo et al. (2014), the CG model of single-stranded DNA obtained by IBI from atomistic simulations could not reproduce experimentally measured radius of gyration. A top-down refinement was subsequently applied to non-bonded interactions to improve the accuracy of the resulting CG model. Such a hybrid bottom-up – top-down approach is useful when fine-grained simulation cannot produce the correct ensemble due to either inaccuracy in the FG model or sampling difficulties. The hybrid approach can also be accomplished by constructing a hybrid objective function before optimization by machine learning algorithms (Leonarski et al., 2013; Zhang et al., 2018; Wang and Gomez-Bombarelli, 2019; Wang et al., 2019; Gkeka et al., 2020). The objective function contains contributions from both the fine-grained simulation and macroscopic measurements. An optimal model should be obtained even though the whole process is not trivial as many hyper parameters are involved in the machine learning algorithm and objective function.

Efforts were also made to derive models in ways similar to the development of classical atomistic force fields. In such practice, a model with a generalized representation of certain atom groups (CG site types) is produced, hoping that these types of CG sites can be used as building blocks in applications of modeling macromolecules. Instead of deducing interaction potential between atom types as in classical force fields, the potential of mean force among these CG site types is calculated through simulations of the moieties constituting the CG sites, using a fine-grained model. The extension of the MARTINI force field to DNA (Uusitalo et al., 2015) follows this modeling philosophy. It is still considered a bottom-up modeling approach since it is based on an atomistic force field model. However it is not a systematic approach, as we described above. Self-consistency is lost in such modeling practices.

While evaluating a CG model's quality, a comprehensive view should be taken to balance various criteria with sound reasons. One should recognize the shortcomings of bottom-up coarse-graining when the resulting CG model fails to reproduce experiments. For such cases, one may seek help from other approaches of deriving the CG model, e.g., by combination with a top-down approach.

2.8. Connection Among Bottom-Up Coarse-Graining Methods

The Henderson uniqueness theorem (Henderson, 1974; Rudzinski and Noid, 2011) states that for liquids with only pair-wise interactions, under given temperature and density, the pair-wise potential, which gives rise to a given radial distribution function, is unique up to a constant. As such, a connection is immediately clear among all structure-based coarse-graining methods, including IBI, IMC and MRG-CG. For an inverse modeling problem with a given set of RDFs and intramolecular distributions, all structure-based modeling methods will arrive at the same set of pair-wise potential up to numerical precision, as long as they can produce the potential as their answer. In practice, IBI has an inherent limitation of ignoring all correlations between pair-wise interactions, resulting in IBI being less capable of producing CG potentials with good precision, especially in complex systems with many CG site types. when the IBI approach can provide a satisfactory CG model, it is usually the most efficient one since the evaluation of the RDF produced by a trial potential is much faster than the evaluation of correlations between RDFs required by other methods.

Although relative entropy minimization is formulated from very different principles compared to structure-based coarse-graining, it has been shown that a CG system described by pair-wise potentials obtained by the relative entropy minimization reproduces the RDFs of the FG system (Chaimovich and Shell, 2011). Hence, relative entropy minimization leads to the same result in CG modeling as structure-based methods taking into account the Henderson uniqueness theorem. Indeed, similarities between relative entropy minimization and other modeling

TABLE 1 | Summary of coarse-grained DNA models.

DNA models	Representation	Main features	Modeling method	Achieved modeling target
IMC DNA	5 beads per two base pairs	Explicit ion; Stable double helical structure	IMC	Reproduces the bending persistence length of dsDNA; Reproduces multivalent ion-induced DNA aggregation
Fan Bonds Model	1 bead per nucleotide	Double helix is maintained by intra- and inter-strand bonds. Explicit ions	MRG-CG or IMC	Bending persistence length of dsDNA can be reproduced with scaled potential or specific bonding structure
"sugar" model	6 beads per nucleotide	Explicitly modeling the sugar ring conformation with a double-well bond; Explicit Ions	Boltzmann Inversion; Energy Relaxation; Empirical adjustments	Reproduced both A-DNA and B-DNA conformations
3SPN DNA Model	3 beads per nucleotide	Specific potential energy terms for base interactions	Top-down modeling for DNA; Relative entropy minimization for ion	Well reproducing DNA melting curve, shapes and curvature

methods have been point out by Chaimovich and Shell (2011). The equivalence becomes even more evident from the fact that in the practical implementation of the relative entropy minimization described by Shell (2008), the method involves inversion of the same cross-correlation matrix (Equations 9, 16) as in the IMC method.

The interconnection between the force matching approach and structure-based coarse-graining has been analyzed by Rudzinski and Noid (2011). It was shown that both approaches could be formulated in terms of an information function that discriminates between the ensembles generated by atomistic and CG models. While the relative entropy approach (and thus other structure-based CG methods) minimizes the average of the information function, the force matching method minimizes the average of its squared gradient. It is why force matching usually produces particle distributions different from the FG reference distributions in practice.

Furthermore, it was shown (Noid et al., 2007) that the kernel that describes the effects of three-particle correlations in the force matching method is equivalent to the kernel of the Yvon-Born-Green (YBG) equation, which relates equilibrium particle 2- and 3-body correlation functions to the additive pair-wise Hamiltonian. In later work by Lu et al. (2013), the force matching algorithm based on the YBG equation can be applied iteratively to reproduce the RDF of liquids as done in structure-based coarse-graining. Considering the Henderson uniqueness theorem, we see that the iterative force matching method (Lu et al., 2013) produces the same CG model if it is applied to minimize the RDF difference between the FG and CG models.

We see that, though developed from different theoretical backgrounds, there is a common theme among the discussed bottom-up coarse-graining methods. When a pair-wise additive potential is adopted, all methods reside in the generalized Yvon-Born-Green hierarchy, either directly, or through modifications as in the later development of force matching. More generally, as proposed by Shell (2008) and later elaborated by Rudzinski and Noid (2011), relative entropy is a fundamental quantity in bottom-up coarse-graining, which connects all structure-based methods and force-based methods.

3. COARSE-GRAINED DNA MODELS

Modeling of DNA, both *in vivo* and *in vitro*, is inherently a multiscale problem, which requires several levels of resolution, from atomistic (and perhaps quantum-chemical) to models describing higher-order structures of DNA in chromatin. Recent years have seen many CG DNA models differing by levels of details developed through both bottom-up and top-down approaches. Here we give a brief description of some models with emphasis on bottom-up models. A summary of these models is provided in **Table 1**.

3.1. IMC Model of dsDNA

We have developed a coarse-grained dsDNA model, focused on the dsDNA double helical shape and a strong emphasis on DNA electrostatic interactions. The first version of this model was proposed by Fan et al. (2013) as a component of a CG nucleosome core particle model. In this model, DNA is represented as a chain of two-base pair units. Each unit contains five CG beads representing two base pairs. Four of the five beads represent phosphate groups with $-1e$ charge, denoted "P" beads. The central bead, called the "D" bead, represents the other atoms of these two base pairs, namely four nucleosides, bearing zero charge. Four types of bonds are defined, between "D" bead and "P" beads in the same unit, between adjacent "P" beads on the same strand of DNA, between "D" beads of adjacent units, and lastly, between corresponding "P" beads across the minor groove. These fragments are put together to form a DNA chain with two helices formed by the phosphate groups (**Figure 2A**).

The total interaction potential consists of bonded and non-bonded interactions:

$$U_{tot} = \sum_{bonds} U_{bond}(r) + \sum_{angles} U_{ang}(\phi) + \sum_{i,j} (U_{el}(r_{ij}) + U^{SR}(r_{ij})) \tag{23}$$

where the electrostatic potential is modeled with screened Coulomb potential with $\epsilon = 78$

$$U_{el}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 \epsilon r_{ij}} \tag{24}$$

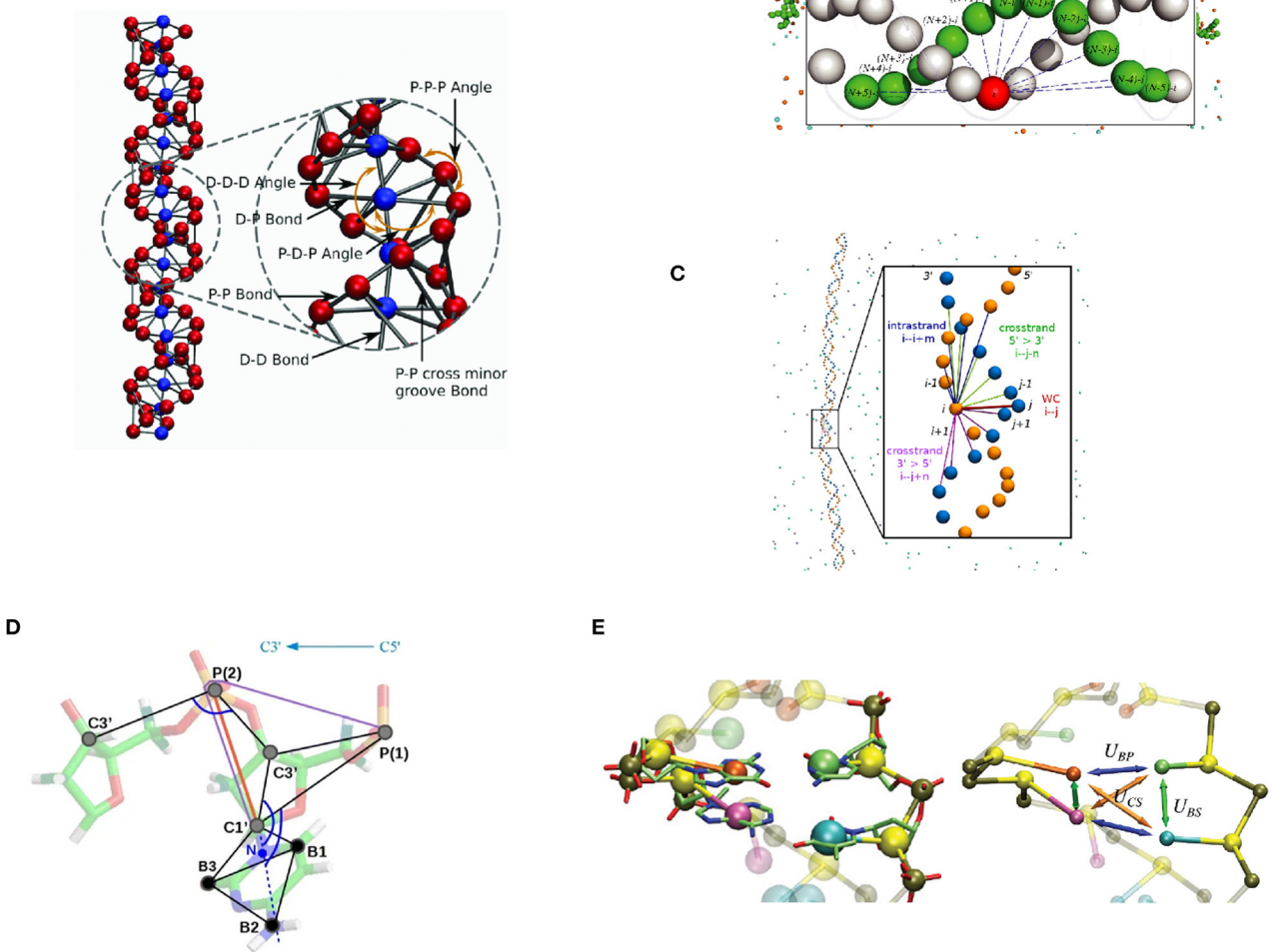


FIGURE 2 | Graphical representation of selected bottom-up coarse-grained DNA models. **(A)** the IMC DNA model (Sun et al., 2019; Minhas et al., 2020). Electrostatic interaction is modeled with explicit ion and phosphate beads (red), in addition to its apparent double-helical structure. Base-pair dynamics is simplified as it is not the primary modeling target. **(B)** Model by Savelyev and Papoian (2010). A double-helical structure is maintained by a series of inter-strand bonds. **(C)** Model by Naômé et al. (2014). Similar to **(B)**. **(D)** The "sugar" DNA model (Kovaleva et al., 2017). Sugar puckering is modeled with a double-well bond potential, so that it is capable of simulating A-DNA and B-DNA transition. **(E)** 3SPN model (Knotts et al., 2007; Sambriski et al., 2009; Hinckley et al., 2013; Freeman et al., 2014). Capable of simulating base-pair dynamics with its detailed Hamiltonian.

In the original formulation (Fan et al., 2013), the model was built within a top-down principle, with empirically chosen parameters for intra- and intermolecular interactions. Subsequently, Korolev et al. (2014) developed it into a standalone model of DNA in which intramolecular bonded interactions were parameterized by the IMC approach based on atomistic simulations of DNA with the CHARMM27 force field. For convenience, bonded potentials, $U_{bond}(r_{ij})$ and $U_{ang}(\phi)$, obtained by IMC in a tabulated form, were fitted to harmonic potentials and used in such form for subsequent CG simulations. This model has successfully reproduced the persistence length of dsDNA in a wide range of salt concentrations (Korolev et al., 2014).

In a recent effort to implement a full bottom-up model and extend the application range of this model, Sun et al. (2019) have recalculated all interaction potentials for DNA and a number of mono- and multivalent ions by IMC, both bonded and non-bonded. The fine-grained simulation was conducted with four segments of 36-bp dsDNA, described by the CHARMM27 force field. The length of the FG simulations is significantly longer than earlier versions of this model. Interaction potentials involving monovalent ions (Na^+, K^+, and Cl^-) and multivalent ions [$Mg(H_2O)_6^{2+}$ and $Co(NH_3)_6^{3+}$] are derived from the same FG simulations to perform explicit ion simulations on a larger scale with the CG model. Except for the

electrostatic potential, all potential terms are derived in tabulated form, as usually done within the IMC algorithm. The model showed its suitability to study multivalent ion-induced DNA aggregation while preserving its accuracy in DNA mechanical properties, i.e., the ion concentration-dependent persistence length (Minhas et al., 2020).

This IMC derived DNA model is designed to work with explicit ions. The model has proven its capabilities in reproducing the mechanical property of dsDNA and phase separation of DNA induced by multivalent ions. The topology of the model preserves the dsDNA's double-helical structure, though it excludes the capability of modeling base pair dynamics, such as bubbling and melting. We will discuss its merits and deficiencies in more detail in the following sections.

A further step of coarse-graining was conducted based on this CG DNA model, producing a mesoscale DNA model (Sun et al., 2019), which showcases the inherent convenience of performing multiscale modeling with bottom-up coarse-graining methods. In this mesoscale model, the dsDNA chain is represented by a chain of spherical beads, each bead representing six base pairs of dsDNA. There are totally three interaction terms in the total potential energy, one bond, one angle and one non-bonded term. With minimal computational resources, simulation of DNA as long as 10 kb was realized. Experimentally observed hexagonal packing in multivalent ion-induced DNA condensates was reproduced in simulations with this mesoscale model. Interestingly, the dynamic process of a single dsDNA chain forming a toroid was described based on simulations within this model.

3.2. The Fan Bonds Model

Savelyev and Papoian (2009a,b) have developed a two-bead per base pair model for dsDNA. This model represents each nucleotide with one bead located at its geometric center. In addition to the bonds that connect nucleotide beads on the same strand, the base-pairing and stacking interactions are collectively modeled by a series of inter-strand bonds, which are called "fan" bonds. These interaction terms are designed to maintain the double-helical structure of dsDNA (**Figure 2B**).

As noted in the theory section above, this Fan Bonds model expands its Hamiltonian over a compact basis set when modeled with the MRG-CG method. The Hamiltonian of this dsDNA model adopts the following form: (Savelyev and Papoian, 2010)

$$
\begin{aligned}
U_{bond,fan} &= \sum_{\alpha=2}^{4} K_\alpha (l - l_0)^\alpha \\
U_{ang} &= \sum_{\alpha=2}^{4} K_\alpha (\theta - \theta_0)^\alpha \\
U_{ion-DNA;ion-ion} &= \sum_{i>j} \left[\frac{A}{r_{ij}^{6;12}} + \sum_{k=1}^{3;5} B^k e^{-C^k (r_{ij} - R_{ij}^k)^2} + \frac{q_i q_j}{4\pi\epsilon_0 \epsilon r_{ij}} \right]
\end{aligned}
\tag{25}
$$

Usage of quartic polynomials for bonds and angles and sums of Gaussians for non-bonded interactions significantly reduces the number of parameters. Indeed, the free energy difference between this CG model and the all-atom model is decreased to $\sim 0.5 k_B T$ in a small number of iterations, which illustrates an efficient algorithm. Finally, the salt concentration-dependent persistence length of dsDNA was reproduced within a scaling factor.

Naômé et al. (2014) have modeled dsDNA with a topology similar to the Fan Bonds model using the IMC approach, while adopting tabulated potentials (Equation 11) to model the system Hamiltonian (**Figure 2C**). Unlike the model by Savelyev and Papoian (2009a,b), where bonded and non-bonded potentials were derived separately, their model follows a more systematic way of deriving all interaction terms from the same calculation, considering cross-correlation among all interaction terms. The inverse problem was solved in two stages, with IBI and IMC respectively. The final model was obtained after a significantly larger number of iterations (∼75 iterations) of than MRG-CG modeling. We clearly see the benefit of having fewer parameters in the MRG-CG method. Various modeling options such as the number of fan bonds and the optimal procedure in solving the inverse problem were explored in their work. The salt concentration-dependent DNA persistence length was well-reproduced with the optimal number of fan bonds, without any scaling factor.

3.3. The "sugar" DNA Model

Kovaleva et al. (2017) have formulated the "sugar" DNA model, designed to include the flexibility of the ribose rings. Each nucleotide is represented by six beads – three for the backbone and three for the base (**Figure 2D**). CG beads are placed on selected atom positions in each nucleotide. The mass of each CG bead is balanced to ensure that the center of mass of the base doesn't move during the mapping operation. The two major conformations of ribose ring, C2′-endo and C3′-endo, are modeled with a double-well potential for P-C1' bond. The transition between these two states corresponds to the transition between C2′-endo and C3′-endo conformations.

Most of the bonded potential functional parameters were derived from atomistic simulations with the AMBER99SB bsc0 force field by Boltzmann inversion, except for the well depth of the aforementioned double-well potential. The depths of two wells of P-C1' potential is set to equal value, as the all-atom sampling is not optimal. In cases where the Boltzmann inversion is not working well, the so-called "relaxation" method is used, where the relevant particle pair is set at a series of distances while relaxing the rest of the system to obtain energy function for this pair. Note that a hybrid approach is adopted in this model to achieve the modeling target. Lastly, using the ion-ion effective potential previously derived by IMC (Lyubartsev and Marčelja, 2002), the "sugar" DNA model successfully modeled the A-DNA and B-DNA states, as well as the transition between the two (**Figure 2D**).

3.4. The 3SPN Model

As an example of a top-down DNA model and hybrid top-down and bottom-up modeling, we discuss the 3-site-per-nucleotide (3SPN) model first developed by Knotts et al. (2007). The first implementation of the 3SPN model (denoted 3SPN.0)

was primarily designed to reproduce the melting temperatures of oligonucleotides. Empirically determined relative interaction strengths among non-bonded interaction terms reduce the parameter set to one single interaction energy ϵ, which is subsequently determined with a trial-and-error approach. In the subsequent development, Sambriski et al. (2009) improved the 3SPN model and derived a new version, 3SPN.1. A new solvent-induced attraction term was introduced in this version of the 3SPN model, and other interaction parameters were further tuned to make the model more precise. As a result, DNA mechanics was improved significantly (**Figure 2E**).

In the subsequent development of the 3SPN model, denoted 3SPN.2 (Hinckley et al., 2013), the authors employed a more detailed interaction Hamiltonian, including a cross-stacking potential to reproduce experimentally determined base interaction energies such as base step energies and base stacking free energies. The resulting model improved the molecular flexibility for both ssDNA and dsDNA. More recently, Freeman et al. (2014) introduced DNA sequence dependence to the 3SPN.2 model. At the same time, additional stability of the helix was implemented using weak dihedral potentials.

Besides developing the 3SPN model by a top-down route, efforts have been made to derive ion-ion, ion-DNA interactions through bottom-up approaches. De Biase et al. (2012) first used a predetermined functional form for ion-ion and ion-DNA interactions, together with 3SPN.1 model. The parameters were derived so that radial distribution functions from all-atom molecular dynamics simulation were satisfactorily reproduced. Subsequently, De Biase et al. (2014) used IMC to derive ion-related potential terms fitted to an empirical functional form. The accuracy of reproducing the RDF was significantly improved in the latter study.

Additionally, the developers of the 3SPN model, Hinckley and de Pablo (2015), developed a transferable coarse-grained ion model for simulations of nucleic acids. Dimethylphosphate (DMP) was adopted as a model molecule for the phosphate group of nucleic acids. Relative entropy minimization was utilized to derive the CG effective potential for ion-ion and ion-phosphate interactions. Ion concentration-dependent persistent length of dsDNA and dsDNA potential of mean force were demonstrated with the new ion model. Although this ionic model only describes ion-ion and ion-phosphate interactions, the authors argued it is a general model, which can be used with other CG models of nucleic acid with explicit charged phosphate sites.

4. DISCUSSION

4.1. Interactions in CG DNA Models

As straightforward as it is theoretically, modeling interactions of DNA in a CG model is not trivial in practice. Since a few types of interactions are involved, it is difficult, if not impossible, to model all aspects of DNA interactions with a reduced number of DOF accurately. We first consider the practice of modeling the DNA conformation ensemble and discuss modeling interaction between DNA and other molecules or ions later.

An ideal model with general applicability would require fairly accurate modeling of four DNA properties, namely electrostatic interaction, sugar puckering, base-pair stacking, and base-pair hydrogen bonding. Long range electrostatic interactions are crucial to DNA chain conformation and mechanical properties. Sugar puckering is essential in the transition between A-DNA and B-DNA, which in turn contributes to DNA thickness and bending flexibility. For nucleic bases, their stacking and hydrogen bonding interactions are anisotropic and directional due to aromaticity. To our knowledge, there is to date no ideal CG DNA model designed to model all these properties simultaneously. All CG DNA models are compromising and focusing on some particular aspect of the interactions in DNA.

Since most CG models are designed with implicit solvent, various screening methods are usually adopted to model electrostatic interaction, sometimes in conjunction with modifications of charge values (Savelyev and Papoian, 2009b). In such cases, the solvent is treated as a uniform medium without structure. The simplest form of electrostatic interaction potential uses a constant relative permittivity with Coulomb's law:

$$U_{el} = \frac{q_i q_j}{4\pi\epsilon_0\epsilon r_{ij}} \tag{26}$$

where ϵ_0 is the vacuum permittivity, ϵ is the relative permittivity. It is used in the IMC dsDNA model (Korolev et al., 2014; Sun et al., 2019; Minhas et al., 2020), the MRG-CG model (Savelyev and Papoian, 2009a, 2010), the "sugar" model for its ion-ion and ion-DNA interactions (Kovaleva et al., 2017), and a few other models. Some models use the Debye-Huckle potential, generally in the form (Savelyev and Papoian, 2009b; Morriss-Andrews et al., 2010; He et al., 2013; Hinckley et al., 2013; Kovaleva et al., 2017)

$$U_{el} = \frac{q_i q_j}{4\pi\epsilon_0\epsilon r_{ij}} \cdot e^{-r_{ij}/\lambda_D} \tag{27}$$

where λ_D is the Debye length of interacting particles. Effectively, interacting particles experience larger permittivity at a longer distance. More complicated forms of distance-dependent permittivity, such as employing a switching function (Kovaleva et al., 2017), are used in some models. Nevertheless, a suitable screened interaction can be engineered to model the electrostatic interaction in CG models where solvent degrees of freedom are missing. However, experience and tweaking might be necessary to obtain an appropriate choice that reproduces a given experimental data-set.

We note here that from the structure-based coarse-graining point of view, electrostatic interaction is part of the force that determines particle correlations alongside short-range interaction, which includes van der Waals attraction, short-range repulsion, and implicit effect of water solvation. With a systematic modeling approach such as IMC, these factors are included in the effective pair interactions as the model's target is to reproduce particle correlation functions. If there was an inaccuracy in the electrostatic interaction, the electrostatic potential's error could be absorbed into the short-range interaction potential and vice versa, such that the final total potential reproduces the target particle correlation functions.

Though sugar puckering is essential to DNA backbone conformation, its modeling is not common in bottom-up

CG DNA models, partially due to the inaccuracy in force field parameters and difficulties in the sampling of the C2′-endo and C3′-endo conformations. Additionally, modeling these interactions requires higher resolution, including more CG beads (and DOFs) compared to what is present in most available CG models. The "sugar" DNA model (Kovaleva et al., 2017) explicitly models sugar puckering with a double-well bond potential and a specifically designed bonding structure along the DNA backbone. Additional empirical modifications are needed upon bottom-up modeling to achieve optimal results, manifesting the difficulties mentioned before. Many CG DNA models consider B-DNA only; hence they have not included sugar puckering in their parameter set.

Due to the planar conformation of nucleic bases, interactions originating from them are anisotropic. To realistically represent this anisotropy, there should be a sufficient number of DOF in the CG description. The simplest way is to represent individual nucleic bases with multiple coarse-grained beads. For example, in the "sugar" DNA model, each base is modeled by three CG beads with balanced mass distribution (Kovaleva et al., 2017) such that the base plane is easily defined. In this way, base stacking and hydrogen bonding potentials can be projected onto a few suitable degrees of freedom. Another popular choice is to use anisotropic potentials at interaction sites. The number of interaction sites is minimal, though the number of degrees of freedom is not necessarily small. Gay-Berne potential (Gay and Berne, 1981; Persson, 2012), a generalized Lennard-Jones potential with anisotropy, is frequently adopted to model ellipsoidal CG beads. For instance, ellipsoidal beads are used in the NARES DNA model (He et al., 2013; Liwo et al., 2014; Yin et al., 2015) and used by Li et al. (2016). One should note that using anisotropic beads does not necessarily result in a better model, as seen in Li et al. (2016). No matter through a bottom-up or top-down approach, determining parameters in these models is not a trivial task. Furthermore, anisotropic potentials require more computational time to compute the forces compared to the models based on isotropic distance-dependent potentials. We also note that anisotropic beads may not be necessary for modeling DNA properties intrinsically related to base interactions. For example, the 3SPN model (Knotts et al., 2007) reproduces the dsDNA salt-dependent melting temperature with isotropic beads.

In a simplified representation, DNA base pair interactions can be approximated by bonded fluctuations, as done in the IMC dsDNA model (Korolev et al., 2014; Sun et al., 2019; Minhas et al., 2020) and the G-quadruplex model by Rebič et al. (2015). Though the details of basepair conformation and dynamics are lost, these representations are incredibly efficient in simulation and are easily extended to large molecules. In cases where base pair dynamics is considered secondary, bonded representation could be superior to more detailed base pair models.

4.2. Mechanical Properties of DNA

DNA mechanical properties are essential for understanding the DNA behavior in chromatin of the cell nucleus and DNA nanomaterial development. These properties are determined by a combination of DNA intramolecular interactions (backbone rigidity, basepair interactions) and the electrostatic polyelectrolyte nature of DNA. To test the performance of any dsDNA model, the bending flexibility characterized by the bending persistence length is usually the first property tested against known experimental data. In simulations, the bending persistence length, L_p, is approximated by an exponential decay of the angular correlation function:

$$\langle \hat{\mathbf{e}}_i \cdot \hat{\mathbf{e}}_{i+n} \rangle = \exp\left(\frac{-nI}{L_p}\right) \tag{28}$$

where $\hat{\mathbf{e}}_i$ is a unit vector along segment i, I is the average segment contour length, angle brackets denote ensemble average. Since the importance of the electrostatic interaction for DNA bending flexibility, testing salt concentration-dependent persistence length is a rigorous way to test the performance of a new model. Though, it should be noted that the approximation of an exponential correlation function as above, albeit good for a worm-like chain polymer, is ignoring the DNA sequence effect and intrinsic DNA curvature (Mitchell et al., 2017).

With the IMC dsDNA model, Korolev et al. (2014) tested the salt concentration-dependent persistence length, where the bonded potential is fitted to harmonic function based on IMC inverted potential. The result showed a very good agreement with experimental data. When the bonded potentials were substituted with accurate tabulated IMC inverted potential (Minhas et al., 2020), the bending persistence length still agrees well with experiments. However, when torsion persistence length was tested with this model, the result showed a significantly larger value than experimentally reported data (Korolev et al., 2014). It was explained by reasoning that while ion-dependent DNA bending is determined mostly by long-range electrostatic forces, the torsion flexibility does not depend much on electrostatics. It is determined mostly by basepair twisting, which is relatively short range. Since the IMC dsDNA model simplifies basepair movement to bonded fluctuations, it is challenging to reproduce DNA twisting satisfactorily.

In the two-bead per basepair models by Savelyev and Papoian (2010) and Naômé et al. (2014), the long-range interactions are implemented with explicit ions similarly to the works by Korolev et al. (2014) and Minhas et al. (2020) while the bonding of CG sites along DNA was different. These models show that persistence length is sensitive to the bonding structure. With an optimal bonding structure, the salt-dependent persistence length can be reproduced well.

Figure 3 compares the result of predictions of the dependence of persistence length on salt for these above mentioned bottom-up CG DNA models. Generally, all bottom-up models provide near quantitative agreement with experimental measurements of persistence length over a wide range of salt concentration. The agreement is particularly good at physiological salt. The predictions at salt concentrations below 1 mM display more variation between models and compared to experiments. The results from the IMC model (Minhas et al., 2020) do a good job over a wide range of salt concentrations and are similar to the data obtained by Naômé et al. (2014) and to those of Hinckley and de Pablo (2015). In the latter work, bottom-up derived CG potentials for the ionic interactions were used, while

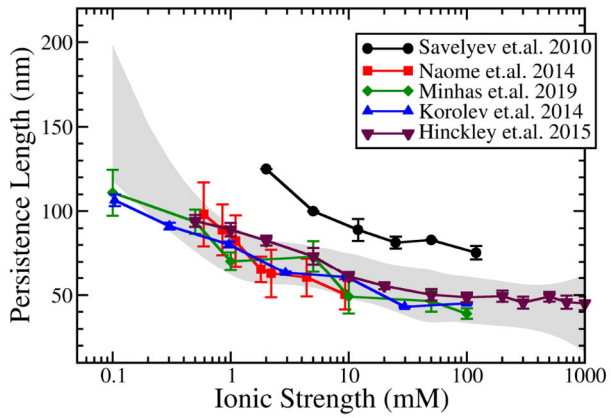


FIGURE 3 | Ionic strength dependent persistence length of dsDNA modeled by coarse-grained DNA models using explicit ions in the simulations. Data from five models are plotted: Savelyev and Papoian (2010), Korolev et al. (2014), Naômé et al. (2014), Hinckley and de Pablo (2015), and Minhas et al. (2020). The gray band corresponds to the spline approximation of the experimental data (Hagerman, 1981; Kam et al., 1981; Manning, 1981; Rizzo and Schellman, 1981; Cairney and Harrington, 1982; Porschke, 1991; Baumann et al., 1997) with a confidence interval of 0.99995.

internal DNA interactions were obtained from the empirically parameterized 3SPN model. The persistence length calculations by Savelyev and Papoian (2010), which displayed consistently larger persistence length values, resulted, however,in good agreement with experimental data following uniform rescaling of all CG DNA structural parameters.

If a more detailed CG DNA model is used, it is more challenging to attribute long-range correlation to specific interactions since these are determined collectively by a group of degrees of freedom. For instance, in Morriss-Andrews et al. (2010), the modeled bending persistence length of single-stranded DNA is lower than experiments.

Another interesting mechanical property is force-extension curve of ssDNA. Maffeo et al. (2014) derived a two-bead per nucleotide model with a hybrid approach, i.e., using IBI to derive a primer model, then refined it by fitting to the radius of gyration of ssDNA. The force-extension curve obtained by this model fits better than the other two top-down models to the experimental curve.

4.3. DNA Aggregation and Compaction Properties

As a highly-charged polyelectrolyte, DNA has been extensively studied in solution both as a standalone subject and as a component of complexes formed with other molecules. Efforts have been made to model such molecular systems and to get insights from a physical perspective. Besides the bending flexibility discussed in section 4.2, quantities such as radius of gyration (Maffeo et al., 2014), melting temperature (Hinckley et al., 2013), and even knotting probabilities (Rieger and Virnau, 2018) are adopted as modeling targets, especially in top-down models. There are additional studies of DNA-protein complexes with top-down models, such as the NARES model for DNA-protein complex (Yin et al., 2015) and free energy associated with nucleosome unwrapping (Lequieu et al., 2016). Although DNA solution structure can be directly simulated once a bottom-up model is acquired, the result may not always agree with experimental data. For instance, in the ssDNA model designed by Maffeo et al. (2014), the direct result of the CG model derived by IBI significantly underestimates the radius of gyration of ssDNA. Correction to the interaction potential was made afterward to obtain agreement with experimental data.

Another application of CG DNA models is to study DNA condensation and phase separation. Positively charged multivalent cations and polyelectrolytes can induce DNA condensation under physiological salt conditions, which is important for understanding DNA compaction in the cell nucleus. Accurate effective potentials between DNA and multivalent ions are crucial to these studies. Córdoba et al. (2017) studied the dependence of DNA packing inside nanometer-sized viral capsids on multivalent cations using the 3SPN.2C CG DNA model incorporating bottom-up effective potentials for ion-phosphate interactions (**Figure 4**). Multivalent cations such as spermidine and magnesium induce attraction between packaged DNA leading to DNA condensation. At high concentrations of spermidine, the condensation reduced the pressure inside the virus capsid. Savelyev and Papoian (2010) used their CG bottom-up "fan" model to predict the structural phase transitions in torsionally stressed DNA nanocircles due to the presence of salt (see **Figure 4B**). The model predicted phase transition to a buckled state in the overtwisted DNA nanocircle under physiological salt conditions.

In work by Sun et al. (2019), a fully bottom-up model of dsDNA in the presence of multivalent ions was built based on atomistic simulations with the CHARMM27 force field. Subsequently the model was used to study DNA condensation in the presence of Cobalt(III)-hexammine ($CoHex^{3+}$) ion. The model successfully reproduced the experimentally observed DNA condensation into hexagonally ordered structures (**Figure 4C**). Furthermore, the developed CG model was used to make one more step in coarse-graining, to obtain a CG DNA model suitable for mesoscale simulation. It resulted in a "super-coarse-grained" DNA model with a simple bead-on-string topology and a single IMC-derived long-range potential between the beads, effectively accounting for the effect of water and ions. The super-CG model of DNA was used to study the formation of toroidal structures formed by long (40k base pairs) DNA in the presence of $CoHex^{3+}$ ions, resulting in excellent agreement with electron microscopy observations of DNA toroids formation. It is worth emphasizing that this model, allowing simulation of DNA on a micrometer length scale, was derived exclusively from atomistic simulations without using any empirical parameters.

5. CONCLUDING REMARKS

To model large biomolecular systems such as the components of organelles of a living cell (e.g., chromatin in the cell nucleus), an atomistic approach based on all-atom MD simulations is

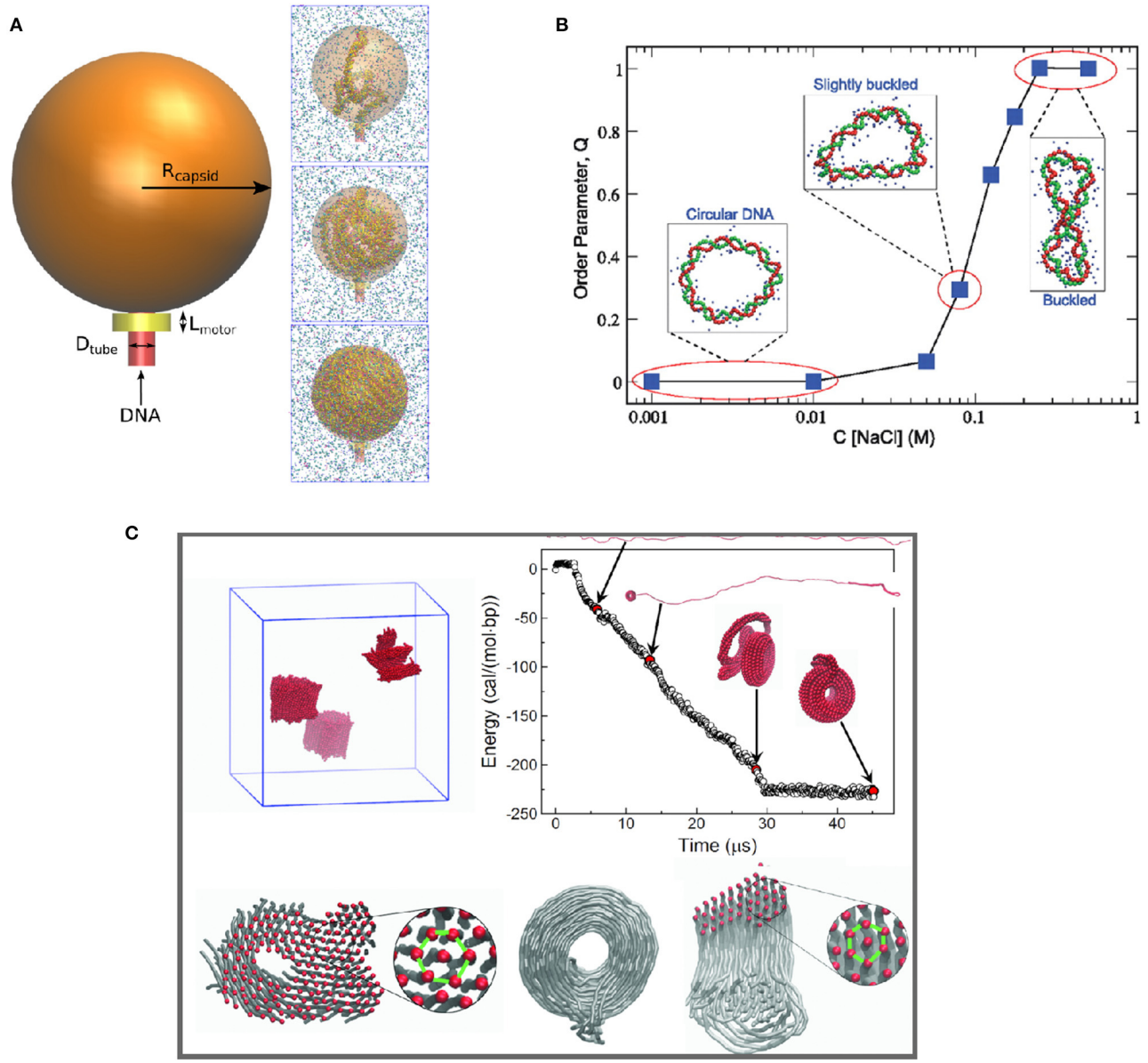


FIGURE 4 | Application of different coarse-grained DNA models illustrating multi-scale phenomena, including **(A)** DNA compaction in virus capsid (Córdoba et al., 2017); **(B)** phase behavior of circular DNA (Savelyev and Papoian, 2010); **(C)** multivalent ion-induced DNA condensation (Sun et al., 2019).

neither computationally feasible nor practically useful since the vast number of DOF makes the analysis complicated. Within the coarse-grained approach, the macromolecular systems can be reduced to a description with effective sites, representing many atoms, which reduces the computational demand and simplifies the analysis by focusing on those DOFs that are of interest for a specific problem at hand. To obtain the effective potentials describing the interactions between the CG sites, we generally have two possibilities, either a top-down approach that fits the potentials to some available experimental data or the bottom-up approach discussed in this review. We have presented an overview of the approach of bottom-up CG computer simulations of DNA, which enables the modeling of multiscale DNA structure, dynamics, and interactions with various bio-macromolecules. As discussed thoroughly here, the available methods are in principle equivalent. They enable rigorous extraction of the effective potentials that reproduces the system's behavior in an FG description based on a given AA force field. These potentials are particularly advantageous when dealing with highly charged systems such as DNA. The IMC method e.g., enables rigorous modeling of DNA mechanical properties and aggregation of very large DNA

assemblies, with a reduction of the number of particles of several orders of magnitude, but still using effective potentials that implicitly incorporate the effects of the solvent water. Further development of this approach to model packaging of DNA in chromatin and even chromosomes is a highly challenging problem, which though in principle, can be handled within the same methodology.

We foresee that as the methodology and computational capability improve, physics-based models, derived by bottom-up modeling or from other physical principles, will be taking up more important roles toward illustrating biophysical processes relevant to experimental studies. It is even more evident when one considers the development in experimental techniques, such as Cryo-EM and single-molecule experiments. These experimental methods are reaching into finer and finer scales to probe the underlying physics. The gap between physical simulations and wet-lab experiments is diminishing. With the inherent multiscale nature of bottom-up coarse-graining, multiple molecular models spanning a few magnitude of length scale can be generated rigorously. More accurate interactions will be modeled to conduct simulations. With a minimal number of empirical parameters, simulations can provide better insights into molecular characteristics at each scale. Though there is still much development to be expected in bottom-up coarse-graining methods, we believe the connections between microscopic molecular characteristics and experimental observations will improve, leading to a deeper understanding of DNA physical properties at large spatial and temporal scales.

AUTHOR CONTRIBUTIONS

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

FUNDING

AL was supported by the Swedish Research Council (Vetenskapsrådet), grant 2017-03950, and computational resources provided by the Swedish National Infrastructure for Computing (SNIC) through National Supercomputer Center (NSC). This work was supported by a Ministry of Education (MOE), Singapore, Academic Research Fund Tier 3 Grant (MOE2019-T3-1-012) to LN.

ACKNOWLEDGMENTS

We acknowledge the generous support of computer time allocation from the National Supercomputing Centre (NSCC) Singapore.

REFERENCES

Baumann, C. G., Smith, S. B., Bloomfield, V. A., and Bustamante, C. (1997). Ionic effects on the elasticity of single dna molecules. *Proc. Natl. Acad. Sci. U.S.A.* 94, 6185–6190. doi: 10.1073/pnas.94.12.6185

Bloomfield, V., Crothers, D., Tinoco, I., Killman, P., Hearst, J., Wemmer, D., et al. (2000). *Nucleic Acids: Structure, Properties, and Functions*. Sausalito, CA: University Science Books.

Cairney, K. L., and Harrington, R. E. (1982). Flow birefringence of t7 phage DNA: dependence on salt concentration. *Biopolymers* 21, 923–934. doi: 10.1002/bip.360210506

Chaimovich, A., and Shell, M. S. (2011). Coarse-graining errors and numerical optimization using a relative entropy framework. *J. Chem. Phys*. 134:094112. doi: 10.1063/1.3557038

Córdoba, A., Hinckley, D. M., Lequieu, J., and Pablo, J. J. (2017). A molecular view of the dynamics of dsDNA packing inside viral capsids in the presence of ions. *Biophys. J*. 112, 1302–1315. doi: 10.1016/j.bpj.2017.02.015

De Biase, P. M., Markosyan, S., and Noskov, S. (2014). Microsecond simulations of dna and ion transport in nanopores with novel ion-ion and ion-nucleotides effective potentials. *J. Comput. Chem*. 35, 711–721. doi: 10.1002/jcc.23544

De Biase, P. M., Solano, C. J. F., Markosyan, S., Czapla, L., and Noskov, S. Y. (2012). BROMOC-D: Brownian dynamics/Monte-Carlo program suite to study ion and dna permeation in nanopores. *J. Chem. Theory Comput*. 8, 2540–2551. doi: 10.1021/ct3004244

Ercolessi, F., and Adams, J. B. (1994). Interatomic potentials from first-principles calculations: The force-matching method. *Europhys. Lett*. 26, 583–588. doi: 10.1209/0295-5075/26/8/005

Eriksson, A., Jacobi, M. N., Nyström, J., and Tunstrøm, K. (2008). Using force covariance to derive effective stochastic interactions in dissipative particle dynamics. *Phys. Rev. E* 77:016707. doi: 10.1103/PhysRevE.77.016707

Fan, Y., Korolev, N., Lyubartsev, A. P., and Nordenskiold, L. (2013). An advanced coarse-grained nucleosome core particle model for computer simulations of nucleosome-nucleosome interactions under varying ionic conditions. *PLoS ONE* 8:e54228. doi: 10.1371/journal.pone.0054228

Freeman, G. S., Hinckley, D. M., Lequieu, J. P., Whitmer, J. K., and de Pablo, J. J. (2014). Coarse-grained modeling of dna curvature. *J. Chem. Phys*. 141:165103. doi: 10.1063/1.4897649

Galindo-Murillo, R., Robertson, J. C., Zgarbova, M., Sponer, J., Otyepka, M., Jurecka, P., et al. (2016). Assessing the current state of amber force field modifications for DNA. *J. Chem. Theory Comput*. 12, 4114–4127. doi: 10.1021/acs.jctc.6b00186

Gay, J. G., and Berne, B. J. (1981). Modification of the overlap potential to mimic a linear site-site potential. *J. Chem. Phys*. 74, 3316–3319. doi: 10.1063/1.441483

Giulini, M., Menichetti, R., Shell, M. S., and Potestio, R. (2020). An information-theory-based approach for optimal model reduction of biomolecules. *J. Chem. Theory Comput*. 16, 6795–6813. doi: 10.1021/acs.jctc.0c00676

Gkeka, P., Stoltz, G., Barati Farimani, A., Belkacemi, Z., Ceriotti, M., Chodera, J. D., et al. (2020). Machine learning force fields and coarse-grained variables in molecular dynamics: Application to materials and biological systems. *J. Chem. Theory Comput*. 16, 4757–4775. doi: 10.1021/acs.jctc.0c00355

Guldbrand, L., Nilsson, L. G., and Nordenskiold, L. (1986). A Monte Carlo simulation study of electrostatic forces between hexagonally packed DNA double helices. *J. Chem. Phys*. 85, 6686–6698. doi: 10.1063/1.451450

Hagerman, P. J. (1981). Investigation of the flexibility of DNA using transient electric birefringence. *Biopolymers* 20, 1503–1535. doi: 10.1002/bip.1981.360200710

Hart, K., Foloppe, N., Baker, C. M., Denning, E. J., Nilsson, L., and MacKerell, A. D. (2012). Optimization of the charmm additive force field for DNA: improved treatment of the bi/bii conformational equilibrium. *J. Chem. Theory Comput*. 8, 348–362. doi: 10.1021/ct200723y

He, Y., Maciejczyk, M., Ołdziej, S., Scheraga, H. A., and Liwo, A. (2013). Mean-field interactions between nucleic-acid-base dipoles can drive the formation of a double helix. *Phys. Rev. Lett*. 110:098101. doi: 10.1103/PhysRevLett.110.098101

Henderson, R. (1974). A uniqueness theorem for fluid pair correlation functions. *Phys. Lett. A* 49, 197–198. doi: 10.1016/0375-9601(74)90847-0

Hess, B., Holm, C., and van der Vegt, N. (2006). Osmotic coefficients of atomistic nacl (aq) force fields. *J. Chem. Phys*. 124:164509. doi: 10.1063/1.2185105

Hijon, C., Espanol, P., Vanden-Eijnden, E., and Delgado-Buscalioni, R. (2010). Mori-Zwanzig formalism as a practical computational tool. *Faraday Discuss*. 144:301–322. doi: 10.1039/B902479B

Hinckley, D. M., and de Pablo, J. J. (2015). Coarse-grained ions for nucleic acid modeling. *J. Chem. Theory Comput*. 11, 5436–5446. doi: 10.1021/acs.jctc.5b00341

Hinckley, D. M., Freeman, G. S., Whitmer, J. K., and de Pablo, J. J. (2013). An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: structure, thermodynamics, and dynamics of hybridization. *J. Chem. Phys*. 139:144903. doi: 10.1063/1.4822042

Izvekov, S., Parrinello, M., Burnham, C. J., and Voth, G. A. (2004). Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: a new method for force-matching. *J. Chem. Phys*. 120, 10896–10913. doi: 10.1063/1.1739396

Izvekov, S., and Voth, G. A. (2005). A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B* 109, 2469–2473. doi: 10.1021/jp044629q

Kam, Z., Borochov, N., and Eisenberg, H. (1981). Dependence of laser light scattering of dna on nacl concentration. *Biopolymers* 20, 2671–2690. doi: 10.1002/bip.1981.360201213

Knotts, T. A., Rathore, N., Schwartz, D. C., and de Pablo, J. J. (2007). A coarse grain model for dna. *J. Chem. Phys*. 126:084901. doi: 10.1063/1.2431804

Kono, H., and Ishida, H. (2020). Nucleosome unwrapping and unstacking. *Curr. Opin. Struct. Biol*. 64, 119–125. doi: 10.1016/j.sbi.2020.06.020

Korolev, N., Luo, D., Lyubartsev, A. P., and Nordenskiold, L. (2014). A coarse-grained dna model parameterized from atomistic simulations by inverse monte carlo. *Polymers* 6, -1675. doi: 10.3390/polym6061655

Korolev, N., Lyubartsev, A. P., and Nordenskiold, L. (2010). Cation-induced polyelectrolyte-polyelectrolyte attraction in solutions of dna and nucleosome core particles. *Adv. Colloid Interface Sci*. 158, 32–47. doi: 10.1016/j.cis.2009.08.002

Korolev, N., Nordenskiold, L., and Lyubartsev, A. P. (2016). Multiscale coarse-grained modelling of chromatin components: DNA and the nucleosome. *Adv. Colloid Interface Sci*. 232, 36–48. doi: 10.1016/j.cis.2016.02.002

Kovaleva, N. A., Koroleva, I. P., Mazo, M. A., and Zubova, E. A. (2017). The "sugar" coarse-grained dna model. *J. Mol. Model*. 23, 1–16. doi: 10.1007/s00894-017-3209-z

Lavery, R., Zakrzewska, K., Beveridge, D., Bishop, T. C., Case, D. A., Cheatham, T.homas, I., et al. (2009). A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res*. 38, 299–313. doi: 10.1093/nar/gkp834

Leonarski, F., Trovato, F., Tozzini, V., Les, A., and Trylska, J. (2013). Evolutionary algorithm in the optimization of a coarse-grained force field. *J. Chem. Theory Comput*. 9, 4874–4889. doi: 10.1021/ct4005036

Lequieu, J., Cordoba, A., Schwartz, D. C., and de Pablo, J. J. (2016). Tension-dependent free energies of nucleosome unwrapping. *ACS Central Sci*. 2, 660–666. doi: 10.1021/acscentsci.6b00201

Li, G., Shen, H., Zhang, D., Li, Y., and Wang, H. (2016). Coarse-grained modeling of nucleic acids using anisotropic gay-berne and electric multipole potentials. *J. Chem. Theory Comput*. 12, 676–693. doi: 10.1021/acs.jctc.5b00903

Liwo, A., Baranowski, M., Czaplewski, C., Golas, E., He, Y., Jagiela, D., et al. (2014). A unified coarse-grained model of biological macromolecules based on mean-field multipole-multipole interactions. *J. Mol. Model*. 20:2306. doi: 10.1007/s00894-014-2306-5

Lu, L., Dama, J. F., and Voth, G. A. (2013). Fitting coarse-grained distribution functions through an iterative force-matching method. *J. Chem. Phys*. 139:121906. doi: 10.1063/1.4811667

Lu, L., Izvekov, S., Das, A., Andersen, H. C., and Voth, G. A. (2010). Efficient, regularized, and scalable algorithms for multiscale coarse-graining. *J. Chem. Theory Comput*. 6, 954–965. doi: 10.1021/ct900643r

Lyubartsev, A., Mirzoev, A., Chen, L., and Laaksonen, A. (2010). Systematic coarse-graining of molecular models by the newton inversion method. *Faraday Discuss*. 144, 43–56. doi: 10.1039/B901511F

Lyubartsev, A. P. (2018). "Inverse Monte Carlo methods," in *Coarse-Grained Modeling of Biomolecules*, ed G. A. Papoian (Boca Raton, FL: CRC Press; Taylor Francis Group), 1–26. doi: 10.1201/9781315374284-1

Lyubartsev, A. P., and Laaksonen, A. (1995). Calculation of effective interaction potentials from radial distribution functions: a reverse Monte Carlo approach. *Phys. Rev. E* 52, 3730–3737. doi: 10.1103/PhysRevE.52.3730

Lyubartsev, A. P., and Marčelja, S. (2002). Evaluation of effective ion-ion potentials in aqueous electrolytes. *Phys. Rev. E* 65:041202. doi: 10.1103/PhysRevE.65.041202

Lyubartsev, A. P., Naômé, A., Vercauteren, D. P., and Laaksonen, A. (2015). Systematic hierarchical coarse-graining with the inverse Monte Carlo method. *J. Chem. Phys*. 143:243120. doi: 10.1063/1.4934095

Maffeo, C., Ngo, T. T. M., Ha, T., and Aksimentiev, A. (2014). A coarse-grained model of unstructured single-stranded dna derived from atomistic simulation and single-molecule experiment. *J. Chem. Theory Comput*. 10, 2891–2896. doi: 10.1021/ct500193u

Manning, G. S. (1981). A procedure for extracting persistence lengths from light-scattering data on intermediate molecular weight DNA. *Biopolymers* 20, 1751–1755. doi: 10.1002/bip.1981.360200815

Minhas, V., Sun, T., Mirzoev, A., Korolev, N., Lyubartsev, A. P., and Nordenskiold, L. (2020). Modeling DNA flexibility: comparison of force fields from atomistic to multiscale levels. *J. Phys. Chem. B* 124, 38–49. doi: 10.1021/acs.jpcb.9b09106

Mirzoev, A., Nordenskiold, L., and Lyubartsev, A. (2019). Magic v.3: an integrated software package for systematic structure-based coarse-graining. *Comput. Phys. Commun*. 237, 263–273. doi: 10.1016/j.cpc.2018.11.018

Mitchell, J. S., Glowacki, J., Grandchamp, A. E., Manning, R. S., and Maddocks, J. H. (2017). Sequence-dependent persistence lengths of DNA. *J. Chem. Theory Comput*. 13, 1539–1555. doi: 10.1021/acs.jctc.6b00904

Morriss-Andrews, A., Rottler, J., and Plotkin, S. S. (2010). A systematically coarse-grained model for DNA and its predictions for persistence length, stacking, twist, and chirality. *J. Chem. Phys*. 132:035105. doi: 10.1063/1.3269994

Naômé, A., Laaksonen, A., and Vercauteren, D. P. (2014). A solvent-mediated coarse-grained model of dna derived with the systematic newton inversion method. *J. Chem. Theory Comput*. 10, 3541–3549. doi: 10.1021/ct500222s

Noid, W. G., Chu, J.-W., Ayton, G. S., Krishna, V., Izvekov, S., Voth, G. A., et al. (2008). The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys*. 128:244114. doi: 10.1063/1.2938860

Noid, W. G., Chu, J.-W., Ayton, G. S., and Voth, G. A. (2007). Multiscale coarse-graining and structural correlations: connections to liquid-state theory. *J. Phys. Chem. B*, 111, 4116–4127. doi: 10.1021/jp068549t

Nordenskiold, L., Korolev, N., and Lyubartsev, A. P. (2008). *DNA-DNA Interactions*. Hoboken, NJ: John Wiley Sons, Ltd. doi: 10.1002/9780470286364.ch8

Persson, R. A. X. (2012). Note: modification of the gay-berne potential for improved accuracy and speed. *J. Chem. Phys*. 136:226101. doi: 10.1063/1.4729745

Porschke, D. (1991). Persistence length and bending dynamics of DNA from electrooptical measurements at high salt concentrations. *Biophys. Chem*. 40, 169–179. doi: 10.1016/0301-4622(91)87006-Q

Postow, L., Crisona, N. J., Peter, B. J., Hardy, C. D., and Cozzarelli, N. R. (2001). Topological challenges to dna replication: conformations at the fork. *Proc. Natl. Acad. Sci. U.S.A*. 98, 8219–8226. doi: 10.1073/pnas.111006998

Rajagopalan, R., and Yakhmi, J. V. (2017). "Chapter 8: Nanotechnological approaches toward cancer chemotherapy," in *Nanostructures for Cancer Therapy, Micro and Nano Technologies*, eds A. Ficai and A. M. Grumezescu (Amsterdam: Elsevier), 211–240. doi: 10.1016/B978-0-323-46144-3.00008-8

Rebič, M., Mocci, F., Laaksonen, A., and Uličný, J. (2015). Multiscale simulations of human telomeric g-quadruplex DNA. *J. Phys. Chem. B* 119, 105–113. doi: 10.1021/jp5103274

Reith, D., Putz, M., and Muller-Plathe, F. (2003). Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem*. 24, 1624–1636. doi: 10.1002/jcc.10307

Rieger, F. C., and Virnau, P. (2018). Coarse-grained models of double-stranded DNA based on experimentally determined knotting probabilities. *React. Funct. Polymers* 131, 243–250. doi: 10.1016/j.reactfunctpolym.2018.08.002

Rizzo, V., and Schellman, J. (1981). Flow dichroism of t7 DNA as a function of salt concentration. *Biopolymers* 20, 2143–2163. doi: 10.1002/bip.1981.360201009

Romiszowski, P., and Yaris, R. (1991). A dynamic simulation method suppressing uninteresting degrees of freedom. *J. Chem. Phys*. 94, 6751–6761. doi: 10.1063/1.460726

Rudzinski, J. F., and Noid, W. G. (2011). Coarse-graining entropy, forces, and structures. *J. Chem. Phys*. 135:214101. doi: 10.1063/1.3663709

Sambriski, E., Schwartz, D., and de Pablo, J. (2009). A mesoscale model of dna and its renaturation. *Biophys. J.* 96, 1675–1690. doi: 10.1016/j.bpj.2008.09.061

Savelyev, A., and Papoian, G. A. (2009a). Molecular renormalization group coarse-graining of electrolyte solutions: application to aqueous NaCl and KCL. *J. Phys. Chem. B* 113, 7785–7793. doi: 10.1021/jp9005058

Savelyev, A., and Papoian, G. A. (2009b). Molecular renormalization group coarse-graining of polymer chains: application to double-stranded DNA. *Biophys. J.* 96, 4044–4052. doi: 10.1016/j.bpj.2009.02.067

Savelyev, A., and Papoian, G. A. (2010). Chemically accurate coarse graining of double-stranded dna. *Proc. Natl. Acad. Sci. U.S.A.* 107, 20340–20345. doi: 10.1073/pnas.1001163107

Schiessel, H. (2003). The physics of chromatin. *J. Phys. Condens. Matter* 15, R699–R774. doi: 10.1088/0953-8984/15/19/203

Shaytan, A. K., Armeev, G. A., Goncearenco, A., Zhurkin, V. B., Landsman, D., and Panchenko, A. R. (2016). Coupling between histone conformations and dna geometry in nucleosomes on a microsecond timescale: atomistic insights into nucleosome functions. *J. Mol. Biol.* 428, 221–237. doi: 10.1016/j.jmb.2015.12.004

Shell, M. S. (2008). The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* 129:144108. doi: 10.1063/1.2992060

Shell, M. S. (2016). *Coarse-Graining With The Relative Entropy*. Hoboken, NJ: John Wiley Sons, Ltd. doi: 10.1002/9781119290971.ch5

Soper, A. (1996). Empirical potential monte carlo simulation of fluid structure. *Chem. Phys.* 202, 295–306. doi: 10.1016/0301-0104(95)00357-6

Stelzl, L. S., Erlenbach, N., Heinz, M., Prisner, T. F., and Hummer, G. (2017). Resolving the conformational dynamics of DNA with Ångstrom resolution by pulsed electron-electron double resonance and molecular dynamics. *J. Am. Chem. Soc.* 139, 11674–11677. doi: 10.1021/jacs.7b05363

Sun, T., Mirzoev, A., Minhas, V., Korolev, N., Lyubartsev, A. P., and Nordenskiold, L. (2019). A multiscale analysis of DNA phase separation: from atomistic to mesoscale level. *Nucleic Acids Res.* 47, 5550–5562. doi: 10.1093/nar/gkz377

Uusitalo, J. J., Ingolfsson, H. I., Akhshi, P., Tieleman, D. P., and Marrink, S. J. (2015). Martini coarse-grained force field: extension to DNA. *J. Chem. Theory Comput.* 11, 3932–3945. doi: 10.1021/acs.jctc.5b00286

Wang, J., Olsson, S., Wehmeyer, C., Perez, A., Charron, N. E., de Fabritiis, G., et al. (2019). Machine learning of coarse-grained molecular dynamics force fields. *ACS Central Sci.* 5, 755–767. doi: 10.1021/acscentsci.8b00913

Wang, L.-P., Martinez, T. J., and Pande, V. S. (2014). Building force fields: an automatic, systematic, and reproducible approach. *J. Phys. Chem. Lett.* 5, 1885–1891. doi: 10.1021/jz500737m

Wang, W., and Gomez-Bombarelli, R. (2019). Coarse-graining auto-encoders for molecular dynamics. *NPJ Comput. Mater.* 5:125. doi: 10.1038/s41524-019-0261-5

Yin, Y., Sieradzan, A. K., Liwo, A., He, Y., and Scheraga, H. A. (2015). Physics-based potentials for coarse-grained modeling of protein-dna interactions. *J. Chem. Theory Comput.* 11, 1792–1808. doi: 10.1021/ct5009558

Zhang, L., Han, J., Wang, H., Car, R., and Weinan, E. (2018). DeePCG: constructing coarse-grained models via deep neural networks. *J. Chem. Phys.* 149:034101. doi: 10.1063/1.5027645

Zhang, Z., Lu, L., Noid, W. G., Krishna, V., Pfaendtner, J., and Voth, G. A. (2008). A systematic methodology for defining coarse-grained sites in large biomolecules. *Biophys. J.* 95, 5073–5083. doi: 10.1529/biophysj.108.139626

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 © 2021 Sun, Minhas, Korolev, Mirzoev, Lyubartsev and Nordenskiöld. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 19 March 2021
doi: 10.3389/fmolb.2021.631854

Multi-Scale Flexible Fitting of Proteins to Cryo-EM Density Maps at Medium Resolution

Marta Kulik[1†‡], *Takaharu Mori*[1] *and Yuji Sugita*[1,2,3*†]

[1]*Theoretical Molecular Science Laboratory, RIKEN Cluster for Pioneering Research, Wako-shi, Japan,* [2]*RIKEN Center for Computational Science, Kobe, Japan,* [3]*RIKEN Center for Biosystems Dynamics Research, Kobe, Japan*

OPEN ACCESS

Edited by:
Peter J Bond,
*Bioinformatics Institute (A*STAR), Singapore*

Reviewed by:
Yong Wang,
University of Copenhagen, Denmark
Simon Olsson,
Chalmers University of Technology, Sweden

***Correspondence:**
Yuji Sugita
sugita@riken.jp

‡Present address:
Biological and Chemical Research Center, Department of Chemistry, University of Warsaw, ul. Żwirki i Wigury 101, 02-089 Warsaw, Poland

†ORCID:
Marta Kulik
0000-0003-2381-7665
Yuji Sugita
0000-0001-9738-9216
Takaharu Mori
0000-0002-8717-2926

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *21 November 2020*
Accepted: *26 January 2021*
Published: *19 March 2021*

Citation:
Kulik M, Mori T and Sugita Y (2021) Multi-Scale Flexible Fitting of Proteins to Cryo-EM Density Maps at Medium Resolution. Front. Mol. Biosci. 8:631854. doi: 10.3389/fmolb.2021.631854

Structure determination using cryo-electron microscopy (cryo-EM) medium-resolution density maps is often facilitated by flexible fitting. Avoiding overfitting, adjusting force constants driving the structure to the density map, and emulating complex conformational transitions are major concerns in the fitting. To address them, we develop a new method based on a three-step multi-scale protocol. First, flexible fitting molecular dynamics (MD) simulations with coarse-grained structure-based force field and replica-exchange scheme between different force constants replicas are performed. Second, fitted Cα atom positions guide the all-atom structure in targeted MD. Finally, the all-atom flexible fitting refinement in implicit solvent adjusts the positions of the side chains in the density map. Final models obtained via the multi-scale protocol are significantly better resolved and more reliable in comparison with long all-atom flexible fitting simulations. The protocol is useful for multi-domain systems with intricate structural transitions as it preserves the secondary structure of single domains.

Keywords: flexible fitting, multi-scale methods, replica exchange, molecular dynamics simulation, cryo-EM density map, all-atom force field, coarse-grained force field, targeted molecular dynamics

INTRODUCTION

The technical advances in single-particle cryo-EM result in a rapid escalation of the number of available density maps at increasingly higher resolutions recently crossing the 1.3 Å barrier (Yip et al., 2020) (Nakane et al., 2020). Also, the lower molecular weight limit of this method already reached much below 50 kDa (Wu and Lander, 2020). In order to extract structural information from the experimental densities, computational techniques such as rigid-docking (Wriggers et al., 1999) (Roseman, 2000), flexible fitting (Tama et al., 2004) (Topf et al., 2008) (Trabuco et al., 2008) (Whitford et al., 2011) (Dou et al., 2017) (Mori et al., 2019), and *de novo* modeling (Lindert et al., 2012) (Wang et al., 2015) (Terashi and Kihara, 2018) are essential to assign or refine the atomic positions into those maps. The choice of the computational method depends on the map resolution (Villa and Lasker, 2014). At near-atomic resolution (typically 3-5 Å), the structural details visible in the density give enough confidence for building the models *de novo*, even though the completeness and accuracy of such models would be lower than those based on atomic resolution density maps (Wang et al., 2015) (Zivanov et al., 2018). However, the density maps do not always reach so high resolutions due to intrinsic flexibility or fluctuation of biomolecules, and the methods other than *de novo* building are often required in such cases.

Flexible fitting has been widely used to guide the initial structure toward the target density map, in which normal mode analysis (Tama et al., 2004) or MD algorithms are utilized (Trabuco et al., 2008)

(Orzechowski and Tama, 2008). Here, a biasing potential is added to the molecular mechanics force field, and its force constant regulates the magnitude of the fitting. The flexible fitting usually requires a starting structure in a different state or closely related to the one prepared from other methods such as homology modeling. The results of the flexible fitting are strongly system-dependent, often fail for complex conformational changes or for globular-shaped molecules when low- or medium-resolution maps (>5 Å) are treated. This is mainly due to degeneracy of the experimental data, which prevents us from determining a unique conformation (Goh et al., 2016). Another common problem is overfitting, which leads to excessive deformations of the fitted structure and usually occurs when the applied force is too strong. On the other hand, a too weak force constant brings poor results of the flexible fitting. This issue underlines the importance of the optimization of the force constant.

REUSfit is one of the useful methods to overcome this problem (Miyashita et al., 2017), which is an extension of the replica-exchange umbrella-sampling method (REUS) (Sugita et al., 2000). Here, copies of the original system are prepared, and different force constants in the biasing potential are assigned to each replica. The flexible fitting is carried out in parallel, where the force constants are exchanged during the simulation. REUSfit showed better results than simple MD-based flexible fitting, as it enables the automatic adjustment of the force constant (Miyashita et al., 2017). However, REUSfit is computationally expensive if it is employed with the all-atom (AA) model with explicit solvent.

To date, various molecular models at different resolutions have been developed in efficient and accurate MD simulations of proteins, ranging from Coarse-grained (CG) Cα models (Kolinski, 2004) (Kmiecik et al., 2016), through AA models in implicit solvent (Schaefer et al., 1998) (Grochowski and Trylska, 2008) (Onufriev and Case, 2019) and in explicit solvent (Zhou, 2003). CG models are computationally efficient, among which the Go model has been widely used for tackling large conformational changes of proteins (Taketomi et al., 1975) (Whitford et al., 2010) (Kobayashi et al., 2015). Implicit solvent model like the GB/SA model treats solvent molecules as the continuum to reduce the computational cost for the solute-solvent or solvent-solvent interaction calculations (Anandakrishnan et al., 2015). CG model is the most suitable for REUSfit, while high-resolution molecular model is eventually desired to refine the protein structure, as it reveals the atomic interactions in biomolecules at a greater level of detail and accuracy.

In this study, we propose a multi-scale protocol in the cryo-EM flexible fitting that can connect CG and AA models in order to reach the final structural models with a quality not accessible by using any of those simulation methods separately. We focus on medium resolution density maps, i.e., around 5-8 Å, since the flexible fitting is generally useful for such maps. We tested our protocol in various systems and found that it is particularly useful for the protein complexes with large conformational transitions, such as Ca^{2+}-ATPase and DNA polymerase.

METHODS

Flexible Fitting Algorithm

In MD simulations with flexible fitting, biomolecular structure is guided toward the EM density map by a biasing potential V_{EM}. This potential is simply added to the force field function and depends on a given force constant k and a cross-correlation coefficient CC between the target density map and the one calculated in the course of the simulation (Orzechowski and Tama, 2008):

$$V_{EM} = k(1 - CC).$$

CC is computed for each (i,j,k) voxel containing the target ρ^{targ} and calculated ρ^{calc} densities (Tama et al., 2004):

$$CC = \frac{\sum_{ijk} \rho^{targ}(i,j,k)\rho^{calc}(i,j,k)}{\sqrt{\sum_{ijk} \rho^{targ}(i,j,k)^2 \sum_{ijk} \rho^{calc}(i,j,k)^2}}$$

Introduction of REUS scheme (Sugita et al., 2000) to the flexible fitting simulations enables to dynamically adjust the force constant during the simulation. This approach attributes higher force constants to the replicas with higher CC, i.e., better fitted to the target density maps. The exchange probability between two replicas i and j, with force constants k_m and k_n and CC to the target map equal to CC^i and CC^j, respectively, is estimated in the following way (Miyashita et al., 2017):

$$w\left(X \to X'\right) \equiv w\left(x_m^{[i]} \middle| x_n^{[j]}\right) = \begin{cases} 1, & \text{for } \Delta \le 0 \\ \exp(-\Delta), & \text{for } \Delta > 0 \end{cases}$$

$$\Delta = \beta(k_n - k_m)\left(CC^j - CC^i\right)$$

The parameter β stands for the inverse temperature and here is the same for all replicas. This method was implemented in GENESIS as REUSfit (Miyashita et al., 2017). For details about the generation of density maps with a Gaussian mixture model and about the domain decomposition and atomic decomposition parallelization strategies developed to speed up the flexible fitting simulations, see (Mori et al., 2019).

Targeted MD Algorithm

In targeted MD, the aim is to simulate the conformational transition of a molecule at initial configuration vector $\boldsymbol{r}_I$ to the known target state given by configuration vector $\boldsymbol{r}_F$. For each configuration, a distance between the current and target configuration can be calculated as (Schlitter et al., 1994):

$$\rho = |r - r_F|$$

In order to force the system to undergo the desired transition, we introduce a force constraint F^c:

$$\phi(r) = |r - r_F|^2 - \rho^2 = 0$$

$$F^c \equiv \lambda \frac{d\phi}{dr} = 2\lambda(r - r_F)$$

The Lagrangian multiplier λ is calculated to satisfy the above equations. At every time step, ρ is diminished by Δρ until the final value ρ_f is reached:

$$\Delta\rho = \frac{\Delta t\left(\rho_0 - \rho_f\right)}{T},$$

where T stands for the total simulation time. In GENESIS, the RMSD is used instead of distance. Therefore, the implemented method is based on internal coordinates rather than on Cartesian coordinates.

Chosen Systems

Seven different systems, containing PDB structures in at least two different conformations, were selected for flexible fitting simulations using simulated density maps: Ca^{2+}-ATPase (Toyoshima and Nomura, 2002) (Toyoshima et al., 2000) (Toyoshima and Mizutani, 2004) (Toyoshima et al., 2007), Na^+ K^+-ATPase (Morth et al., 2007) (Kanai et al., 2013), Adenylate Kinase (Müller et al., 1996) (Müller and Schulz, 1992), Ribose Binding Protein (Björkman et al., 1994) (Björkman and Mowbray, 1998), Maltodextrin Binding Protein (Sharff et al., 1992) (Quiocho et al., 1997), Diphtheria Toxin (Bennett et al., 1994) (Bennett and Eisenberg, 1994) and CO Dehydrogenase (Darnault et al., 2003). In the case of simulations with the experimental density maps, two systems were chosen, for which PDB structures in two states, as well as an experimental Cryo-EM density map for one of those states, are available: Magnesium Transporter CorA (Matthies et al., 2016) and DNA Polymerase (Fernandez-Leiro et al., 2015).

CG Simulations

The MMTSB Go model Web Server was used to prepare the Cα model parameters based on the potential energy function developed by Karanicolas and Brooks (Karanicolas and Brooks, 2003). For the structures containing more than one protein chain, in-home scripts were prepared to combine parameters for different chains, taking into account the inter- and intra-chain interactions. Additionally, the native contact values between τ500 and the tail of PolIIIα were increased threefold to prevent τ500 from detaching from the polymerase complex, as this interaction was previously reported stable upon the studied conformational transition (Fernandez-Leiro et al., 2015). The initial and target PDB structures were superimposed using the Cα atoms. The simulated density maps were generated at 5 Å resolution with 2 Å voxel size using the target PDBs and emmap_generator tool from the GENESIS package (Mori et al., 2019). The resolution and voxel size of the experimental density maps were left unchanged and the negative values were removed with the SITUS voledit tool (Wriggers et al., 1999).

The flexible fitting molecular dynamics simulations with replica exchange (REUSfit) were performed using 32 replicas at force constants 500; 524; 555; 593; 636; 684; 737; 794; 856; 921; 991; 1,065; 1,144; 1,228; 1,318; 1,415; 1,519; 1,632; 1,755; 1,889; 2,035; 2,195; 2,371; 2,564; 2,777; 3,011; 3,269; 3,553; 3,865; 4,209; 4,586; 5,000 kcal/mol. The distribution of the force constants was based on a preliminary survey to ensure the sufficient overlap of the biasing potential between replicas. The cutoff distance for the native and non-native contact interaction terms was set to 20 Å. All bond lengths were fixed with the SHAKE algorithm (Ryckaert et al., 1977). The temperature was kept constant at 200 K using the Berendsen thermostat (Berendsen et al., 1984). The exchanges were attempted every 2 ps and each replica was simulated for 250 ns with a time step of 20 fs, except for DNA Polymerase, where 10 fs was used as a time step. The density map was recalculated at every time step with the truncation of the Gaussian function to zero for the densities below 1% of the maximum value (Tama et al., 2004). The REUSfit of each system were repeated 20 times with different initial velocities. The highest CC structures in each simulation were chosen as the best models in each run.

AA Simulations

CHARMM C36 force field parameters for proteins (Best et al., 2012) were prepared using psfgen plugin in VMD (Humphrey et al., 1996) with the patches for disulfide bonds where necessary. GBSA implicit solvent model with OBC2 parameters (Onufriev et al., 2004) was used with salt concentration 0.15 M and the surface tension coefficient 0.005 kcal·mol^{-1}·Å^{-2}. The dielectric constant was set to 78.5. For the AA simulations, the 1 Å voxel size of the simulated density maps was used. The experimental density maps were the same as at CG level. In the second step of the multi-scale protocol, the Cα positions from 5 best models from CG simulations were used as a target in targeted MD simulations of AA structures in implicit solvent. The simulations in NVT ensemble with Langevin thermostat with the friction coefficient gamma set at 5 ps^{-1} were performed for 1 ns with time step 2 fs. The final structure after targeted MD will have low RMSD value with respect to the target but we perform additional alignment using the Cα atoms to shift the center of mass of the system.

The last step of the multi-scale simulations is the refinement of the last frame of the targeted MD trajectories with the flexible fitting simulations to the density maps. The simulation is performed for 1 ns using a similar set of parameters as the Targeted MD simulation. For comparison, the AA flexible fitting simulations starting from original PDB structure are carried out in 5 replicates and referred to as the simple protocol. The best models from each of those simulations are chosen according to the highest CC. All simulations were done in GENESIS version 1.4 (https://www.r-ccs.riken.jp/labs/cbrt/).

Analysis

Various tools were used for the analysis of the best final models M1, M2, M3, SP and also S0 and RF: CC without rigid body docking was calculated in the collage program from Situs package (Wriggers et al., 1999); RMSD in Gromacs 2016.4 (Abraham et al., 2015) with gmx rms tool; CaBLAM and Ramachandran outliers and MolProbity score in Phenix (Williams et al., 2018) (Adams et al., 2010).

CaBLAM and Ramachandran outliers and MolProbity score depend on the relation between the geometry of the final models and a high-quality protein dataset from PDB database, called Top8000. CaBLAM is a method to evaluate the geometry of the protein backbone using the virtual angles and dihedrals created

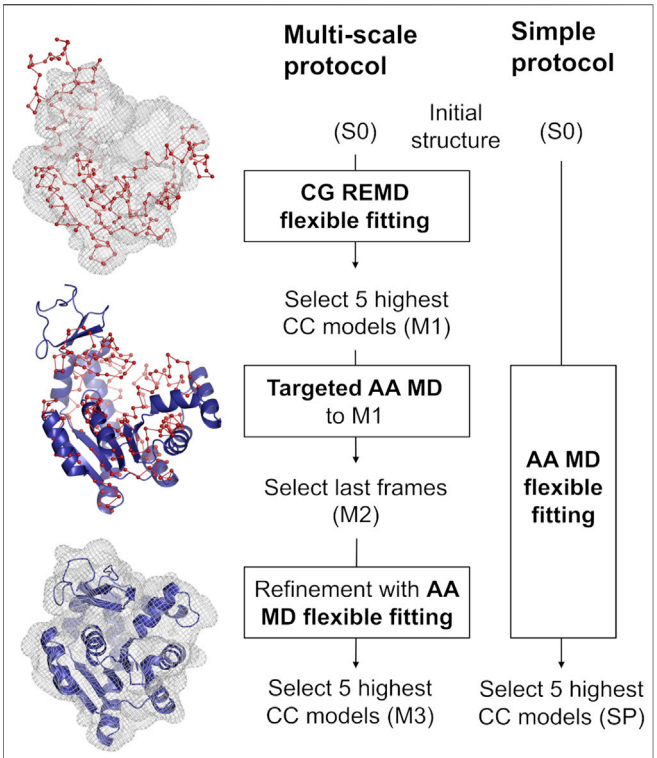


FIGURE 1 | Comparison between the multi-scale protocol and a simple flexible fitting protocol. The positions of Cα atoms in the CG model are shown as red spheres joined by a ribbon representing the pseudobonds. The blue cartoon depicts the protein modeled with AA in implicit solvent. Density map is shown in wireframe representation.

by the Cα atoms and carbonyl groups [details in (Williams et al., 2018)]. Ramachandran outliers use the φ and ψ angles of the protein backbone for analysis.

Secondary structure score is determined by comparing the secondary structure of the fitted model and the target, namely by counting the Hamming distance between two strings, representing the coded secondary structure features for each residue, as calculated by mkdssp version 2.2.1 (Touw et al., 2015) (Kabsch and Sander, 1983) and in-house scripts. The Hamming distance is normalized by dividing it by the length of the string and representing it as a percentage.

RESULTS

Multi-Scale Protocol for Flexible Fitting

Our multi-scale protocol involves three steps (**Figure 1**, **left scheme**). In step 1, REUSfit with the Cα Go model (Karanicolas and Brooks, 2003) is employed. In this model, one amino acid is represented by a single bead at their Cα atom position, and the beads interact via the structure-based potential that is constructed from the native structure. The interactions between side chains are scaled for respective amino acid types using Miyazawa-Jernigan contact energies (Miyazawa and Jernigan, 1996). In REUSfit, the cross-correlation-coefficient (CC) between the calculated and experimental density maps is included in the biasing potential. We carry out five individuals runs, and the best models, called M1, are chosen from each run according to the highest CC. In step 2, we carry out targeted MD (Schlitter et al., 1994) starting from the initial atomic structure toward M1, where the implicit solvent model (GB/SA) (Onufriev et al., 2004) is used to reduce the computational cost. Note that the experimental density map is not used in this step. Since targeted MD imposes the restraint on the root-mean-square deviation (RMSD) with respect to the target structure, the Cα coordinates in the AA model do not exactly match the target CG model. Therefore, additional alignment to M1 is necessary to shift the center of mass of the obtained structure after the targeted MD. The final atomic structure is called M2. In step 3, M2 is further refined with the CC-based flexible fitting using the experimental density map. Here, the GB/SA implicit solvent model is again used as in step 2. The final models are chosen according to their CC and named M3.

We compare the best final models obtained via two protocols, starting from the same initial structure S0: the multi-scale protocol and one, long, all-atom flexible fitting simulation called a simple protocol (SP) (**Figure 1**, **right scheme**). The computational condition is same as in step 3 of the multi-scale protocol.

The two protocols are applied to a range of systems with simulated and experimental density maps, shown in **Table 1**. Simulated density maps serve as a test of accuracy and precision of fitting. The density maps were generated from X-ray crystal structures of the target biomolecules at 5 Å resolution. Tested systems include a small protein such as Adenylate Kinase as well as relatively large ones, for instance, DNA polymerase. All tested systems are classified as having two significantly different states in initial and target structures. The transitions between the two states represent varying levels of difficulty in each case. Particularly, Ca^{2+}-ATPase and Diphtheria Toxin are difficult to fit correctly (initial RMSD >10 Å), while Ribose Binding Protein and Maltodextrin Binding Protein are relatively easy (initial RMSD <4 Å). For the experimental density maps, we have chosen two systems, for which both the cryo-EM density maps and fitted models were available, together with a starting structure in a different physiological state, involving complex conformational transition between the two states. Two systems fulfilling all the criteria were found, with cryo-EM density maps at resolution in the range 5-8 Å: magnesium transporter CorA (Matthies et al., 2016) and DNA polymerase (Fernandez-Leiro et al., 2015).

Large Conformational Transitions can be Fitted Well With Multi-Scale Protocol

First, we focus on Ca^{2+}-ATPase, which requires large conformational transitions during the flexible fitting. Ca^{2+}-ATPase undergoes a series of structural changes between E2, E1·2Ca^{2+}, E1P, and E2P states (**Supplementary Figure S1**), linked with Ca^{2+} transport against the concentration gradient (Toyoshima, 2008). If the flexible fitting is performed in these

TABLE 1 | Systems used in the flexible fitting simulations using either simulated or experimental density maps. The RMSD was calculated between the Cα atoms of the initial structure and the target structure, for the number of residues (and chains), specified in the last column.

System	State (PDB ID) of the initial → target structures	Density map, resolution [Å]	RMSDt (Cα) [Å]	# of residues (chains)
Ca^{2+}-ATPase	E2 (1iwo) → E1·2Ca^{2+} (1su4)	simulated, 5.0	14.0	994 (1)
	E1·2Ca^{2+} (1su4) → E1·ATP (1vfp)	simulated, 5.0	13.7	994 (1)
	E1·ATP (1vfp) → E2P (2zbf)	simulated, 5.0	11.2	994 (1)
	E2P (2zbf) → E2 (1iwo)	simulated, 5.0	5.5	994 (1)
Na^+ K^+-ATPase	K^+-E2 (3kdp) → Na^+-E1P·ADP (3wgu)	simulated, 5.0	11.6	1,280 (2)
	Na^+-E1P·ADP (3wgu) → K^+-E2 (3kdp)	simulated, 5.0	11.6	1,280 (2)
Adenylate kinase	open (4ake) → closed (1ake)	simulated, 5.0	7.1	214 (1)
Ribose binding protein	closed (2dri) → open (1urp)	simulated, 5.0	4.1	271 (1)
Maltodextrin binding protein	open (1omp) → closed (1anf)	simulated, 5.0	3.8	370 (1)
Diphtheria toxin	open (1ddt) → closed (1mdt)	simulated, 5.0	15.6	523 (2)
CO dehydrogenase	closed (1oaoC) → open (1oaoD)	simulated, 5.0	7.9	729 (1)
Magnesium transporter CorA	closed (3jcf) → open (3jcg)	Experimental (EMD-6552), 7.1	9.7	1,653 (5)
DNA polymerase	Bound (5fkv) → free (5fku)	Experimental (EMD-3201), 8.3	14.2	2,211 (7)

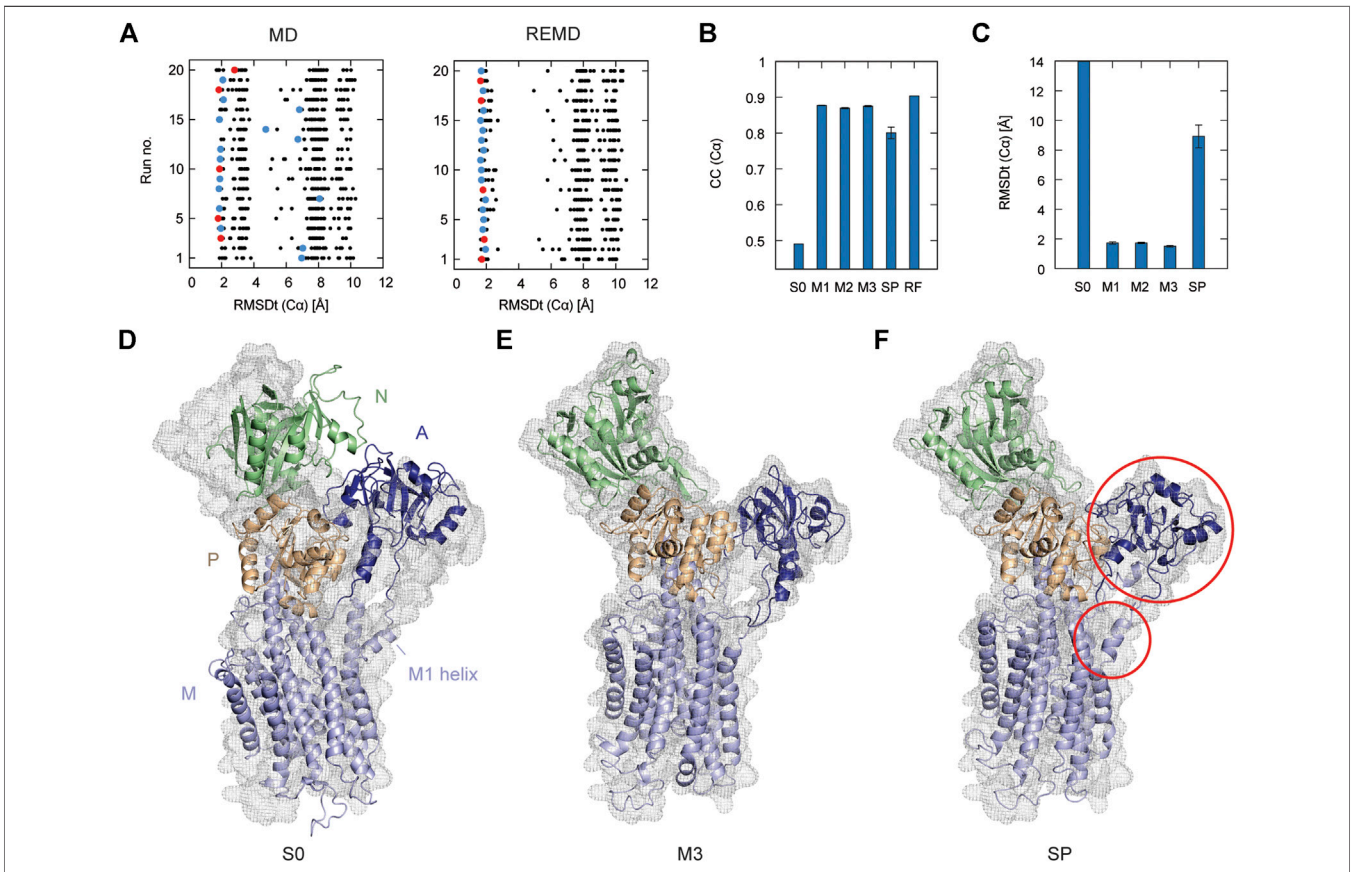


FIGURE 2 | Flexible fitting of Ca^{2+}-ATPase from E2 to E1·2Ca^{2+} state. **(A)** RMSDt of Cα atoms of best CC models of each of 32 MD replicas at different force constants for 20 independent simulation runs. The best CC replica in each run is colored and the chosen 5 best models are highlighted in red. MD and REMD panels show simulations without and with exchanges between replicas, respectively. **(B)** Average CC of the initial structure (S0) and 5 best models in the multi-scale protocol stages: CG REMD simulations (M1), targeted MD (M2), AA MD (M3) and in the simple protocol (SP). Standard deviation is shown as an error bar. **(C)** RMSDt of the (Cα) atoms of the same structures. **(D)** Initial structure used for simulations, superimposed on the target density map. **(E**, **F)** The highest CC model obtained using the multi-scale protocol and simple protocol, respectively.

states, the movements of cytoplasmatic domains (A, P, and N) and transmembrane helices in domain M are challenging for us, since the A domain rotates by 90° from the E1P to E2P state and the inclination of the N domain with respect to the P domain in the E1·2Ca^{2+} and E1·AMPPCP crystal structures changes by almost 90° as well.

TABLE 2 | Quality of the best models obtained using the simple and multi-scale protocol for the simulated density maps for Ca^{2+}-ATPase and Diphtheria Toxin.

System	Str.	c.c.	RMSDt (Cα) [Å]	RMSDt (AA) [Å]	MolProbity score	Ramachandran outlier [%]	Rotamer [%]	CABLAM [%]	Sec. struct. index [%]
Ca^{2+} ATPase									
E2 → E1·2Ca_{2+}	MS3	0.96 (0.95 ± 0.00)	1.4 (1.5 ± 0.0)	2.3 (2.4 ± 0.1)	1.67 (1.74 ± 0.05)	2.7 (3.3 ± 0.3)	6.3 (7.3 ± 0.7)	4.0 (5.3 ± 1.0)	22 (23 ± 1)
	SP	0.91 (0.90 ± 0.02)	8.8 (8.9 ± 0.8)	9.2 (9.4 ± 0.8)	1.85 (1.80 ± 0.04)	4.2 (4.8 ± 0.4)	8.5 (8.7 ± 0.7)	7.2 (7.5 ± 1.3)	31 (32 ± 1)
E1·2Ca_{2+} → E1·ATP	MS3	0.96 (0.95 ± 0.00)	1.5 (1.5 ± 0.1)	2.3 (2.3 ± 0.1)	1.79 (1.82 ± 0.03)	2.3 (2.4 ± 0.4)	9.1 (9.7 ± 1.1)	4.7 (4.0 ± 0.6)	21 (23 ± 1)
	SP	0.94 (0.94 ± 0.00)	3.0 (2.8 ± 0.3)	3.6 (3.5 ± 0.2)	1.83 (1.82 ± 0.05)	3.7 (3.6 ± 0.4)	9.0 (8.8 ± 1.0)	5.1 (5.8 ± 0.5)	22 (24 ± 1)
E1·ATP → E2P	MS3	0.96 (0.96 ± 0.00)	1.4 (1.4 ± 0.1)	2.3 (2.3 ± 0.1)	1.72 (1.75 ± 0.06)	2.5 (2.7 ± 0.4)	8.5 (7.9 ± 0.7)	2.8 (3.7 ± 0.5)	21 (21 ± 1)
	SP	0.93 (0.92 ± 0.02)	4.4 (5.1 ± 1.5)	4.9 (5.6 ± 1.5)	1.85 (1.84 ± 0.04)	4.3 (4.0 ± 0.7)	9.4 (8.4 ± 0.8)	6.3 (6.5 ± 0.8)	26 (28 ± 1)
E2P → E2	MS3	0.96 (0.96 ± 0.00)	1.3 (1.3 ± 0.0)	2.1 (2.1 ± 0.0)	1.56 (1.69 ± 0.02)	2.0 (2.0 ± 0.3)	6.5 (7.4 ± 0.6)	2.9 (3.2 ± 0.6)	21 (23 ± 1)
	SP	0.96 (0.95 ± 0.00)	1.4 (1.5 ± 0.1)	2.2 (2.3 ± 0.1)	1.63 (1.70 ± 0.05)	1.7 (2.1 ± 0.4)	6.5 (7.7 ± 1.2)	3.6 (3.7 ± 0.6)	24 (25 ± 2)
Diphtheria toxin									
	MS3	0.97 (0.97 ± 0.00)	0.8 (0.9 ± 0.1)	1.6 (1.7 ± 0.1)	1.71 (1.74 ± 0.03)	2.7 (3.0 ± 0.3)	7.8 (7.7 ± 1.0)	5.4 (5.2 ± 0.5)	18 (19 ± 2)
	SP	0.95 (0.95 ± 0.00)	9.9 (9.6 ± 0.3)	10.3 (9.9 ± 0.3)	1.93 (1.86 ± 0.05)	6.0 (5.4 ± 0.6)	8.3 (7.7 ± 0.5)	10.5 (8.9 ± 1.5)	38 (33 ± 3)

In **Figure 2A**, the left panel shows 20 individual runs of the flexible fitting from E2 to E1·2Ca^{2+} states using the CG model at different force constants (500–5,000 kcal/mol), which corresponds to REUSfit without replica exchange. The final fitted models with the highest CC are colored and chosen as the best models. Their RMSD with respect to the target structure (RMSDt) around 2 Å represent the successful simulations with well-fitted domain A. However, there are multiple failed attempts. Introducing the exchanges between the replicas brings substantial improvement, and all the attempts are successful (right panel in **Figure 2A**). Each fitting run of the multi-scale protocol is highly repeatable due to exchanges between replicas. Therefore, a single fitting simulation following the multi-scale procedure is sufficient to obtain a reliable result, which is visible in the low values of standard deviation of multiple replicates in **Table 2**. Best models from the simple protocol reveals significantly higher standard deviations. Similar results were obtained in the previous work using 10 Å resolution maps (Miyashita et al., 2017). Thus, the structures with the correct Cα atoms positions (M1) can be further used in the multi-scale protocol in targeted MD and refined to obtain M2 and M3 models, respectively.

In **Figure 2B**, we summarize the results of the multi-scaling protocol, comparing the averaged CC of the obtained M1, M2, and M3 models. For comparison, CC of the initial structure S0, reference structure RF, and the results of the simple protocol (SP) are also shown. The best models obtained using our multi-scale protocol (CC = ~0.88) are significantly better than that in the simple protocol (CC = ~0.80), indicating a successful fitting to the density map. RMSDt of the M3 model is also smaller (1.5 Å) than that of the SP model (8.9 Å) (**Figure 2C**). In addition, the MolProbity score (Keedy et al., 2009), CaBLAM (Williams et al., 2018), and Ramachandran outliers of the M3 model are better than the SP model (**Supplementary Figure S2**). The RMSDt of the SP models are closer to the S0 structure, implying that during the single-step fitting, smaller conformational changes occurred than during the three-step multi-scale protocol. In fact, RMSDi (RMSD with respect to the initial structure) of the M2 and M3 models is larger than that of the SP model (**Supplementary Figure S2**).

In the multi-scale protocol, all three steps show similar averaged values of the CC (Cα) and RMSDt (Cα), around 0.88 and 1.5–1.7 Å, suggesting that the first CG simulation with REUSfit is the most important to determine the positions of the Cα atoms. Targeted MD is also working well to superimpose the all-atom model to the CC model by means of a gradual decrease of the RMSD of Cα atoms to their target positions. The protein side chains follow the conformational changes induced by the shift of Cα atoms. The detailed conformation of the protein side chains can be refined in step3, as the density map is not present during the targeted MD. We found that RMSDt of all non-hydrogen atoms is improved from 2.7 to 2.4 Å (**Supplementary Figure S2**). Although each domain undergoes almost rigid-body rotation and thus most side chains are already well positioned in the targeted MD, further refinement seems to be possible in step3.

One of the main differences between the M3 and SP models is the orientation of the A domain. In the initial structure, the N and A domains are contacting each other (**Figure 2D**), and they are separated during the fitting. The A domain acts as an actuator of

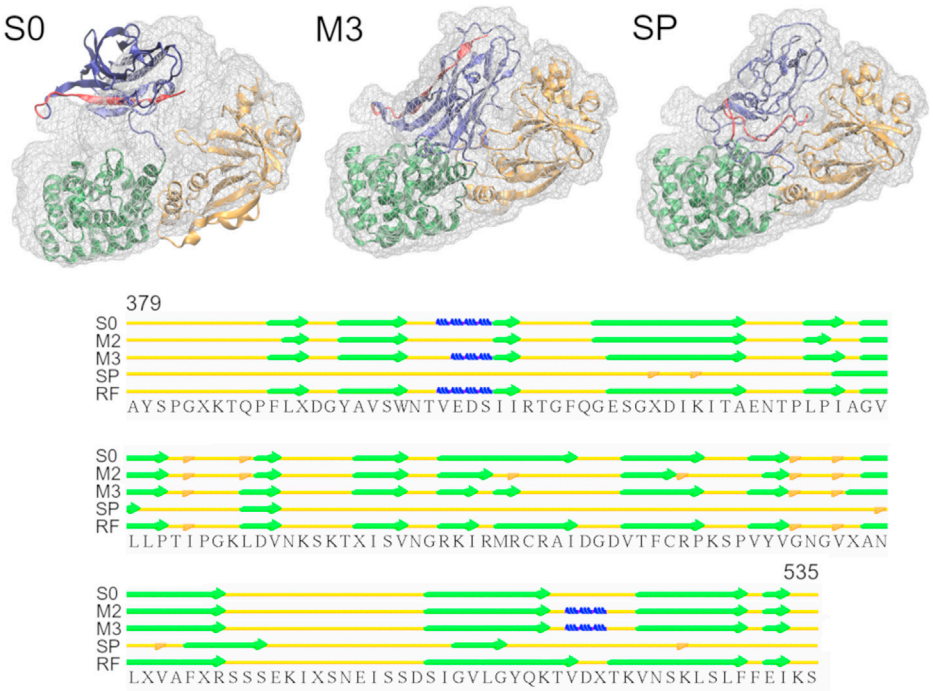


FIGURE 3 | Simulations of Diphtheria Toxin. Initial structure (S0), the highest CC models obtained using the multi-scale protocol (M3) and the simple protocol (SP) are superimposed with the target density map. The R domain is shown in blue with residues 520-535 highlighted in red. The graph below depicts the secondary structure elements in the R domain of the structures from M3, SP and the reference structure (RF), as assigned by the Stride web server. β-strands, 3-10 helices and isolated β-bridges are shown in green, blue and orange, respectively.

the gating of the ion pathway and it has to rotate in order to regulate the Ca^{2+} binding and release. In the CG simulation, this shift occurs with a domain rotation. This success extends to the all-atom level in the next stage of the multi-scale protocol, resulting in a correct final model (**Figure 2E** and **Supplementary Movie S1**). On the other hand, in the simple protocol the A domain was simply fitted to the density without correct rotation, and none of the simulations was successful (**Figure 2F** and **Supplementary Movie S2**). Since the orientation of the A domain can affect the transmembrane domain through the loop connecting them, the conformation of the N-terminal side of the M1 helix is also reproduced in the M3 model.

Interestingly, troublesome regions in the flexible fitting in the other physiological states of Ca^{2+}-ATPase and Na^{+} K^{+}-ATPase are similar, including the transmembrane M1 and M2 helices and A domain. (**Supplementary Figure S3**). In these cases, the multi-scale protocol still performs significantly better than the simple protocol (see **Table 2** below). The main reason behind this is the rotation of the A domain between E1 and E2-type states, performed correctly in the multi-scale protocol and incorrectly in the simple protocol. Without the proper rotation, the secondary structure of the A domain in SP models becomes deformed to accommodate to the density map. In the transition from E1·2Ca^{2+} to E1·ATP state, the transmembrane helices M1 and M2 undergo shifting and rearrangement with bending of the N-terminal of the M1 helix. This is reasonably well reflected in M3 models but not in SP models, where M1 helix is bent in opposite way than expected.

In order to represent the degree of disruption of the secondary structure of a protein, we are introducing in this work a new measure called a "secondary structure score." This score is based on a Hamming distance between two strings, containing the coded secondary structure elements of each residue of two structures that are compared (see the details in the Methods section). The smaller is the value of the secondary structure score, normalized and indicated as a percentage, the more similar are two structures. The last panel of the **Supplementary Figure S2** shows the secondary structure scores calculated between the target structure and each model M2, M3 and SP. Here, it is well visible that the secondary structure of the best SP model (31%) deviates much more from the target structure than the best M3 model (22%).

Diphtheria Toxin also requires correct domain rotation during the fitting. The initial structure is shown in the S0 panel in **Figure 3**, where RMSDt is 15.6 Å. To emphasize the change in orientation of the relevant domain (R domain), its C-terminal fragment was highlighted in red. The M3 and SP panels are the highest-CC models obtained from the multi-scale and simple protocols, respectively. The M3 model showed correct orientation of the domain R [average RMSDt (Cα) = 0.9, RMSDi (Cα) = 15.6, and CC (Cα) = 0.89], while it was wrong in the SP model [average RMSDt (Cα) = 9.6, RMSDi (Cα) = 11.7, and CC (Cα) = 0.84]. In addition, there was a disruption of the secondary structure of several β-strands in the R domain of the SP model (bottom panel in **Figure 3** and **Supplementary Figure S4**, average secondary structure score = 33%). Such disruption is not observed in the

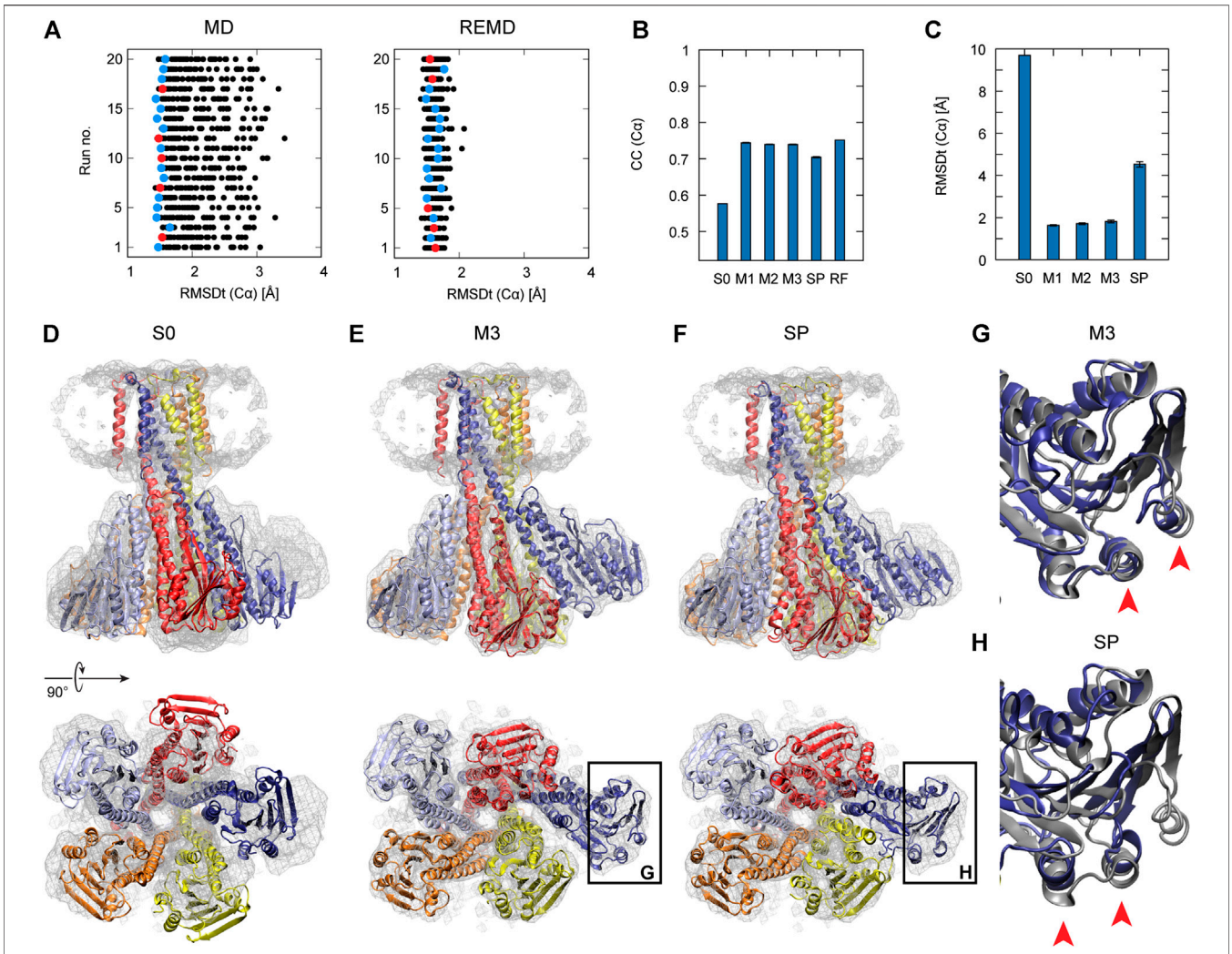


FIGURE 4 | Fitting of Magnesium Transporter CorA to experimental density map. **(A)** RMSDt of Cα atoms of best CC models for runs of simulations without (MD) and with exchanges between replicas (REMD), respectively. 5 best models are highlighted in red. **(B)** Average CC of the initial structure (S0) and 5 best models in the multi-scale protocol stages (M1, M2, M3) and in the simple protocol (SP) with standard deviation as an error bar. **(C)** Average RMSDt of the (Cα) atoms. **(D)** The side view and bottom view of the initial structure with the target density map. **(E)** The highest CC models obtained using the multi-scale and **(F)** simple protocol. Panels **(G, H)** show the enlarged view on the chain A with the reference structure in dark gray.

multi-scale protocol (average sec. struct. score = 19%), presumably because the CG simulations with the Go model enable keeping the native contacts within the domain. A slight improvement of the secondary structure during the final refinement in the multi-scale protocol may be caused by the adjustments of the side chains into the density map, which may influence the backbone geometry and hydrogen bonds between strands.

Simple Conformational Changes in Small Proteins do not Require Multi-Scale Procedure

The flexible fitting simulations were also performed for several small systems: Adenylate Kinase, Maltodextrin Binding Protein, Ribose Binding Protein, and CO Dehydrogenase. Even though the two conformational states of those systems are classified as significantly different and referred to as "open" and "closed" states in the literatures, the overall transition between them is very simple and does not involve large rotation or shift in the component domains. For those systems, the best models obtained from the multi-scale protocol are similar to those from the simple protocol (**Supplementary Figure S5**). As a representative, let us take Ribose Binding Protein. The CC (Cα) in the M1, M2, M3, SP, and RF models is 0.89–0.90, and the RMSDt (Cα) is ~0.6 Å. We found that the multi-scale protocol gives structures of a comparable quality in terms of the MolProbity scores. The average score is 1.34 ± 0.11 in M3, while 1.23 ± 0.08 in SP (see **Supplementary Table S1**). The average values of Ramachandran, rotamer and CaBLAM outliers are similar within the standard deviation. However, the best M3 models selected basing on the CC are slightly better than the best SP models. Although the multi-scale protocol is applicable to

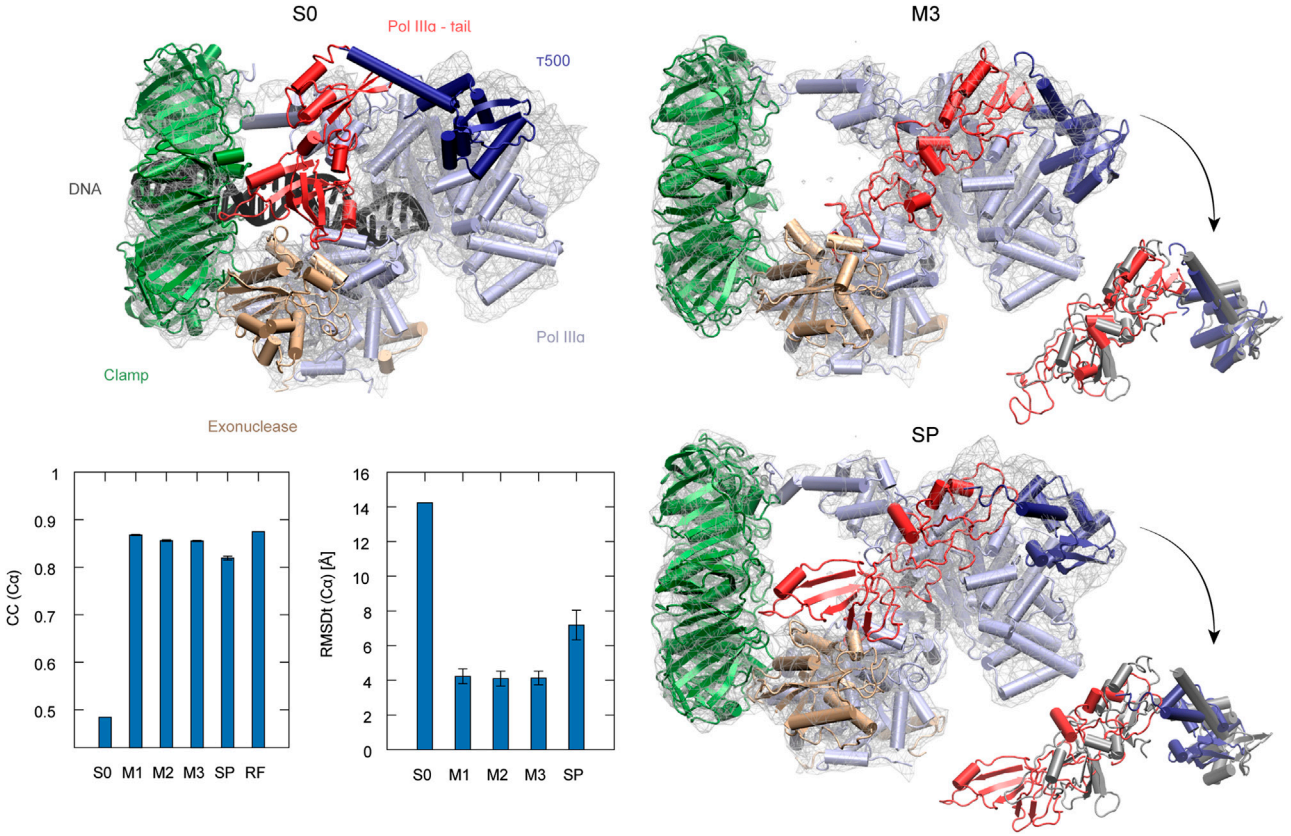


FIGURE 5 | DNA polymerase fitting. S0: Initial structure with the target density map. Position of DNA is shown in black (not included in simulations). M3 and SP are the highest CC models obtained with the multi-scale and simple protocols, respectively. τ500 and tail of PolIIIα are additionally superimposed with the reference structure (RF) in dark gray in the insets. The average CC and RMSDt for 5 best models in each run are shown in the graphs with standard deviation marked as error bars.

small proteins, the simple protocol seems to be enough in most cases. The multi-scale protocol is useful to search large conformational space. Involving the CG models in the protocol is much more beneficial for large proteins and their complexes.

The Multi-Scale Protocol Performs Well for Experimental Cryo-EM Density Maps

CorA is responsible for magnesium uptake in prokaryotes. In the closed Mg^{2+}-bound state, it forms a symmetrical pentamer. While moving to the open Mg^{2+}-free state, the 5-fold symmetry is lost. The density map of the latter state is used here as the target of flexible fitting simulations. In order to assess the performance of our protocol by calculating RMSDt, we have used the model fitted by rigid body methods and deposited by the density map authors in the RCSB PDB database. We realize that this model may not be highly accurate due to its low resolution but that is the only "answer" model that could be used for comparison and we may possibly contribute to its refinement.

To examine the usefulness of REUSfit in the multi-scale protocol, we first compare CG MD with and without replica exchange. In **Figure 4A**, we see that REUSfit produced all the models in the RMSDt range 1–2 Å, suggesting that we can select a reliable model even from one simulation without considering the effect of the force constant. Since the targeted MD worked well in step 2, the all-atom model was superimposed to the CG model effectively. As a result, the multi-scale protocol gives significantly better models than simple protocol in terms of both CC (Cα): 0.74 vs 0.70 and RMSDt (Cα): 1.8 vs 4.5 Å (**Figures 4B,C**). Similar tendency was obtained in the averaged RMSDt for non-hydrogen atoms, CaBLAM outliers, MolProbity score, and secondary structure score (Figure S7). The reason behind it is the difficulty of fitting is the skew of chain A (dark blue in **Figures 4D–F**). The comparison of two chosen helices of chain A with the target structures shows that they were fitted correctly by the multi-scale protocol in the CG step of the protocol using REUSfit and trapped in a wrong density patch in the best model obtained with the simple protocol (red arrows in **Figures 4G,H**). The REUSfit method is able to rescue the system from being trapped in local density minima by the efficient conformational sampling.

Step 3 did not improve the structure with respect to the target model, as both CC and RMSDt values of the M2 and M3 models are similar. As already mentioned, the target model is not a high-resolution structure of the system. Also, the resolution of the density map is 7.1 Å. Therefore, it is difficult to conclude whether the final refinement step brings any improvement to the structure. It is expected that for the target density maps of

TABLE 3 | Quality of the best models obtained using the simple and multi-scale protocol for the experimental density maps.

System	Str.	c.c.	RMSDt (Cα) [Å]	RMSDt (AA) [Å]	MolProbity score	Ramachandran outlier [%]	Rotamer outlier [%]	CABLAM outlier [%]	Sec. struct. index [%]
Magnesium transporter CorA	MS3	0.77 (0.77 ± 0.00)	1.7 (1.8 ± 0.1)	2.9 (3.0 ± 0.0)	1.70 (1.70 ± 0.02)	2.8 (3.0 ± 0.6)	7.3 (7.3 ± 0.4)	5.3 (4.5 ± 0.5)	18 (19 ± 1)
	SP	0.74 (0.74 ± 0.00)	4.4 (4.5 ± 0.1)	5.1 (5.1 ± 0.1)	1.75 (1.76 ± 0.01)	2.9 (3.3 ± 0.5)	7.2 (7.7 ± 0.3)	4.8 (5.9 ± 0.7)	22 (22 ± 1)
DNA polymerase	MS3	0.85 (0.85 ± 0.00)	4.1 (4.1 ± 0.4)	4.5 (4.5 ± 0.3)	1.84 (1.80 ± 0.03)	3.6 (3.6 ± 0.4)	9.3 (8.2 ± 0.7)	6.7 (6.3 ± 0.4)	21 (21 ± 1)
	SP	0.82 (0.82 ± 0.00)	6.3 (7.2 ± 0.8)	6.6 (7.5 ± 0.8)	1.84 (1.81 ± 0.02)	4.4 (3.9 ± 0.4)	8.5 (8.1 ± 0.4)	6.1 (6.5 ± 0.3)	23 (24 ± 1)

higher resolution, step 3 would contribute more to the quality of the structure.

The second target using the experimental density map is the DNA polymerase III from *E. coli*. It serves as an enzyme for DNA replication. The complex consists of several factors of different functionality, such as the clamp for sliding along the DNA and exonuclease working as a proofreader. To build the DNA lagging strand, the polymerase needs to quickly and repeatedly release and reposition DNA. In order to understand this process, several structures of the complex have been solved, both in DNA-bound and DNA-free states (Fernandez-Leiro et al., 2015). Here, we performed flexible fitting to a density map of a DNA-free state starting from the DNA-bound state structure. DNA was removed from the original PDB structure before the simulations (Dark gray in S0 panel of **Figure 5**).

The best model obtained with the simple protocol reveals many inaccuracies with respect to the density map in the regions of the τ500 domain and neighboring tail of PolIIIα, when compared with the modeled reference structure (SP panel in **Figure 5**). Particularly, the tail of PolIIIα that undergoes a rearrangement upon DNA binding is not in a correct position, significantly stretching out of the contours of the density map. Applying the multi-scale protocol allowed avoiding such problems. The secondary structure of those domains is retained with their simultaneous shift toward the correct position within the target density map (M3 panel in **Figure 5**). Both CC and RMSDt of the Cα atoms of the best final models are significantly improved by using the multi-scale protocol, from 0.82 to 0.85 and from 7.2 to 4.1, respectively (bottom left panel in **Figure 5**). Similarly, as in the case of the CorA, the CG simulations with REUSfit largely contribute to the correct fitting.

Overall Performance of Multi-Scale Protocol

In **Tables 2**, **3** and **Supplementary Table S1**, we summarize the CC, RMSD, and other parameters of the best models obtained from the simple and multi-scale protocols for all test sets. In the cases of Adenylate Kinase, Maltodextrin Binding Protein, Ribose Binding Protein, and CO Dehydrogenase, which include a simple rotation or shift of the component domains, the two protocols are almost comparable. However, in the case of Ca^{2+}-ATPase, Na^{+} K^{+}-ATPase, Diphtheria Toxin, Magnesium Transporter, and DNA polymerase, which includes complicated conformational changes, multi-scale protocol showed better scores than the simple protocol. In the simple protocol, RMSDt of all the heavy atoms is high, reaching even 9.2 Å for Ca^{2+}-ATPase and 10.3 Å for Diphtheria Toxin. In some cases, we also observe disruptions of the secondary structures in the rotated/shifted domain, resulting in higher secondary structure scores (>30%), and thus, poor MolProbity score and high Ramachandran and CaBLAM outliers. In the multi-scale protocol, RMSDt came down below 2.5 Å in all systems with simulated density maps and to 2.9 and 4.5 Å in the experimental density maps systems. The secondary structure scores are mostly less than 30%, and the main chain geometry is consistently at least 0.5% better than for the SP best models. We suggest that multi-scale protocol is useful for large proteins or their complexes in which a complex conformational transition, such as a rotation or domain rearrangement, occurs between two states.

DISCUSSION

To date, several methods relying on MD simulations have been developed in order to flexibly fit protein structures to the target density maps. Here, we use the CC-based approach (Tama et al., 2004). This can be combined with structure-based force fields, as in MDfit (Whitford et al., 2011), which enables us to prevent the structure from breaking secondary or tertiary structure contacts during the fitting due to native contact energy term. However, new stable contacts cannot be created during the fitting, even if there is a contact in the target structure. MDFF uses the physics-based force field such as CHARMM, which allows for creating new contacts, but requires secondary structure restraint (Villa and Lasker, 2014) (Trabuco et al., 2008). In our approach, we aim to design a fitting method that does not only keep secondary and tertiary structure relations, but also allows for the refinement of side chain positions, including breaking or creating new contacts. Therefore, we use a structure-based model for the first stage of the protocol, and CHARMM force field with the CC-based biasing potential for the third stage, joined together by targeted MD. One advantage of our method is that in the second and third stages we do not introduce any secondary structure restraints, as in CC-based methods, but we take advantage of those restraints at an earlier step by using the structure-based model. Although the protocol requires the correct secondary structure in the initial model, it allows the changes of the secondary structure in the target structures via the non-restraint MD simulations.

Another advantage of our protocol is that we are searching a large conformational space in the first stage with REUSfit (Miyashita et al., 2017), where the computational cost is reduced with the CG model.

REUSfit increases the reliability of the fitting, since exchanging the force constants between replicas can promote a feedback loop to output the structures with the highest CC in different fitting trials. Extending the normal MD simulation or using other methods increasing the sampling such as accelerated MD would also lead to a wide conformational search (Hamelberg et al., 2004). However, it may not choose and promote the same fitted pose in many runs when dealing with a complex conformational transition.

For systems with large and complex conformational changes, it is difficult to obtain reliable results just using a simple method, as the results may not be converged (Ahmed et al., 2012). Frequently, wrong fitting poses are obtained, similarly as in the case of Diphtheria Toxin presented here. Thus, for such cases, our multi-scale approach combining CG and AA model is powerful to reach a final model with reliable backbone positions by REUSfit and also containing hydrogen atoms. The consideration of the hydrogen atoms and hydrogen bonds is of importance for proteins and it allows for refinement of the side chains. Introducing a protocol consisting of three steps may be seen as a disadvantage of our method due to the relative complexity of setting up the files for both CG and AA simulations and running 3 different types of simulations consecutively in comparison with a single flexible fitting run. All necessary functionalities are implemented within the GENESIS suite of programs (Jung et al., 2015) (Kobayashi et al., 2017) (Mori et al., 2019).

The protocol uses targeted MD to convey the information about the correct Cα atoms positions after structure-based model fitting to the AA simulations. The most straightforward way one could imagine for going from Cα only model to AA model would be to rebuild the protein backbone and side chains basing on conventional geometrical principles. Various tools were examined in step 2, including for example PULCHRA (Rotkiewicz and Skolnick, 2008) and SCWRL4 (Krivov et al., 2009). In our experiences, none of them was free of severe errors in rebuilt structures, especially in the case of the slightly distorted main chain after flexible fitting simulation. Common problems included strange dihedral angles of the backbone, overlapping side chains and ring penetration (i.e., insertion of a covalent bond into an aromatic ring). We consider that targeted MD is generally applicable to back mapping in multi-scale simulations that combine CGMD and AAMD when at least one reliable experimental structure is available. Another method that could be taken into consideration for retrieving atomic detail from CG model is a recently developed back mapping method using Bayesian inference and restrained MD (Peng et al., 2019).

In this study, we tried to obtain the one best model from one experimental density map. Understanding the experimental data as an ensemble average of the multiple conformations is also important, since many conformations can reproduce the same experimental data (degeneracy problem) (Rieping et al., 2005) (Cossio and Hummer, 2013) (Olsson et al., 2017). In the cryo-EM, multiple states of proteins can be classified with the 3D reconstruction protocol for a large amount of 2D images (Bai et al., 2015). However, the obtained 3D density map is still an ensemble average of molecules with sufficient conformational fluctuations in solution. In addition, noise or random errors might be contained in the map to some extent, even if high-resolution density maps are obtained. To address these issues in the structure modeling, various inferential algorithms have been proposed (Bonomi et al., 2017). Particularly, metainference calculates the ensemble average using multiple-replica simulations, and also considers effects of the noise and errors in the energy function based on the Bayesian inference (Bonomi et al., 2016), which was recently applied to the cryo-EM structure modeling (Bonomi et al., 2018). Introduction of the ensemble average, noise, or errors will be one of the future extensions of our protocol.

CONCLUSION

We have designed a multi-scale protocol by integrating the CGMD with the AAMD in a single flexible fitting workflow. Performed simulations for several systems containing intricate domain rotations, domain shifts and transitions between open and closed states show that we can improve the accuracy and reliability of the flexible fitting procedure in comparison with a single, long AA flexible fitting simulation. The protocol performs well both for simulated and experimental density maps from cryo-EM and it is recommended to use with large proteins or their complexes.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author. The simulation protocol will be available as a tutorial "Multi-scale Cryo-EM flexible fitting" at the GENESIS webpage (https://www.r-ccs.riken.jp/labs/cbrt/).

AUTHOR CONTRIBUTIONS

Conceptualization, YS; Methodology and Visualization, MK, TM, and YS; Investigation, MK; Resources, YS; Writing—Original Draft, MK; Writing—Review & Editing, MK, TM, and YS; Funding Acquisition, YS; Supervision, YS.

FUNDING

This work was supported by the RIKEN Pioneering Project Dynamic Structural Biology and MEXT/Kakenhi (Grant number 19H05645 to YS). MD simulations were carried out at RIKEN on HOKUSAI BigWaterFall.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.631854/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

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., et al. (2010). PHENIX: a comprehensive python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* 266 (Pt. 2), 213–221. doi:10.1107/S0907444909052925

Ahmed, A., Whitford, P. C., Sanbonmatsu, K. Y., and Tama, F. (2012). Consensus among flexible fitting approaches improves the interpretation of cryo-EM data. *J. Struct. Biol.* 177, 561–570. doi:10.1016/j.jsb.2011.10.002

Anandakrishnan, R., Drozdetski, A., Walker, R. C., and Onufriev, A. V. (2015). Speed of conformational change: comparing explicit and implicit solvent molecular dynamics simulations. *Biophys. J.* 108, 1153–1164. doi:10.1016/j.bpj.2014.12.047

Bai, X. C., Rajendra, E., Yang, G., Shi, Y., and Scheres, S. H. (2015). Sampling the conformational space of the catalytic subunit of human γ-secretase. *eLife* 4, e11182. doi:10.7554/eLife.11182

Bennett, M. J., Choe, S., and Eisenberg, D. (1994). Refined structure of dimeric diphtheria toxin at 2.0 A resolution. *Protein Sci.* 3 (9), 1444–1463. doi:10.1002/pro.5560030911

Bennett, M. J., and Eisenberg, D. (1994). Refined structure of monomeric diphtheria toxin at 2.3 A resolution. *Protein Sci.* 3 (9), 1464–1475. doi:10.1002/pro.5560030912

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

Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., et al. (2012). Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ1 and χ2 dihedral angles. *J. Chem. Theory Comput.* 8 (9), 3257–3273. doi:10.1021/ct300400x

Björkman, A. J., Binnie, R. A., Zhang, H., Cole, L. B., Hermodson, M. A., and Mowbray, S. L. (1994). Probing protein-protein interactions. the ribose-binding protein in bacterial transport and chemotaxis. *J. Biol. Chem.* 269 (48), 30206–30211. doi:10.1016/s0021-9258(18)43798-2

Björkman, A. J., and Mowbray, S. L. (1998). Multiple open forms of ribose-binding protein trace the path of its conformational change. *J. Mol. Biol.* 279 (3), 651–664. doi:10.1006/jmbi.1998.1785

Bonomi, M., Camilloni, C., Cavalli, A., and Vendruscolo, M. (2016). Metainference: a Bayesian inference method for heterogeneous systems. *Sci. Adv.* 2 (1), e1501177. doi:10.1126/sciadv.1501177

Bonomi, M., Heller, G. T., Camilloni, C., and Vendruscolo, M. (2017). Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* 42, 106–116. doi:10.1016/j.sbi.2016.12.004

Bonomi, M., Pellarin, R., and Vendruscolo, M. (2018). Simultaneous determination of protein structure and dynamics using cryo-electron microscopy. *Biophys. J.* 114 (7), 1604–1613. doi:10.1016/j.bpj.2018.02.028

Cossio, P., and Hummer, G. (2013). Bayesian analysis of individual electron microscopy images: towards structures of dynamic and heterogeneous biomolecular assemblies. *J. Struct. Biol.* 184 (3), 427–437. doi:10.1016/j.jsb.2013.10.006

Darnault, C., Volbeda, A., Kim, E. J., Legrand, P., Vernède, X., Lindahl, P. A., et al. (2003). Ni-Zn-[Fe4-S4] and Ni-Ni-[Fe4-S4] clusters in closed and open α subunits of acetyl-CoA synthase/carbon monoxide dehydrogenase. *Nat. Struct. Biol.* 10 (4), 271–279. doi:10.1038/nsb912

Dou, H., Burrows, D. W., Baker, M. L., and Ju, T. (2017). Flexible fitting of atomic models into cryo-EM density maps guided by helix correspondences. *Biophys. J.* 112 (12), 2479–2493. doi:10.1016/j.bpj.2017.04.054

Fernandez-Leiro, R., Conrad, J., Scheres, S. H., and Lamers, M. H. (2015). cryo-EM structures of the *E. coli* replicative DNA polymerase reveal its dynamic interactions with the DNA sliding clamp, exonuclease and τ. *eLife* 4, 1–16. doi:10.7554/eLife.11134

Goh, B. C., Hadden, J. A., Bernardi, R. C., Singharoy, A., McGreevy, R., Rudack, T., et al. (2016). Computational methodologies for real-space structural refinement of large macromolecular complexes. *Annu. Rev. Biophys.* 45, 253–278. doi:10.1146/annurev-biophys-062215-011113

Grochowski, P., and Trylska, J. (2008). Continuum molecular electrostatics, salt effects, and counterion binding-A review of the Poisson-Boltzmann theory and its modifications. *Biopolymers* 89, 93–113. doi:10.1002/bip.20877

Hamelberg, D., Mongan, J., and McCammon, J. A. (2004). Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* 120, 11919–11929. doi:10.1063/1.1755656

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graph* 14 (1), 33–38. doi:10.1016/0263-7855(96)00018-5 27-8

Jung, J., Mori, T., Kobayashi, C., Matsunaga, Y., Yoda, T., Feig, M., et al. (2015). GENESIS: a hybrid-parallel and multi-scale molecular dynamics simulator with enhanced sampling algorithms for biomolecular and cellular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 5, 310–323. doi:10.1002/wcms.1220

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22 (12), 2577–2637. doi:10.1002/bip.360221211

Kanai, R., Ogawa, H., Vilsen, B., Cornelius, F., and Toyoshima, C. (2013). Crystal structure of a Na+-bound Na+,K+-ATPase preceding the E1P state. *Nature* 502 (7470), 201–206. doi:10.1038/nature12578

Karanicolas, J., and Brooks, C. L. (2003). Improved gō-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *J. Mol. Biol.* 334, 309–325. doi:10.1016/j.jmb.2003.09.047

Keedy, D. A., Williams, C. J., Headd, J. J., Arendall, W. B., Chen, V. B., Kapral, G. J., et al. (2009). The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. *Proteins* 77 (Suppl. 9), 29–49. doi:10.1002/prot.22551

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., and Kolinski, A. (2016). Coarse-grained protein models and their applications. *Chem. Rev.* 116 (14), 7898–7936. doi:10.1021/acs.chemrev.6b00163

Kobayashi, C., Jung, J., Matsunaga, Y., Mori, T., Ando, T., Tamura, K., et al. (2017). GENESIS 1.1: a hybrid-parallel molecular dynamics simulator with enhanced sampling algorithms on multiple computational platforms. *J. Comput. Chem.* 38, 2193–2206. doi:10.1002/jcc.24874

Kobayashi, C., Matsunaga, Y., Koike, R., Ota, M., and Sugita, Y. (2015). Domain motion enhanced (DoME) model for efficient conformational sampling of multidomain proteins. *J. Phys. Chem. B* 119, 14584–14593. doi:10.1021/acs.jpcb.5b07668

Kolinski, A. (2004). Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.* 51 (2), 349–371. doi:10.18388/abp.2004_3575

Krivov, G. G., Shapovalov, M. V., and Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* 77, 778–795. doi:10.1002/prot.22488

Lindert, S., Alexander, N., Wötzel, N., Karakaş, M., Stewart, P. L., and Meiler, J. (2012). EM-fold: de novo atomic-detail protein structure determination from medium-resolution density maps. *Structure* 20, 464–478. doi:10.1016/j.str.2012.01.023

Matthies, D., Dalmas, O., Borgnia, M. J., Dominik, P. K., Merk, A., Rao, P., et al. (2016). Cryo-EM structures of the magnesium channel CorA reveal symmetry break upon gating. *Cell* 164, 747–756. doi:10.1016/j.cell.2015.12.055

Miyashita, O., Kobayashi, C., Mori, T., Sugita, Y., and Tama, F. (2017). Flexible fitting to cryo-EM density map using ensemble molecular dynamics simulations. *J. Comput. Chem.* 38 (16), 1447–1461. doi:10.1002/jcc.24785

Miyazawa, S., and Jernigan, R. L. (1996). Residue—residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256, 623–644. doi:10.1006/jmbi.1996.0114

Mori, T., Kulik, M., Miyashita, O., Jung, J., Tama, F., and Sugita, Y. (2019). Acceleration of cryo-EM flexible fitting for large biomolecular systems by efficient space partitioning. *Structure* 27, 161–174. doi:10.1016/j.str.2018.09.004 e3

Morth, J. P., Pedersen, B. P., Toustrup-Jensen, M. S., Sørensen, T. L., Petersen, J., Andersen, J. P., et al. (2007). Crystal structure of the sodium-potassium pump. *Nature* 450, 1043–1049. doi:10.1038/nature06419

Müller, C. W., Schlauderer, G. J., Reinstein, J., and Schulz, G. E. (1996). Adenylate kinase motions during catalysis: an energetic counterweight

balancing substrate binding. *Structure* 4 (2), 147–156. doi:10.1016/s0969-2126(96)00018-4

Müller, C. W., and Schulz, G. E. (1992). Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor Ap5A refined at 1.9 A resolution. a model for a catalytic transition state. *J. Mol. Biol.* 224 (1), 159–177. doi:10.1016/0022-2836(92)90582-5

Nakane, T., Kotecha, A., Sente, A., McMullan, G., Masiulis, S., Brown, P. M. G. E., et al. (2020). Single-particle cryo-EM at atomic resolution. *Nature* 587, 152–156. doi:10.1038/s41586-020-2829-0

Olsson, S., Wu, H., Paul, F., Clementi, C., and Noé, F. (2017). Combining experimental and simulation data of molecular processes via augmented Markov models. *Proc. Natl. Acad. Sci. U.S.A.* 114 (31), 8265–8270. doi:10.1073/pnas.170480311410.1073/pnas.1704803114

Onufriev, A., Bashford, D., and Case, D. A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* 55, 383–394. doi:10.1002/prot.20033

Onufriev, A. V., and Case, D. A. (2019). Generalized born implicit solvent models for biomolecules. *Annu. Rev. Biophys.* 48, 275–296. doi:10.1146/annurev-biophys-052118-115325

Orzechowski, M., and Tama, F. (2008). Flexible fitting of high-resolution x-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations. *Biophys. J.* 95 12, 5692–705. doi:10.1529/biophysj.108.139451

Peng, J., Yuan, C., Ma, R., and Zhang, Z. (2019). Backmapping from multiresolution coarse-grained models to atomic structures of large biomolecules by restrained molecular dynamics simulations using Bayesian inference. *J. Chem. Theory Comput.* 15, 3344–3353. doi:10.1021/acs.jctc.9b00062

Quiocho, F. A., Spurlino, J. C., and Rodseth, L. E. (1997). Extensive features of tight oligosaccharide binding revealed in high-resolution structures of the maltodextrin transport/chemosensory receptor. *Structure* 5 (8), 997–1015. doi:10.1016/s0969-2126(97)00253-0

Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential structure determination. *Science* 309, 303–306. doi:10.1126/science.1110428

Roseman, A. M. (2000). Docking structures of domains into maps from cryo-electron microscopy using local correlation. *Acta Crystallogr. D Biol. Crystallogr.* 56, 1332–1340. doi:10.1107/s0907444900010908

Rotkiewicz, P., and Skolnick, J. (2008). Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* 29, 1460–1465. doi:10.1002/jcc.20906

Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 23, 327–341. doi:10.1016/0021-9991(77)90098-5

Schaefer, M., Bartels, C., and Karplus, M. (1998). Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model. *J. Mol. Biol.* 284, 835–848. doi:10.1006/jmbi.1998.2172

Schlitter, J., Engels, M., and Krüger, P. (1994). Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *J. Mol. Graph* 12, 84–89. doi:10.1016/0263-7855(94)80072-3

Sharff, A. J., Rodseth, L. E., Spurlino, J. C., and Quiocho, F. A. (1992). Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of the maltodextrin binding protein involved in active transport and chemotaxis. *Biochemistry* 31 (44), 10657–10663. doi:10.1021/bi00159a003

Sugita, Y., Kitao, A., and Okamoto, Y. (2000). Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.* 113, 6042–6051. doi:10.1063/1.1308516

Taketomi, H., Ueda, Y., and Gō, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulation. I. the effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.* 7 (6), 445–459. doi:10.1111/j.1399-3011.1975.tb02465.x

Tama, F., Miyashita, O., and Brooks, C. L. (2004). Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J. Mol. Biol.* 337, 985–999. doi:10.1016/j.jmb.2004.01.048

Terashi, G., and Kihara, D. (2018). De novo main-chain modeling for em maps using MAINMAST. *Nat. Commun.* 9 (1), 1618. doi:10.1038/s41467-018-04053-7

Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., and Sali, A. (2008). Protein structure fitting and refinement guided by cryo-EM density. *Structure* 16, 295–307. doi:10.1016/j.str.2007.11.016

Touw, W. G., Baakman, C., Black, J., te Beek, T. A., Krieger, E., Joosten, R. P., et al. (2015). A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* 43, D364–D368. doi:10.1093/nar/gku1028 Database issue

Toyoshima, C. (2008). Structural aspects of ion pumping by Ca2+-ATPase of sarcoplasmic reticulum. *Arch. Biochem. Biophys.* 476, 3–11. doi:10.1016/j.abb.2008.04.017

Toyoshima, C., and Mizutani, T. (2004). Crystal structure of the calcium pump with a bound ATP analogue. *Nature* 430, 529–535. doi:10.1038/nature02680

Toyoshima, C., and Nomura, H (2002). Structural changes in the calcium pump accompanying the dissociation of calcium. *Nature* 418, 605–611. doi:10.1038/nature00944

Toyoshima, C., Nakasako, M., Nomura, H., and Ogawa, H. (2000). Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature* 405, 647–655. doi:10.1038/35015017

Toyoshima, C., Norimatsu, Y., Iwasawa, S., Tsuda, T., and Ogawa, H. (2007). How processing of aspartylphosphate is coupled to lumenal gating of the ion pathway in the calcium pump. *Proc. Natl. Acad. Sci.* 104, 19831–19836. doi:10.1073/pnas.0709978104

Trabuco, L. G., Villa, E., Mitra, K., Frank, J., and Schulten, K. (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* 16 (5), 673–683. doi:10.1016/j.str.2008.03.005

Villa, E., and Lasker, K. (2014). Finding the right fit: chiseling structures out of cryo-electron microscopy maps. *Curr. Opin. Struct. Biol.* 25, 118–125. doi:10.1016/j.sbi.2014.04.001

Wang, R. Y.-R., Kudryashev, M., Li, X., Egelman, E. H., Basler, M., Cheng, Y., et al. (2015). De novo protein structure determination from near-atomic-resolution cryo-EM maps. *Nat. Methods* 12, 335–338. doi:10.1038/nmeth.3287

Whitford, P. C., Ahmed, A., Yu, Y., Hennelly, S. P., Tama, F., Spahn, C. M. T., et al. (2011). Excited states of ribosome translocation revealed through integrative molecular modeling. *Proc. Natl. Acad. Sci. U.S.A.* 108, 18943–18948. doi:10.1073/pnas.1108363108

Whitford, P. C., Geggier, P., Altman, R. B., Blanchard, S. C., Onuchic, J. N., and Sanbonmatsu, K. Y. (2010). Accommodation of aminoacyl-tRNA into the ribosome involves reversible excursions along multiple pathways. *RNA* 16, 1196–1204. doi:10.1261/rna.2035410

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., et al. (2018). MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci.* 27, 293–315. doi:10.1002/pro.3330

Wriggers, W., Milligan, R. A., and McCammon, J. A. (1999). Situs: a package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.* 125, 185–195. doi:10.1006/jsbi.1998.4080

Wu, M., and Lander, G. C. (2020). How low can we go? structure determination of small biological complexes using single-particle cryo-EM. *Curr. Opin. Struct. Biol.* 64, 9–16. doi:10.1016/j.sbi.2020.05.007

Yip, K. M., Fischer, N., Paknia, E., Chari, A., and Stark, H. (2020). Atomic-resolution protein structure determination by cryo-EM. *Nature* 587, 157–161. doi:10.1038/s41586-020-2833-4

Zhou, R. (2003). Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins* 53, 148–161. doi:10.1002/prot.10483

Zivanov, J., Nakane, T., Forsberg, B. O., Kimanius, D., Hagen, W. J., Lindahl, E., et al. (2018). New tools for automated high-resolution cryo-EM structure determination in RELION-3. *elife* 7, e42166. doi:10.7554/eLife.42166

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 © 2021 Kulik, Mori and Sugita. 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 Molecular Biosciences

BRIEF RESEARCH REPORT
published: 25 March 2021
doi: 10.3389/fmolb.2021.644644

Molecular Basis of Class B GPCR Selectivity for the Neuropeptides PACAP and VIP

Chenyi Liao[1,2], *Jacob M. Remington*[1], *Victor May*[3] *and Jianing Li*[1]*

[1]Department of Chemistry, University of Vermont, Burlington, VT, United States, [2]State Key Laboratory of Molecular Reaction Dynamics, Dalian Institute of Chemical Physics, Dalian, China, [3]Department of Neuroscience, University of Vermont, Burlington, VT, United States

The related neuropeptides PACAP and VIP, and their shared PAC1, VPAC1 and VPAC2 receptors, regulate a large array of physiological activities in the central and peripheral nervous systems. However, the lack of comparative and molecular mechanistic investigations hinder further understanding of their preferred binding selectivity and function. PACAP and VIP have comparable affinity at the VPAC1 and VPAC2 receptor, but PACAP is 400–1,000 fold more potent than VIP at the PAC1 receptor. A molecular understanding of the differing neuropeptide-receptor interactions and the details underlying the receptor transitions leading to receptor activation are much needed for the rational design of selective ligands. To these ends, we have combined structural information and advanced simulation techniques to study PACAP/VIP binding selectivity, full-length receptor conformation ensembles and transitions of the PACAP/VIP receptor variants and subtypes, and a few key interactions in the orthosteric-binding pocket. Our results reveal differential peptide-receptor interactions (at the atomistic detail) important for PAC1, VPAC1 and VPAC2 receptor ligand selectivity. Using microsecond-long molecular dynamics simulations and the Markov State Models, we have also identified diverse receptor conformational ensembles and microstate transition paths for each receptor, the potential mechanisms underlying receptor open and closed states, and the interactions and dynamics at the transmembrane orthosteric pocket for receptor activation. These analyses reveal important features in class B GPCR structure-dynamics-function relationships, which provide novel insights for structure-based drug discovery.

Keywords: G protein-coupled receptor, membrane protein, signaling, molecular dynamics, conformational transition, ligand selectivity

OPEN ACCESS

Edited by:
Joanna Trylska, University of Warsaw, Poland

Reviewed by:
Yinglong Miao, University of Kansas, United States
Shuguang Yuan, Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, China

***Correspondence:**
Jianing Li
jianing.li@uvm.edu

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *21 December 2020*
Accepted: *16 February 2021*
Published: *25 March 2021*

Citation:
Liao C, Remington JM, May V and Li J (2021) Molecular Basis of Class B GPCR Selectivity for the Neuropeptides PACAP and VIP. Front. Mol. Biosci. 8:644644. doi: 10.3389/fmolb.2021.644644

INTRODUCTION

Pituitary adenylate cyclase-activating peptide (PACAP, *ADCYAP1*, P18509) and vasoactive intestinal polypeptide (VIP, *VIP*, P01282) are two important neuropeptides for neural development, body calcium homeostasis, glucose metabolism, circadian rhythm, thermoregulation, inflammation, feeding behavior, pain modulation, stress and related endocrine response (Harmar et al., 2012; Bortolato et al., 2014; Culhane et al., 2015; Liao et al., 2019b; Liao et al., 2019c). Their important roles are mediated by activating three class B G protein-coupled receptors (GPCRs) including the PAC1 (*ADCYAP1R1* and related splice variants), VPAC1 (*VIPR1)*, and VPAC2 (*VIPR2)* receptors, which share over 60% sequence similarity within this GPCR subtype.

Interestingly, PACAP and VIP display distinct selectivity for the PAC1 and VPAC1/2 receptors (Vaudry et al., 2009): PACAP (PACAP38 and the C-terminally truncated PACAP27) and VIP have comparable high affinity for the VPAC1 and VPAC2 receptors (Gottschall et al., 1990; Lam et al., 1990), whereas the PACAP peptides are 400–1,000 fold more potent than VIP as agonists for PAC1 receptor (Gottschall et al., 1990; Dautzenberg et al., 1999).

From recent progress in X-ray crystallography and cryogenic electron microscopy (cryo-EM) (Hollenstein et al., 2013; Yang et al., 2015; Jazayeri et al., 2017; Zhang et al., 2017a; Zhang et al., 2017b; Zhang et al., 2018; Duan et al., 2020; Kobayashi et al., 2020; Liang et al., 2020; Wang et al., 2020), several class B receptor structures have been established. We have now integrated the structural information with simulations, theory and available experimental findings to examine: *1)* differences in intrinsic PAC1 and VPAC1/2 receptor dynamics; *2)* differential PACAP and VIP interactions and energetic impacts at the receptor subtypes; and *3)* the interactions and dynamical features of PACAP-induced activation of the PAC1null receptor. Class B GPCRs have been suggested to have receptor features and dynamics distinct from those described for class A receptors, presumably because of the presence of the extracellular domain (ECD) (Liao et al., 2017), the apparent absence of "toggle switches" (Li et al., 2013), and other signature features. The class B receptors are likely to adopt unique conformational changes in the heptahelical transmembrane (7TM) domain following neuropeptide binding for receptor activation. Different from some class B receptors (Hollenstein et al., 2014; Yang et al., 2015; Graaf et al., 2016; Jazayeri et al., 2016; Jazayeri et al., 2017; Song et al., 2017; Zhang et al., 2017a; Zhang et al., 2017b; Liang et al., 2018), the detailed and comparative examination of the PAC1 and VPAC1/2 receptor function and actions have not been well investigated. Accordingly, we have integrated long-timescale molecular dynamics (MD) simulations with other computational and theoretical techniques to study the conformational ensembles and transitions of two PAC1 receptor variants (namely PAC1null and PAC1s) (Liao et al., 2019c), VPAC1, and VPAC2 receptors, as well as their interactions with the neuropeptides PACAP and VIP. The PAC1null receptor has no intracellular loop 3 (ICL3) inserts (neither hip nor hop inserts) from alternative splicing, and represents one of the most prevalent receptor isoforms in the central nervous system (CNS). The PAC1s receptor is similar to the PAC1null variants, but has a 21-amino acid deletion (residues 89–109) in the ECD. For reasons unknown, the PAC1s receptor is not highly expressed in the CNS (or any tissue) and yet this variant was chosen for structural determination (Sun et al., 2007; Kumar et al., 2011; Wang et al., 2020). Distinct from structural determination, our simulations and analyses detail the receptor conformational transitions and peptide interactions that are biologically relevant to the receptor subtype. While the PAC1 receptor has been recognized as an emerging target for stress-related disorders (Ressler et al., 2011), our studies, for the first time, reveal the molecular basis underlying PACAP receptor selectivity which may be essential for the rational design of effective therapeutic agents.

MODELS AND METHODS

Model Preparation and MD Simulation Setup

The ligand-free and ligand-binding full-length receptor systems. Our homology modeling was carried out with the protein modeling program *Prime* (Schrödinger, *Inc.*) (Zhao et al., 2011; Li et al., 2011) which built the full-length VPAC1 and VPAC2 including the ECD, the 7TM, as well as the linker loop (see details in **Supplementary Table S2**). The PAC1s model was modified from our previous PAC1null receptor model (Liao et al., 2017) by deleting the 21-amino acid in ECD. For each receptor, we generated a number of initial models which were distinct in the ECD orientation to better sample different conformational states (**Supplementary Figure S1**). The Membrane Builder in CHARMM-GUI (Jo et al., 2008) was employed to build the protein/membrane complex systems, each of which contains a receptor with/without a ligand, a lipid bilayer of ~ 225 POPC molecules, ~ 27,000 TIP3P water molecules, counter ions, and 0.12 M NaCl, totaling ~ 120,000 atoms in a periodic box ~$94 \times 94 \times 147 Å^3$.

The ligand-ECD complex systems. To understand the distinct binding affinity of PACAP/VIP for PAC1null, PAC1s, VPAC1 and VPAC2 receptors, we simulated the ECD of each receptor and PACAP/VIP in a solvent environment. PACAP/VIP was initially placed near the position based on known peptide-binding modes to class B GCPRs (Sun et al., 2007; Grace et al., 2010; Parthier et al., 2007; Pioszak and Xu, 2008) (**Supplementary Figure S2**). We used the PDB Reader of CHARMM-GUI (Jo et al., 2008) to prepare protein then solvated and neutralized with addition of 0.12 M NaCl in VMD (Humphrey et al., 1996) (see our simulation summary in **Supplementary Table S1**).

Our simulations were performed with the CHARMM36-cmap force field (Best et al., 2012). Each system went through energy minimization, 285-ns multiple-step equilibrations, and MD simulations using the NAMD package (Phillips et al., 2005). Both equilibration and production runs were performed in the NPT ensemble (310K, 1 bar, Langevin thermostat and Nose-Hoover Langevin barostat) with a time step of 2 fs. All of the bond lengths to the hydrogen atoms in the protein or lipid molecules were constrained. The particle mesh Ewald (PME) technique was used for the electrostatic calculations. The van der Waals and short-range electrostatics were cut off at 12.0 Å with switch at 10.0 Å. Each system was carried out 1000 ns for 4–5 replicas.

Enhanced sampling for ligand-binding simulations. Adaptive tempering was employed to enhance conformational sampling under PACAP or PACAP6-38 insertion for ~ 400 ns, in which the simulation temperature was dynamically updated as a continuously random variable in the range of 310–360 K with the Langevin equation (Zhang and Ma, 2010) implemented in NAMD (**Supplementary Figure S3**).

MARKOV STATE MODEL

Markov state models (MSMs) of molecular kinetics were applied on the collection of MD trajectories, from which we identified the transition pathways and estimated the long-time statistical dynamics between the conformational states of interest (Pande et al., 2010; Prinz et al., 2011). The ligand-free full-length receptor MD trajectories were used for MSM construction. We used the MSMBuilder 3.8.0 program Bowman et al., 2009a; Bowman et al., 2009b to construct the transition matrices of MSM. Trajectories that contain stable conformational states (measured by C_α root-mean-square deviation (RMSD) in **Supplementary Figure SM2**) were prepared by saving the C_α coordinates of a GPCR protein with atomic indices. Thus, we have 4,519–14,281 conformations in each trajectory. We grouped the conformational ensembles into a set of clusters, called microstates, based on C_α RMSD using the *k*-centers algorithm with 33–55 cluster centers Bowman et al., 2009b.

With the microstate discretization, we used the maximum likelihood estimation to build the reversible transition matrices at the lag time series. The evolution of the transition probability between closed and open conformational states of a receptor protein was evaluated at a series of lag times τ, 2τ, …, $n\tau$ by performing the *Chapman-Kolmogorov* test (Prinz et al., 2011) which confirmed that our model at the chosen lag time is Markovian (Chodera et al., 2007; Noé et al., 2009; Prinz et al., 2011) (**Supplementary Figures SM4,SM5**). Finally, this was achieved at a lag time of 1.2–2.4 ns as judged by flat lines in the *implied timescales* plots (**Supplementary Figure SM3**) (see Ref. Liao et al., 2019a for our detailed implementation of the *Chapman-Kolmogorov* test).

Using the transition-path theory (TPT) (Weinan and Vanden-Eijnden, 2006; Berezhkovskii et al., 2009; Metzner et al., 2009; Noé et al., 2009), we calculated the transition path connecting between the stable conformational states in a transition matrix and estimated the time quantity to travel from one set of states to the others by the lag time divided by the minimum net flux in the transition path (**Supplementary Table S3**). Thus, the transition time refers to the time quantity to complete one transition between two stable conformational states. The division of macrostates were calculated from the eigenfunction structure using the *Perron Cluster Cluster Analysis* (PCCA) method (Noé et al., 2007). **Supplementary Figure SM6** illustrates the 3D projections of the microstates and macrostate divisions in accordance to ECD orientation and position by backbone center-of-mass (COM) distances between ECD and extracellular loop 1 (ECL1) ($d_{ECD\text{-}ECL1}$) and extracellular loop 3 (ECL3) ($d_{ECD\text{-}ECL3}$), respectively. Detailed constructions and validation of MSM were described in the **Supplementary Material**.

Data Analysis

Binding free energy (ΔG_b) of PACAP/VIP peptide bound to the ECD and per-residue free energy decomposition were computed over the last 500 ns MD trajectories using the Molecular Mechanics Generalized Born Surface Area (MM-GBSA) method (Miller et al., 2012) by MMPBSA.py tool (version: 14.0) in the Amber package (Case, 2018). The second modified Bondi radii set were used in the GB method. ΔG_b was given as an average of five replicas in each system. Conformational analysis and distance analysis were performed with TCL scripts implemented in VMD 1.9.1 (Humphrey et al., 1996) and plotted by matplotlib. The average distances involving a group of residues were calculated by the distance between the backbone COMs of groups of residues. Water density map were calculated using MDAnalysis python package. The tilt angle of ECD is defined as the angle between the vector along the N-terminal helix and the Z-axis. For example, the vector along the N-terminal helix of PAC1 ECD is calculated by summing the C-O vectors along the helical residues 30–47.

RESULTS AND DISCUSSION

Conformation Ensembles and Transitions of PAC1 and VPAC1/2 Receptors

Building on our previous study of PAC1null receptor (Liao et al., 2017), we first examined the conformations of full-length PAC1null, PAC1s, VPAC1, and VPAC2 receptors, as well as the transitions within the conformational ensembles, which are key in understanding the differential actions of these receptors beyond amino acid sequences. With MD simulations totaling 20 microseconds to sample different conformations, we constructed MSM transition pathways between conformational microstates for PAC1s and VPAC1/2 receptors to compare with those shown for the PAC1null receptor described previously (Liao et al., 2017) (**Figure 1**). Similar to the PAC1null receptor, the ligand-free PAC1s and VPAC1/2 receptors display a wide diversity of conformational states with respect to the ECD orientation (θ) and ECD-7TM backbone COM distance (d), which reflect in part the actions of the linker region that tethers the ECD to the 7TM. Free energy maps based on the conformation density are shown in **Supplementary Figure S4**, where the low energy regions correspond to the stable conformational states in **Figure 1** are circled with the same color. The energy barriers between the stable conformational states vary from 0.5 to 3.5 kcal/mol. We define the closed (θ > 80 degree, d < 55 Å), semi-open (θ = ~60 degree, d < 55 Å), and open states (θ = ~40 degree, d < 55 Å) to distinguish these conformational states. ECD-unrestricted conformations refer to those with ECD not touching the 7TM (d > 55 Å), which include most intermediate states (colored in yellow in **Figure 1**), as well as some stable conformational states in systems of VPAC1 and VPAC2, e.g., the state in red in **Figure 1C** and the state in green in **Figure 1D**. Without binding to 7TM, they are capable to change to other stable states much faster. The conformational states in red and orange in **Figure 1D** that have the ECD-ECL1 backbone COM distance ~10 Å shorter than ECD-ECL3 backbone COM distance are labeled as the on-side conformations.

For the PAC1null receptor, we previously showed that the 21-amino acid loop in ECD interacts with ECL3 to potentially sustain the open time of the receptor (Liao et al., 2017). Consistently in this work, with the deletion of the 21 amino acids in PAC1s, the receptor microstate transitions between the open conformational state and closed conformational states (**Figure 1A**) are faster by a factor of 2 (transition time up to

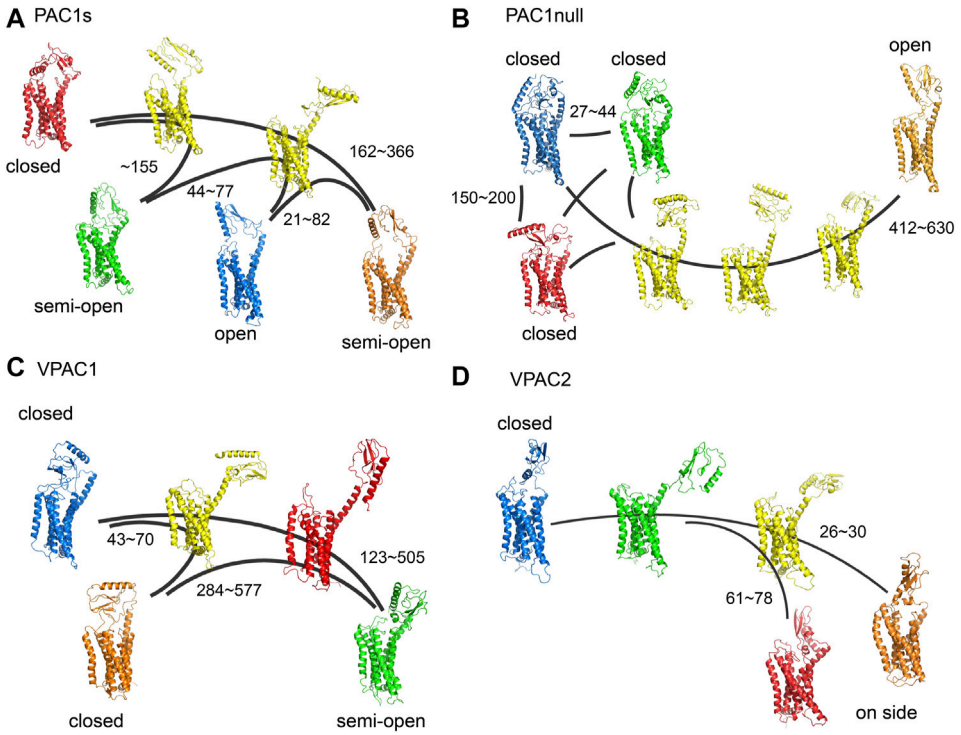


FIGURE 1 | (A) Transition pathways for **(A)** PAC1s, **(B)** PAC1null, **(C)** VPAC1 and **(D)** VPAC2 receptors. Stable conformational states are colored in blue, red, green, and orange respectively. Intermediate states are colored in yellow. Most intermediate states are ECD-unrestricted conformations. The VPAC1 conformation in red and the VPAC2 conformation in green also belong to the ECD-unrestricted conformational ensemble. PAC1null is remade from our recent study (Liao et al., 2017). The kinetic transition time between certain conformational states are labeled with units in nanoseconds.

400 μs for PAC1s and up to 1,000 μs for PAC1null, **Supplementary Table S3**). The free energy map also exhibits ~3 kcal/mol higher energy barrier between the open and closed states in PAC1null compared with it in PAC1s (**Supplementary Figure S4**). While the functional implications of these differences remain to be investigated, the open state of PAC1null, once formed, may be sustained longer than that for PAC1s to facilitate peptide trapping and PAC1null receptor activation. Overall, the conformational microstate transitions for the VPAC2 receptor appear the fastest (0.8–80 μs) compared to the VPAC1 (40–600 μs) or the PAC1 receptor variants (21–366 μs), in agreement with the free energy difference between states in **Supplementary Figure S4**. Therefore, despite significant primary amino acid sequence homology, the different PAC1 and VPAC receptors display unique intrinsic ensemble dynamics and conformational microstate transition times.

Several of the ligand-free receptor conformations resemble the peptide-bound receptor states identified in the structural solutions with respect to ECD orientation (θ) and ECD-7TM (d) or ECD-ECL backbone COM distances. For example, both the PAC1null and PAC1s receptors could adopt stable open conformations (θ = ~40°, 45 < d < 55Å), which have been seen in the peptide-bound PAC1 (PDBID: 6M1I) and other Class B receptor structures, including the glucagon receptor (GCGR, PDBID: 5YQZ), parathyroid hormone 1 receptor (PTH1, PDBID: 6FJ3), and glucagon-like peptide-1 receptor (GLP-1, PDBID: 5VAI) with minimum RMSD of 4.2Å (**Supplementary Figure S5**). In addition, the "semi-open" conformations of the VPAC1 receptor are similar to the CGRP-bound calcitonin receptor-like receptor (CLR) structure (PDBID: 6E3Y, θ = ~60°, d < 55Å) with RMSD of 5.4 Å. Moreover, the VPAC2 receptor "on-side" conformations with ECD in contact with ECL1 share similar features (θ < 80°, short ECD-ECL1 distance) with an antagonist-bound GCGR structure (PDBID: 5XEZ) with minimum RMSD of 4.6 Å. Beyond the relative orientation and position of ECD to 7TM, the 7TM conformations in the ligand-free receptor approximated those for the inactive GPCR structures (with lowest RMSD = 2.0–2.5Å in 7TM) than the peptide-bound active receptors (**Supplementary Figures S6,S7**), consistent with the actions of peptides to initiate transmembrane dynamics for receptor activation. In aggregate, these findings suggest that the PAC1 and VPAC1/2 receptors, even without peptide binding, have dynamic transitions between the open and closed conformational states, which prepare them for ligand binding and subsequent receptor activation/deactivation.

Distinct Affinity of PACAP and VIP to the Receptor ECDs

For Class B receptors, the ECD acts as a high-affinity peptide trap and subsequent peptide-ECD dynamics result in the presentation

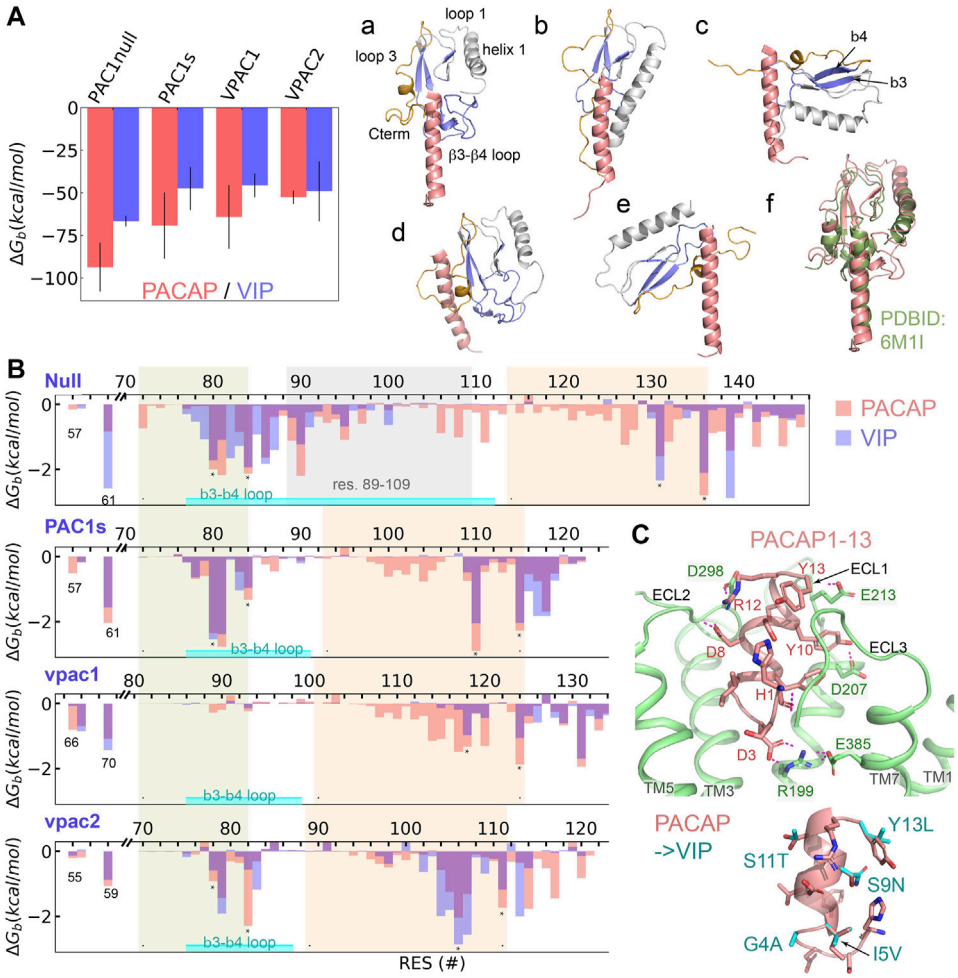


FIGURE 2 | Binding selectivity and recognition of PACAP and VIP on PAC1null, PAC1s, VPAC1 and VPAC2 receptors. **(A)** Binding free energy (ΔG_b) of PACAP and VIP on PAC1null, PAC1s, VPAC1 and VPAC2 receptors respectively by averaging the ΔG_b values from five replicas with bars representing the standard deviation. Representative binding conformations are displayed on the right, where residues to calculate free energy decomposition in **(B)** are colored in blue; only residues 1–27 are displayed for PACAP and VIP. **(B)** ECD per-residue free energy decomposition (average of five replicas) in PAC1 and VPAC1/2 binding PACAP (red column) and VIP (blue column). β3–β4 including the loop between is labeled. The 21 amino acids (residues 89–109) are labeled in PAC1null. **(C)** Close view of the N-terminal residues 1–13 of PACAP interacting with transmembrane orthosteric pocket of PAC1 receptor. Hydrogen bonds are shown in dashed lines. Comparison of PACAP with VIP at residues 1–13 by direct mutations is shown at the bottom right.

of the peptide N-terminus to the orthosteric binding site in the 7TM to initiate receptor activation. Our MSM analyses suggest the dynamic processes (peptide binding, large-scale ECD rotation, and peptide insertion) occur on a time scale greater than tens of microseconds. Thus, we applied a total of 40 microseconds MD simulations on the ligand-ECD systems; with the ECD-peptide complex trajectories we employed the MM-GBSA method (Miller et al., 2012) to estimate the binding free energy (ΔG_b) of PACAP or VIP to the ECD of PAC1null, PAC1s, VPAC1 and VPAC2 receptors, as well as the ΔG_b contribution per residue. Our results (**Figure 2A**) show that PACAP binding to PAC1null receptor is the most favorable among receptor and peptide combinations. Specifically, ΔG_b of PACAP is around 1.4 times of ΔG_b of VIP in binding PAC1null ECD; the difference decreases in PAC1s and VPAC1, and become comparable for PACAP and VIP binding VPAC2 ECD (**Figure 2A**). This provides an energetic basis for PAC1 receptor binding selectivity for PACAP ($K_d \approx 0.5$ nM) over VIP (K_d >500 nM) (Vaudry et al., 2009), which corresponds to that PACAP-binding free energy (−13.2 kcal/mol with K_d converted by $\Delta G_b = RT\ln K_d$, T = 310 K) is about 1.5 times of VIP (−8.9 kcal/mol) in binding PAC1null, while VPAC1 and VPAC2 receptors appear to exhibit comparable binding affinity ($K_d \approx 1$ nM) for PACAP and VIP (Vaudry et al., 2009). Notably, the experimental binding energies were obtained with full-length receptor systems, and the interactions between the ligand and 7TM may also affect the magnitude of the ΔG_b values for comparison with our calculation estimations. Therefore, here we compare the relative values of ΔG_b for the different systems as opposed to the absolute magnitude which differs from prior experiments.

To examine the ligand-ECD complex conformations, we performed structural clustering on the last 200- ns trajectories and identified six complex conformational states based on ECD orientation and the bound peptide in a vertical pose for receptor activation (**Figure 2A**; a-e). State (a) is found in PACAP binding PAC1null, PAC1s and VPAC2 receptors, which was resolved in the PAC1 receptor cyro-EM structure (PDBID: 6M1I, **Figure 2** state(f)). State (b) in which the ECD tile angle is decreased compared to state (a) is found prevalently in PACAP/VIP binding the VPAC1 receptor. While the PAC1s receptor adopts state (a) upon PACAP binding, it establishes state (c) if bound to VIP. Similarly, the VPAC2 receptor adopts state (a) in PACAP binding, but it alters to state (e) with VIP, which is comparable to the bound ligand bound CRF1 conformation (PDBID: 2L27).

According to previous theoretical and experimental findings (Vaudry et al., 2009; Wang et al., 2020), we examined the interacting residues between the receptor ECD (**Figure 2B**) and the C-terminal region of PACAP/VIP (amino acid residues 13–26). The PACAP and VIP C-terminal hydrophobic $V^{19}xxY^{22}LxxV/I^{26}$ sequence appears to contribute to the hydrophobic interactions with PAC1null/PAC1s and VPAC1/2 receptor residues $I61^{PAC}/L70^{VPAC1}/I59^{VPAC2}$; the PACAP/VIP $R^{14}K^{15}$ residues contribute to hydrogen bonding with the ECD (**Supplementary Figure S10**). The β3-β4 loop in both PAC1null/PAC1s and VPAC2 receptor (environs $L^{80}FxxF^{84,PAC}$ and $V^{78}FxxF^{82,VPAC2}$ respectively) interacts with PACAP/VIP (**Figure 2**, state (a)). In contrast, the ECD helix 1 rather than the β3-β4 loop in VPAC1 receptor interacts with the peptide (**Figure 2**, state (b)). In the different receptor ECD C-terminal region, segments $F^{131}xxxxF^{136,PAC1null}$, $F^{110}xxxxF^{115,PAC1s}$, $Y^{118}xxxxxL^{124,VPAC1}$, and $V^{106}xxxxY^{111,VPAC2}$ also contribute to ligand-receptor hydrophobic interactions.

For the PAC1null receptor, there are a few notable differences in PACAP and VIP binding. Firstly, both states (a) and (d) in **Figure 2** are identified in PAC1null-PACAP binding; state (a) is similar to the cyro-EM structure (Wang et al., 2020) and state (d) is close to the Nuclear Magnetic Resonance (NMR) model (Sun et al., 2007). PAC1null receptor VIP binding can also be in state (a). Secondly, PAC1null receptor hydrophobic residues $72–140^{PAC1null}$ play major role in PACAP binding, such as M^{72}, $L^{80}FxIF^{84}xP^{86}$, VW^{90}, I^{95}, $L^{106}xLxxM^{111}$, F^{127}, F^{131}, F^{136}, and Y^{139} (**Figure 2B**); only E^{142} and D^{145} at the ECD C-terminal region contributes significantly to hydrogen bonding with the ligand. Different from PACAP, VIP can also interact with charged/polar residues $78–87^{PAC1null}$ such as E^{79}, R^{82}, N^{85} and D^{87}. Lastly, PACAP interacts with residues $92–130^{PAC1null}$ extensively, while VIP interacts with few. This also supports the finding that another short variant with deletion of residues $53–109^{PAC1null}$ significantly reduces VIP binding affinity over PACAP (Dautzenberg et al., 1999). These remarkable interaction differences together contribute to the binding preference for PACAP. Compared with PAC1null, PAC1s missing the 21-amino acid loop loses the interaction formed by residues 86–111 of PAC1null which results in decrease in ligand-ECD binding affinity.

Integrating our finding on ligand-ECD interactions, PAC1 receptor cyro-EM structure (Wang et al., 2020) and biochemical data, the C-terminal region of PACAP/VIP plays a major role in binding with ECD, while the peptide N-terminus interacts with the receptor ECLs and 7TM activation site. Based on alignment with the cyro-EM structure (Wang et al., 2020), we simulated $PACAP_{1-13}$ insertion in PAC1 7TM, without PACAP C-terminal residues 13–26 associations to the ECD. Within a few hundred nanoseconds, PACAP residues 1–4 rearranges to form a $D3^{PACAP}$-$R199^{PAC1}$ interacting pair (**Figure 2C**). The depth of peptide N-terminus into the 7TM is not significantly changed, but $H1^{PACAP}$ reorients to form hydrogen bond with ECL3, while PACAP residues D8, Y10, R12 and Y13 interact with ECLs (**Figure 2C**). These interactions result in outward conformational transition of TM6 at the intracellular face of receptor. However, in longer simulations, PACAP1-13 is also observed to shift away from its primary position toward the V-gap between TM5 and TM6, allowing TM6 to transition back to former inactive conformational state. Thus, these observations show the importance of PACAP C-terminus-ECD binding to constraint the peptide N-terminus within the 7TM orthosteric interaction site and sustain the activation process.

PACAP-Induced Activation of PAC1null Receptor

To further understand how PACAP interacts with the full-length receptor, we simulated the PAC1null receptor (a homology model of the inactive state (Siu et al., 2013; Liao et al., 2017)) bound to the $PACAP_{1-38}$ peptide, with the first 13 residues partially inserted into the receptor 7TM. We use the Wootten numbering (Wootten et al., 2013) in this section to better demonstrate the residue positions in PAC1null receptor. Given the timescale of PAC1R activation, we combined microsecond scale MD simulation and adaptive tempering (Zhang and Ma, 2010) to enhance conformational sampling under these conditions. The N-terminus of PACAP underwent conformational rearrangement in the orthosteric site with a clear TM6 outward movement of ~ 4 Å with the TM6-TM3 distance close to the relaxed GCGR and GLP-1 receptor structures (**Supplementary Figures S11,S12**), indicating an "active" state.

A detailed examination suggests that the formation of this state results from several key interactions. As shown in **Figure 3B** (top panel), $D3^{PACAP}$ and PAC1null $R199^{2.60}$ near the receptor extracellular face appears to migrate closer together in the first 700 ns, but becomes entrapped in a separate conformation for another 1500 ns. During adaptive tempering, the average distance between $H1^{PACAP}$ and the surrounding receptor $K206/D207^{ECL1}$, $E374/E380/E385^{ECL3}$, and $K154^{1.40}$ decreases gradually and appears to form hydrogen bond (H-bond) networks with an average small distance of ~5 Å. The hydrogen bonding between the $D3^{PACAP}$ sidechain and $H1^{PACAP}$ amino terminus weakens during adaptive tempering resulting in the ability for $D3^{PACAP}$ to form a new pairing with receptor $R199^{2.60}$. As a result of $D3^{PACAP}$-$R199^{2.60}$ interactions, the average distance between $R199^{2.60}$ and $N240^{3.43}$, $H365^{6.52}$ and $Q392^{7.49}$ increases

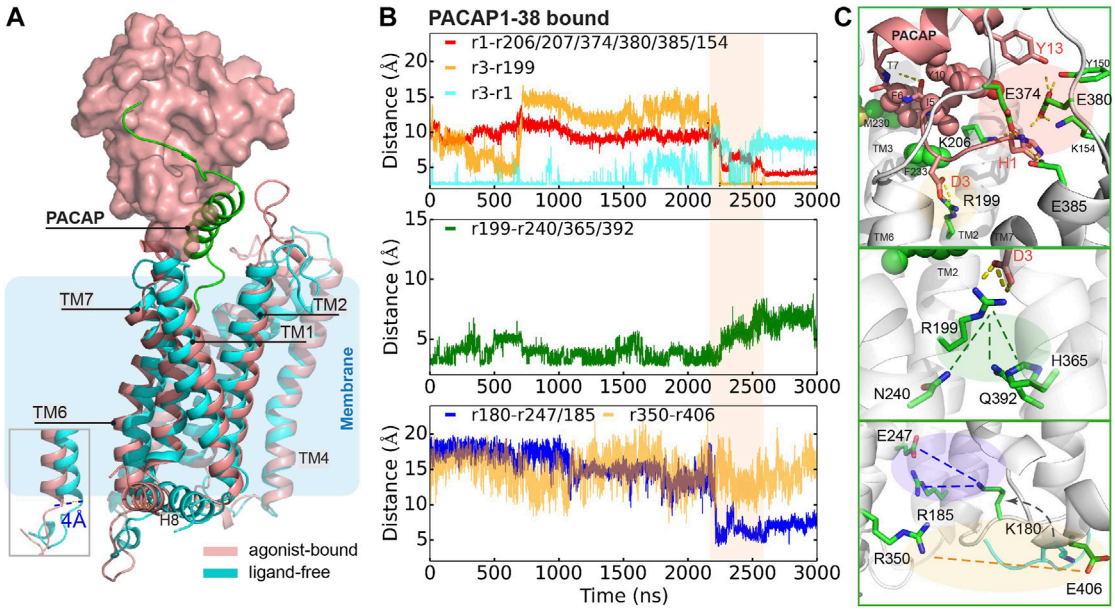


FIGURE 3 | Representative conformation of PACAP-activated PAC1null receptor. **(A)** Structural alignment of PACAP-activated (red) and ligand-free (cyan) conformations. Ligand-free conformation is from our recent study(Liao et al., 2017) with only the 7TM shown. PACAP is shown in green cartoon and ECD is shown in surface representation. **(B)** Time evolution of average distance or pair distance of key charged/polar residues locating at the extracellular side, middle transmembrane, and intracellular side of the PAC1null receptor, which are displayed in **(C)** accordingly. Adaptive tempering period is in light orange strip background.

significantly (**Figure 3B**, middle panel), indicating weakened contacts between TM2, TM3, TM6, and TM7 (**Figure 4D**). At the receptor intracellular face, $K180^{ICL1}$ transitions from a solvent orientation to face E247/R185, and decreasing $K180^{ICL1}$-$E247^{3.50}$/$R185^{ICL1}$ distance (**Figure 3B**, bottom panel) converts the PAC1null receptor from an apparent "inactive" to "active" conformation that allows G protein interactions. Direct mutagenesis on residues $4–13^{PACAP}$ (i.e., N-terminal G4A, I5V, S9N, S11T, Y13L found in VIP sequence) in the "active" conformation caused the revocation of the spreading H-bond network around H1 and ECLs and led to inward movement of TM6 (**Supplementary Figure S12**) and stronger TM2, TM3, TM6 and TM7 interactions (**Figure 4D**).

To investigate the effects of the first few residues at the N-terminal, we removed PACAP residues 1-5 from the PACAP-bound PAC1null complex, to simulate the effects of the PAC1 receptor antagonist PACAP6-38. Starting from an active conformation, we employed 45 ns adaptive tempering to accelerate the process, as conventional MD simulations show that longer times may be needed to relax an "active" conformational state (**Supplementary Figure 12B**). As shown in **Figures 4A,B** a 4~6 Å inward movement of TM6 was observed, indicating the closing of the intracellular G protein binding site into an inactive state. Compared to the "active" conformation, $E374^{ECL3}$ is released and shifted to face the solvent. A return of enhanced TM2, TM3, TM6 and TM7 interactions are subsequently observed (**Figure 4D**), resulting in decreased average distance between $R199^{2.60}$ and $N240^{3.43}$, $H365^{6.52}$ and $Q392^{7.49}$, and the distance between the $R350^{ICL3}$-$E406^{H8}$ pair (**Figure 4A**).

Key interactions between the N-terminus of PACAP or PACAP6-38 and PAC1null receptor is summarized in **Figure 4C** comparing with residue interactions in the ligand-free conformations from our previous study (Liao et al., 2017). Compared with the recent PAC1 receptor cryo-EM structure (Wang et al., 2020), key contact like $D3^{PACAP}$-$R199^{2.60}$ was also found in our simulations, although the N-terminus of PACAP underwent conformational rearrangements in our simulations. Together, the C-$term^{ligand}$-ECD and N-$term^{ligand}$-7TM interactions present the unique structure features of class B GPCR in activation: ligand-ECD binding initially constraints the peptide close to 7TM; under optimized binding state, interactions between residues 8–15 of PACAP/VIP and ECLs help dock the N-terminus of the peptide into the pocket and initiate activation (Schäfer et al., 1999).

CONCLUSION

In summary, we have used advanced simulation and modeling techniques combined with prior experimental knowledge to examine the differential binding selectivity, conformation ensembles and microstate dynamic transitions among the PACAP/VIP receptor variants and subtypes, and a few key interactions in the orthosteric-binding pocket. The PAC1null, PAC1s, VPAC1 and VPAC2 receptors demonstrate unique conformation and transition states. Expectedly, from residue interactions and free energy calculations, PACAP C-terminal binding to PAC1null receptor ECD is favored over VIP. The PAC1s receptor variant harboring a 21-amino acid ECD deletion

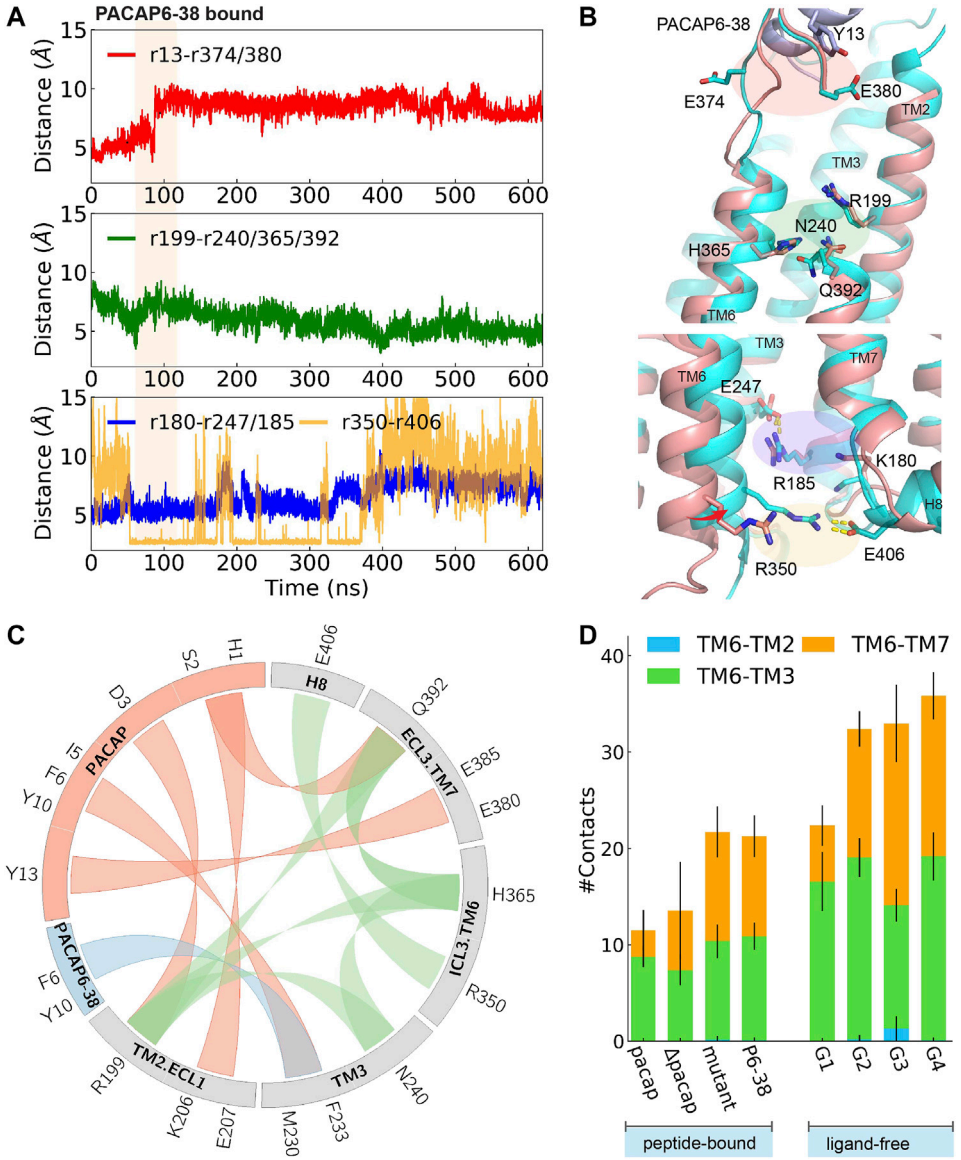


FIGURE 4 | Time evolution of average distance or pair distance of key charged/polar residues locating at extracellular side, middle transmembrane, and intracellular side of PAC1null binding PACAP6-38, which are illustrated in **(B)** accordingly. Adaptive tempering period is in light orange strip background. **(B)** Final snapshot of PACAP6-38 bound PAC1null (cyan) superposed on PACAP-activated PAC1null structure (red), in which an inward shifting of TM6 was observed. **(C)** Wheel plot of major polar and hydrophobic interactions in the N-terminus of PACAP bound PAC1null (in red), N-terminus of PACAP6-38 bound PAC1null (in blue), and residue interactions within ligand-free PAC1null (in green). **(D)** Stack histogram of number of contacts between the intracellular half of TM6 (residues 344–360) and TM2, TM3, and TM7 in PAC1null receptor with: PACAP, PACAP deleted (ΔPACAP), PACAP6-38 (P6-38), and PACAP mutant (G4A, I5V, S9N, S11T, Y13L). Collections of the last 120 ns of each replica were used for the contact calculations for the PACAP, ΔPACAP, P6-38, and PACAP mutant systems. G1-G4 represent the four major ligand-free conformational states in **Figure 1B**. We used the last 960 ns of each 2- μs ligand-free system (Liao et al., 2017) for the contact calculations.

has been enigmatic because after identification from cloning, the variant appears absent or expressed at very low levels in all tissues examined to date; the PAC1null receptor by contrast is dominant. Our analyses show that the ECD 21-amino acid loop in the PAC1null receptor allows several functional advantages including potential enhancement of the receptor open state to facilitate peptide entrapment and increasing peptide binding affinity. The simulations reveal the receptor conformational changes upon ECD-bound peptide dynamics and peptide N-terminal presentation to the 7TM orthosteric activation site. Notably, the interaction of PACAP residues 8–15 to the ECLs is observed to help dock the peptide N-terminal into the orthosteric pocket and facilitate $D3^{PACAP}$–$R199^{PAC1null}$ interactions that result in 7TM transitions and TM6 dynamics to initiate activation. Interestingly, the ECD not only functions for high affinity binding but also appears to constraint the orientation of the peptide N-terminus in the orthosteric site to maintain the activation state. These

observations may be common to other Class B GPCRs and provide mechanistic insights for the development of therapeutics.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

CL and JR performed simulations and data analysis. All authors discussed and approved the contents of the manuscript and contributed to its final version by reading and editing.

FUNDING

The work was supported by grants R01-GM129431 (JL) from the National Institutes of Health.

ACKNOWLEDGMENTS

Computational resources were provided by VACC, PSC-ANTON, and XSEDE.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.644644/full#supplementary-material.

REFERENCES

Berezhkovskii, A., Hummer, G., and Szabo, A. (2009). Reactive flux and folding pathways in network models of coarse-grained protein dynamics. *J. Chem. Phys.* 130 (20), 205102. doi:10.1063/1.3139063

Best, R. B., Zhu, X., Shim, J., Lopes, P. E., Mittal, J., Feig, M., et al. (2012). Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ(1) and χ(2) dihedral angles. *J. Chem. Theor. Comput.* 8 (9), 3257–3273. doi:10.1021/ct300400x

Bortolato, A., Doré, A. S., Hollenstein, K., Tehan, B. G., Mason, J. S., and Marshall, F. H. (2014). Structure of class B GPCRs: new horizons for drug discovery. *Br. J. Pharmacol.* 171 (13), 3132–3145. doi:10.1111/bph.12689

Bowman, G. R., Beauchamp, K. A., Boxer, G., and Pande, V. S. (2009a). Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* 131 (12), 124101. doi:10.1063/1.3216567

Bowman, G. R., Huang, X., and Pande, V. S. (2009b). Using generalized ensemble simulations and markov state models to identify conformational states. *Methods* 49 (2), 197–201. doi:10.1016/j.ymeth.2009.04.013

Case, D. A. (2018) *Amber 2018*. (San Francisco: University of California).

Chodera, J. D., Singhal, N., Pande, V. S., Dill, K. A., and Swope, W. C. (2007). Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* 126, 155101. doi:10.1063/1.2714538

Culhane, K. J., Liu, Y., Cai, Y., and Yan, E. C. (2015). Transmembrane signal transduction by peptide hormones via family B G protein-coupled receptors. *Front. Pharmacol.* 6, 264. doi:10.3389/fphar.2015.00264

Dautzenberg, F. M., Mevenkamp, G., Wille, S., and Hauger, R. L. (1999). N-terminal splice variants of the type I PACAP receptor: isolation, characterization and ligand binding/selectivity determinants. *J. Neuroendocrinol* 11, 941–949. doi:10.1046/j.1365-2826.1999.00411.x

Duan, J., Shen, D. D., Zhou, X. E., Bi, P., Liu, Q. F., Tan, Y. X., et al. (2020). Cryo-EM structure of an activated VIP1 receptor-G protein complex revealed by a NanoBiT tethering strategy. *Nat. Commun.* 11, 4121. doi:10.1038/s41467-020-17933-8

Gottschall, P. E., Tatsuno, I., Miyata, A., and Arimura, A. (1990). Characterization and distribution of binding sites for the hypothalamic peptide, pituitary adenylate cyclase-activating polypeptide. *Endocrinol.* 127, 272–277. doi:10.1210/endo-127-1-272

Graaf, C., Donnelly, D., Wootten, D., Lau, J., Sexton, P. M., Miller, L. J., et al. (2016). Glucagon-like peptide-1 and its class B G protein-coupled receptors: a long march to therapeutic successes. *Pharmacol. Rev.* 68, 954–1013. doi:10.1124/pr.115.011395

Grace, C. R., Perrin, M. H., Gulyas, J., Rivier, J. E., Vale, W. W., and Riek, R. (2010). NMR structure of the first extracellular domain of corticotropin-releasing factor receptor 1 (ECD1-CRF-R1) complexed with a high affinity agonist. *J. Biol. Chem.* 285, 38580–38589. doi:10.1074/jbc.M110.121897

Harmar, A. J., Fahrenkrug, J., Gozes, I., Laburthe, M., May, V., Pisegna, J. R., et al. (2012). Pharmacology and functions of receptors for vasoactive intestinal peptide and pituitary adenylate cyclase-activating polypeptide: IUPHAR Review 1. *Br. J. Pharmacol.* 166, 4–17. doi:10.1111/j.1476-5381.2012.01871.x

Hollenstein, K., de Graaf, C., Bortolato, A., Wang, M. W., Marshall, F. H., and Stevens, R. C. (2014). Insights into the structure of class B GPCRs. *Trends Pharmacol. Sci.* 35, 12–22. doi:10.1016/j.tips.2013.11.001

Hollenstein, K., Kean, J., Bortolato, A., Cheng, R. K., Doré, A. S., Jazayeri, A., et al. (2013). Structure of class B GPCR corticotropin-releasing factor receptor. *Nature* 499, 438–443. doi:10.1038/nature12357

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

Jazayeri, A., Doré, A. S., Lamb, D., Krishnamurthy, H., Southall, S. M., Baig, A. H., et al. (2016). Extra-helical binding site of a glucagon receptor antagonist. *Nature* 533, 274–277. doi:10.1038/nature17414

Jazayeri, A., Rappas, M., Brown, A. J. H., Kean, J., Errey, J. C., Robertson, N. J., et al. (2017). Crystal structure of the GLP-1 receptor bound to a peptide agonist. *Nature* 546 (7657), 254–258. doi:10.1038/nature22800

Jo, S., Kim, T., Iyer, V. G., and Im, W. (2008). CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* 29 (11), 1859–1865. doi:10.1002/jcc.20945

Kobayashi, K., Shihoya, W., Nishizawa, T., Kadji, F. M. N., Aoki, J., Inoue, A., et al. (2020). Cryo-EM structure of the human PAC1 receptor coupled to an engineered heterotrimeric G protein. *Nat. Struct. Mol. Biol.* 27 (3), 274–280. doi:10.1038/s41594-020-0386-8

Kumar, S., Pioszak, A., Zhang, C., Swaminathan, K., and Xu, H. E. (2011). Crystal structure of the PAC1R extracellular domain unifies a consensus fold for hormone recognition by class B G-protein coupled receptors. *PLoS ONE* 6 (5), e19682. doi:10.1371/journal.pone.0019682

Lam, H. C., Takahashi, K., Ghatei, M. A., Kanse, S. M., Polak, J. M., and Bloom, S. R. (1990). Binding sites of a novel neuropeptide pituitary-adenylate-cyclase-activating polypeptide in the rat brain and lung. *Eur. J. Biochem.* 193 (3), 725–729. doi:10.1111/j.1432-1033.1990.tb19392.x

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 (10), 2794–2812. doi:10.1002/prot.23106

Li, J., Jonsson, A. L., Beuming, T., Shelley, J. C., and Voth, G. A. (2013). Ligand-dependent activation and deactivation of the human adenosine A(2A) receptor. *J. Am. Chem. Soc.* 135 (23), 8749–8759. doi:10.1021/ja404391q

Liang, Y. L., Belousoff, M. J., Zhao, P., Koole, C., Fletcher, M. M., Truong, T. T., et al. (2020). Toward a structural understanding of class B GPCR peptide binding and activation. *Mol. Cell* 77 (3) 656–668. doi:10.1016/j.molcel.2020.01.012

Liang, Y. L., Khoshouei, M., Glukhova, A., Furness, S. G. B., Zhao, P., Clydesdale, L., et al. (2018). Phase-plate cryo-EM structure of a biased agonist-bound human GLP-1 receptor-Gs complex. *Nature* 555 (7694), 121–125. doi:10.1038/nature25773

Liao, C., May, V., and Li, J. (2019a). Assessment of conformational state transitions of class B GPCRs using molecular dynamics. *Methods Mol. Biol.* 1947, 3–19. doi:10.1007/978-1-4939-9121-1_1

Liao, C., May, V., and Li, J. (2019b). PAC1 receptors: shapeshifters in motion. *J. Mol. Neurosci.* 68 (3), 331–339. doi:10.1007/s12031-018-1132-0

Liao, C., de Molliens, M. P., Schneebeli, S. T., Brewer, M., Song, G., Chatenet, D., et al. (2019c). Targeting the PAC1 receptor for neurological and metabolic disorders. *Curr. Top. Med. Chem.* 19 (16), 1399–1417. doi:10.2174/1568026619666190709092647

Liao, C., Zhao, X., Brewer, M., May, V., and Li, J. (2017). Conformational transitions of the pituitary adenylate cyclase-activating polypeptide receptor, a human class B GPCR. *Sci. Rep.* 7 (1), 5427. doi:10.1038/s41598-017-05815-x

Metzner, P., Schütte, C., and Vanden-Eijnden, E. (2009). Transition path theory for Markov jump processes. *Multiscale Model. Simul.* 7 (3), 1192–1219. doi:10.1137/070699500

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. Theor. Comput.* 8 (9), 3314–3321. doi:10.1021/ct300418h

Noé, F., Horenko, I., Schütte, C., and Smith, J. C. (2007). Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J. Chem. Phys.* 126 (15), 155102. doi:10.1063/1.2714539

Noé, F., Schütte, C., Vanden-Eijnde, E., Reich, L., and Weikl, T. R. (2009). Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. United States* 106 (45), 19011–19016. doi:10.1073/pnas.0905466106

Pande, V. S., Beauchamp, K., and Bowman, G. R. (2010). Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* 52 (1), 99–105. doi:10.1016/j.ymeth.2010.06.002

Parthier, C., Kleinschmidt, M., Neumann, P., Rudolph, R., Manhart, S., Schlenzig, D., et al. (2007). Crystal structure of the incretin-bound extracellular domain of a G protein-coupled receptor. *Proc. Natl. Acad. Sci. United States* 104 (35), 13942–13947. doi:10.1073/pnas.0706404104

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26 (16), 1781–1802. doi:10.1002/jcc.20289

Pioszak, A. A., and Xu, H. E. (2008). Molecular recognition of parathyroid hormone by its G protein-coupled receptor. *Proc. Natl. Acad. Sci. United States* 105 (13), 5034–5039. doi:10.1073/pnas.0801027105

Prinz, J. H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., et al. (2011). Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* 134 (17), 174105. doi:10.1063/1.3565032

Ressler, K. J., Mercer, K. B., Bradley, B., Jovanovic, T., Mahan, A., Kerley, K., et al. (2011). Post-traumatic stress disorder is associated with PACAP and the PAC1 receptor. *Nature* 470 (7335), 492–497. doi:10.1038/nature09856

Schäfer, H., Zheng, J., Morys-Wortmann, C., Fölsch, U. R., and Schmidt, W. E. (1999). Structural motifs of pituitary adenylate cyclase-activating polypeptide (PACAP) defining PAC1-receptor selectivity. *Regul. Pept.* 79 (2-3), 83–92. doi:10.1016/s0167-0115(98)00147-5

Siu, F. Y., He, M., de Graaf, C., Han, G. W., Yang, D., Zhang, Z., et al. (2013). Structure of the human glucagon class B G-protein-coupled receptor. *Nature* 499 (7459), 444–449. doi:10.1038/nature12393

Song, G., Yang, D., Wang, Y., de Graaf, C., Zhou, Q., Jiang, S., et al. (2017). Human GLP-1 receptor transmembrane domain structure in complex with allosteric modulators. *Nature* 546, 312. doi:10.1038/nature22378

Sun, C., Song, D., Davis-Taber, R. A., Barrett, L. W., Scott, V. E., Richardson, P. L., et al. (2007). Solution structure and mutational analysis of pituitary adenylate cyclase-activating polypeptide binding to the extracellular domain of PAC1-Rs. *Proc. Natl. Acad. Sci. United States* 104 (19), 7875–7880. doi:10.1073/pnas.0611397104

Vaudry, D., Falluel-Morel, A., Bourgault, S., Basille, M., Burel, D., Wurtz, O., et al. (2009). Pituitary adenylate cyclase-activating polypeptide and its receptors: 20 Years after the discovery. *Pharmacol. Rev.* 61 (3) 283–357. doi:10.1124/pr.109.001370

Wang, J., Song, X., Zhang, D., Chen, X., Li, X., Sun, Y., et al. (2020). Cryo-EM structures of PAC1 receptor reveal ligand binding mechanism. *Cell Res.* 30 (5), 436–445. doi:10.1038/s41422-020-0280-2

Weinan, E., and Vanden-Eijnden, E. (2006). Towards a theory of transition paths. *J. Stat. Phys.* 123, 503–523. doi:10.1007/s10955-005-9003-9

Wootten, D., Simms, J., Miller, L. J., Christopoulos, A., and Sexton, P. M. (2013). Polar transmembrane interactions drive formation of ligand-specific and signal pathway-biased family B G protein-coupled receptor conformations. *Proc. Natl. Acad. Sci. United States* 110 (13), 5211–5216. doi:10.1073/pnas.1221585110

Yang, L., Yang, D., de Graaf, C., Moeller, A., West, G. M., Dharmarajan, V., et al. (2015). Conformational states of the full-length glucagon receptor. *Nat. Commun.* 6, 7859. doi:10.1038/ncomms8859

Zhang, C., and Ma, J. (2010). Enhanced sampling and applications in protein folding in explicit solvent. *J. Chem. Phys.* 132 (24), 244101. doi:10.1063/1.3435332

Zhang, H., Qiao, A., Yang, D., Yang, L., Dai, A., de Graaf, C., et al. (2017a). Structure of the full-length glucagon class B G-protein-coupled receptor. *Nature* 546 (7657), 259–264. doi:10.1038/nature22363

Zhang, Y., Sun, B., Feng, D., Hu, H., Chu, M., Qu, Q., et al. (2017b). Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein. *Nature* 546 (7657), 248–253. doi:10.1038/nature22394

Zhang, H., Qiao, A., Yang, L., Van Eps, N., Frederiksen, K. S., Yang, D., et al. (2018). Structure of the glucagon receptor in complex with a glucagon analogue. *Nature* 553 (7686), 106–110. doi:10.1038/nature25153

Zhao, S., Zhu, K., Li, J., and Friesner, R. A. (2011). Progress in super long loop prediction. *Proteins* 79 (10), 2920–2935. doi:10.1002/prot.23129

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 © 2021 Liao, Remington, May and Li. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 12 April 2021
doi: 10.3389/fmolb.2021.636599

Characterization of Weak Protein Domain Structure by Spin-Label Distance Distributions

*Irina Ritsch[1], Laura Esteban-Hofer[1], Elisabeth Lehmann[2], Leonidas Emmanouilidis[2], Maxim Yulikov[1], Frédéric H.-T. Allain[2] and Gunnar Jeschke[1]**

[1]Department of Chemistry and Applied Biosciences, ETH Zürich, Zürich, Switzerland, [2]Department of Biology, ETH Zürich, Zürich, Switzerland

Function of intrinsically disordered proteins may depend on deviation of their conformational ensemble from that of a random coil. Such deviation may be hard to characterize and quantify, if it is weak. We explored the potential of distance distributions between spin labels, as they can be measured by electron paramagnetic resonance techniques, for aiding such characterization. On the example of the intrinsically disordered N-terminal domain 1–267 of fused in sarcoma (FUS) we examined what such distance distributions can and cannot reveal on the random-coil reference state. On the example of the glycine-rich domain 188–320 of heterogeneous nuclear ribonucleoprotein A1 (hnRNP A1) we studied whether deviation from a random-coil ensemble can be robustly detected with 19 distance distribution restraints. We discuss limitations imposed by ill-posedness of the conversion of primary data to distance distributions and propose overlap of distance distributions as a fit criterion that can tackle this problem. For testing consistency and size sufficiency of the restraint set, we propose jack-knife resampling. At current desktop computers, our approach is expected to be viable for domains up to 150 residues and for between 10 and 50 distance distribution restraints.

Keywords: low-complexity domains, RNA-binding proteins, ensemble structure, EPR spectroscopy, random coil, intrinsically disordered, liquid-liquid phase separation

OPEN ACCESS

Edited by:
Fabio Trovato,
Freie Universität Berlin, Germany

Reviewed by:
Paul Robustelli,
Dartmouth College, United States
Claudio Luchinat,
University of Florence, Italy

****Correspondence:***
Gunnar Jeschke
gjeschke@ethz.ch

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *01 December 2020*
Accepted: *19 February 2021*
Published: *12 April 2021*

Citation:
Ritsch I, Esteban-Hofer L, Lehmann E, Emmanouilidis L, Yulikov M, Allain FH-T and Jeschke G (2021) Characterization of Weak Protein Domain Structure by Spin-Label Distance Distributions. Front. Mol. Biosci. 8:636599. doi: 10.3389/fmolb.2021.636599

INTRODUCTION

Based on Anfinsen's influential thermodynamic hypothesis, function of proteins was considered for a long time to depend exclusively on domains that are well represented by a single conformer with minimum free energy. Structure of such domains can be specified at atomic resolution and can be characterized by well-established techniques, such as x-ray crystallography, NMR spectroscopy, and cryo-electron microscopy. However, meanwhile it is also well-established knowledge that a substantial fraction of the genome codes for intrinsically disordered domains (IDDs) or proteins (IDPs) that are also functional, without relying for their function on three-dimensional structure defined at atomic or near-atomic resolution (Uversky, 2019). Often, IDDs and IDPs are described as structure-less, implying that they exhibit random-coil behavior similar to the one postulated for chemically denatured proteins (Fitzkee and Rose, 2004). However, this assumption is at odds with the observation that many disordered domains contain evolutionary conserved residues and that mutations in such domains can be pathogenic. This applies even to disordered domains that have not been found to fold upon binding to other proteins or nucleic acids, suggesting that conformation space of proteins cannot be described by a dichotomy of atomic-resolution structure and complete

absence of structure. Instead, there exists a continuum between order and disorder (Uversky et al., 2005; Habchi et al., 2014). IDDs are ubiquitous in nucleic-acid binding proteins, as has been demonstrated early for transcription factors (Liu et al., 2006). More recently, formation of membrane-less organelles by RNA-binding proteins has been associated with the presence of IDDs (Brangwynne et al., 2015; Uversky, 2017). Despite impressive development of characterization techniques for IDDs over the past two decades, quantification of weak deviation from random-coil behavior has remained a complicated problem. In a dynamic picture, such weak order corresponds to correlated motion of residues or groups of residues with large sequence separation that would not be expected in a random coil (Kurzbach et al., 2016). In an ensemble-average picture, this correlated motion is expected to cause heterogeneity of segmental radius of gyration across the peptide sequence. In other words, sections of the chain will be more or less compact, depending on where in the sequence they are situated.

While the radius of gyration is easy to measure for the whole chain, such measurements are not feasible for sections of the chain. However, it is well known from polymer physics that the radius of gyration and the root mean square (RMS) end-to-end distance both scale by an exponential law with respect to the number of residues, with the scaling exponent coinciding for ideal chains. The end-to-end distance of chain sections is accessible by distance measurements between spin labels using pulsed dipolar spectroscopy (PDS) techniques, in particular by the double electron resonance (DEER) experiment. Such measurements are well established for proteins (Schiemann and Prisner, 2007; Jeschke, 2012). The primary data that they provide can be processed into distance distributions (Jeschke et al., 2002; Bowman et al., 2004; Chiang et al., 2005), which can in turn be predicted from polymer physics models (Zheng et al., 2018) and from ensemble models of a domain (Jeschke, 2016). Such characterization by distance distribution measurements has been applied to IDPs in the past (Drescher, 2012; Geist et al., 2013). Here we develop a general approach for using distance distribution data in ensemble modeling and for analyzing such ensembles in terms of weak order.

The first approaches for experimentally informed ensemble modeling of intrinsically disordered proteins utilized small-angle x-ray scattering curves (Bernado et al., 2007) or ensemble mean values of NMR parameters, such as various chemical shifts, $^3J_{HNHa}$ couplings, residual dipolar couplings, paramagnetic relaxation enhancements, and nuclear Overhauser enhancements (Marsh and Forman-Kay, 2012). The maximum entropy principle allows for combining molecular dynamics (MD) simulations with any type of constraints or restraints for which a forward model exists that can compute the measured quantity from atomic coordinates of a conformer (Cesari et al., 2018). Alternatively, ensembles resulting from unbiased MD simulations or other sampling techniques can be reweighted based on the maximum entropy principle (Köfinger et al., 2019). A somewhat similar approach, which has been discussed for DEER data, estimates the maximum occurrence of a conformer in an ensemble that is consistent with experimental restraints (Gigli et al., 2018). Principles of modeling proteins ensembles have been recently reviewed (Bonomi et al., 2017).

Our approach is based on the idea that knowledge of the probability distribution of experimental restraints, as is the case with DEER distance distributions, provides us with additional information compared to ensemble mean values. Hence we can apply our restraints as "distribution restraints". In a first modeling step, distribution restraints can guide sampling, as they allow for testing to what extent a single conformer is consistent with experimental data (Jeschke, 2016). In a second step, they can be used, together with other types of restraints, in ensemble reweighting (Jeschke, 2021). For application work, such integration with other types of data is desired and may often be necessary in order to sufficiently restrain the model and test its reliability. In the present study, we develop an approach that provides ensemble models from only distance distribution data and tests their robustness.

This development faces three obstacles. As a first obstacle, the reference state of complete disorder is assumed, but not proved to correspond to a Flory self-avoiding walk random coil. We need to test whether this approximation is sufficiently good in the range where label-to-label distance distribution measurements can be performed. To that end, we analyze data from a recent study on liquid-liquid phase separation (LLPS) of the intrinsically disordered N-terminal domain (NTD) of fused in sarcoma (FUS), where sections 10–29 and 105-128 exhibited featureless, Gaussian-like distance distributions (Emmanouilidis et al., 2021). Specifically, we address the question whether PDS EPR can distinguish between Gaussian-like distributions and self-avoiding walk distributions with variable scaling exponent ν (SAW-ν distributions) within the segment-length range where its distance distribution fidelity is best. This question is important for specifying restraints in modeling. As a second obstacle, conversion of primary data to distance distributions is an ill-posed problem (Jeschke et al., 2002; Chiang et al., 2005; Gigli et al., 2018), which may be particularly difficult to solve for very broad distributions that extend to rather long distances (Jeschke et al., 2004). We address this problem by systematic analysis of 19 distance distributions obtained on heterogeneous nuclear ribonucleoprotein A1 (hnRNP A1), a human protein predominantly active in mRNA regulation (Jean-Philippe et al., 2013). HnRNP A1 consists of a glycine-rich domain (188–320), which is an intrinsically disordered low-complexity domain (LCD) with a similar amino acid composition as the NTD of FUS, tethered to two folded RNA recognition motifs (RRMs). We analyze DEER data obtained with one or both label sites in the LCD of hnRNP A1 with different approaches and check how this influences ensemble modeling. Apart from addressing data analysis, hnRNP A1 also provides a reference case for analyzing weak structure in flexible domains tethered to folded domains with known structure. This is achieved by including the previously reported solution structure of the RRMs (Barraud and Allain, 2013) in the modeling. As a third obstacle, it is difficult to estimate how many distance distributions are required for sufficiently restraining an ensemble model in the regime of weak order. In order to answer this question, we perform jack-knife resampling, where systematically one of the

19 restraints for hnRNP A1 is left out and the corresponding distance distribution is predicted by an ensemble model fitted to the other 18 restraints. Finally, we discuss limitations of our approach as well as possible extensions.

MATERIALS AND METHODS

Sample Preparation for hnRNP A1

A construct of His_6-tagged wild-type hnRNPA1 (P09651, isoform A1-A), was already available in the Allain group at ETH Zurich, and a two-step affinity chromatography protocol was adapted from (Barraud and Allain, 2013) to purify all mutants. Further experimental details can be found in (Ritsch, 2019). In summary, point mutations of the partially buried native Cys in the folded domains to non-reactive residues (C43S, C175A), and to introduce pair-wise engineered Cys spin labeling sites were introduced by sequential application of site-directed mutagenesis with PCR primers with the desired point mutation. Spin labeling was performed with MTSL (2,2,5,5-tetramethyl-3-pyrroline-3-methylmethanethiosulfonate) in 50 mM sodium phosphate buffer, pH 6.5, 100 mM L-arginine, and 100 mM L-glutamate at approximately 10 times molar excess and 10 μM hnRNP A1 concentration. Excess spin label was removed on a PD10 desalting column (GE Healthcare), and the labeled protein was concentrated in 10 kDa MWCO centricons (Amicon Ultra-4 Centrifugal Filter Units, Merck and Cie). DEER samples were prepared by mixing with d8-glycerol (1:1 ratio, v:v), and transferring ~35 μL solution at approximately 25 μM hnRNP A1 concentration to 3 mm outer diameter quartz capillaries and flash-freezing by immersion into liquid nitrogen.

DEER Measurements on hnRNP A1

The DEER measurements were performed at a home-built high power Q-band spectrometer (≈34 GHz) controlled by a Bruker Elexsys E580 bridge in a home-built TE001-type resonator at 50 K. The 4-pulse DEER pulse sequence was used, with pulse lengths of $t(\pi) = t\,(\pi/2) = 16$ ns. The pump pulse length was either 16 or 12 ns (on the maximum of the nitroxide spectrum), and a pump/detection frequency separation of 100 MHz was used. The first refocusing delays was $t_1 = 400$ ns, and the second refocusing delay was set individually for each sample (as long as possible to be able to still detect a reasonably strong echo). The time-step was either 12 or 8 ns.

Distance Distribution Analysis

The sample preparation and DEER measurements of FUS are described in detail in (Emmanouilidis et al., 2021). Distance distribution analysis for FUS was performed using DeerLab 0.9.0 (downloadable at jeschkelab.github.io/DeerLab/). Briefly, we used a multi-pathway kernel $K(t, r)$ of the form

$$K(t,r) = \left[\Lambda_0 + \lambda_1 K_0(t,r) + \lambda_2 K_0\left(t - T_{0,2}, r\right)\right] e^{-k\left(\lambda_1 |t| + \lambda_2 \left|t - T_{0,2}\right|\right)}$$

where K_0 is the elementary dipolar kernel, Λ_0 accounts for the contribution of unmodulated dipolar pathways, λ_1 and λ_2 describe the amplitudes of the modulated dipolar pathways, $T_{0,2}$ is the refocusing time of the additional modulated dipolar pathway, and k is the background decay rate. While K_0 is fixed, Λ_0, λ_1, λ_2, and k are fit parameters. We assumed an exponential background function. The distance distribution $P(r)$ was fitted by Tikhonov regularization (using either the Bayesian information criterion or the residual method for regularization parameter selection), by fitting a single Gaussian distribution by varying mean distance and full width at half maximum Γ, or by fitting the SAW-ν model by varying RMS end-to-end distance and scaling exponent ν. Standard deviations σ_r were computed as $\sigma_r = \Gamma/\sqrt{8ln2}$. The dispersed phase data were taken not from the biphasic sample, but from monophasic dispersed FUS 1–267 sample in 0.6 M urea that had been prepared for obtaining initial values and lower and upper bounds for fitting biphasic data, as shown in (Emmanouilidis et al., 2021). Distance distribution analysis for hnRNP A1 was performed by single-Gaussian fitting in DeerAnalysis 2019 using default settings, by single-Gaussian and Tikhonov analysis in DeerLab 0.8b, and by DeerNet from Spinach version 2.5.5446. For Tikhonov analysis, regularization parameters were either selected by the AIC criterion or fixed at α = 5. In DeerAnalysis and DeerLab, a monoexponential background function was assumed corresponding to a homogeneous distribution of remote spin labels in three-dimensional space. DeerNet is trained with stretched exponential background functions (Worswick et al., 2018), thus also allowing for fractal dimensionality lower than three of the spatial distribution of remote labels. Results of single-Gaussian analysis with LongDistances[1] were provided by Christian Altenbach and results of multi-Gaussian analysis by DD (Stein et al., 2015) were provided by Eric Hustedt.

Generation of Raw Ensembles

Raw ensembles were generated in MMM using the Domain Ensemble Modeller described in (Jeschke, 2016). Briefly, in unrestrained mode the Domain Ensemble Modeller generates peptide backbone models that conform to residue-specific Ramachandran statistics for backbone torsions ϕ and ψ as provided by (Hovmoller, 2002). In restrained mode, each distance distribution restraint is tested in a given conformer as soon as backbone coordinates for both sites are available. The mean spin label position at the newly generated site is predicted from backbone coordinates, the label-to-label distance is computed, and a probability is estimated for this distance to be consistent with a Gaussian restraint. If the product of these probabilities for all evaluated restraints drops below a certain threshold, the conformer is discarded and generation of a new conformer is started. As such rejection of conformers typically occurs well before they are completely generated, the approach improves sampling of the part of conformational space that is consistent with the restraints.

Restraints were set as $\langle r \rangle$, $\sqrt{2}\sigma_r$ and an acceptance threshold corresponding to probability 0.75 was used. For FUS, a fixed number of conformers was generated in a single run. For FUS

[1]http://www.biochemistry.ucla.edu/Faculty/Hubbell/.

1–267, this number was 90 and in all other cases it was 2,500. For unrestrained FUS 1–100 and 67–166, two runs with 2,500 conformers each were performed. For hnRNP A1, PDB structure 2LYV of UP1 (Barraud and Allain, 2013) was used for the RRMs. Coordinates of only residues 1–187 were kept. In all runs, LCD models were generated for all 20 models in the NMR structure, expending 1 h computation time on eight processor cores per NMR model. This generated about 400 conformers. For the unrestrained ensemble, 48 min were expended per NMR model, leading to 2,146 conformers.

Ensemble Fitting

Ensemble fitting was performed with the *module_ensemble_fit* function of MMMx (github.com/gjeschke/MMMx, commit 55a7fef). Distance distribution files were provided in a four column format with the first column being the distance axis (units of nanometers), the second column being the distance distribution, and the third and fourth column the lower and upper bound of the uncertainty band (used only for plotting). The only change with respect to the corresponding module in MMM, as described in (Jeschke, 2021), is adaptive block size in iterative fitting. If less than 10% of the specified block size are available for the next iteration, block size is increased by 50%. If the number of retained conformers later drops, block size is reduced to the originally specified value, which was 100 in our computations.

RESULTS

The Random-Coil Reference State

Probing the Reference State by DEER Distance Distributions

Before we can address the problem of weak structure, we need to establish what can be inferred by DEER distance distributions on an unstructured protein. In particular, we are interested in the question what level of detail can be realistically interpreted in an ensemble model informed by such distributions. We analyze this on the example of the NTD of FUS, where we used protein constructs that do not contain any well-structured domains and are thus well suited as a fully disordered reference system.

We define structure as deviation from the maximum-entropy state of a protein domain under given solvent conditions. From a polymer physics view, this maximum-entropy state corresponds to a random coil. Since protein domains are not homopolymers, some deviation from simple polymer physics models is expected. For instance, Ramachandran angle preferences are residue-specific even in loop regions of proteins (Hovmoller et al., 2002), suggesting that the maximum-entropy state of a domain depends on sequence. Therefore, the random-coil reference state itself must be characterized first before conclusions can be drawn on weak structure.

It has been demonstrated as early as 2004 that randomization of backbone torsion angles of structured proteins leads to ensembles whose RMS end-to-end distances R and radii of gyration R_G conform to expectations for random coils, with a scaling law $R \propto N^{\nu}$ where N is the number of residues and $\nu \approx 0.6$ for chemically denatured proteins (Fitzkee and Rose, 2004). These conditions correspond to a Flory random coil in a good solvent where it maximally expands ($\nu = 3/5$). A single-molecule Förster resonance excitation transfer (FRET) study found $\nu \approx 0.62$ for denatured and IDPs, whereas foldable sequences in water exhibited $\nu \approx 0.48$, close to the condition of a Θ-solvent ($\nu = 1/2$) that offsets monomer-monomer interactions (Hofmann et al., 2012). Accordingly, the distribution $P(r)$ of end-to-end distances in IDDs is expected to conform to a self-avoiding random walk. Dependence of chain dimension on solvent quality is then quantified by the scaling exponent ν. An expression for $P(r)$ as a function of R and ν has been derived (SAW-ν model) and compared to molecular dynamics (MD) simulations (Zheng et al., 2018). The distribution is skewed with a steeper flank toward short distances than toward long distances.

In recent work on the NTD of FUS, we found distance distributions between spin labels for the two chain sections 10–29 and 105–128 that were experimentally indistinguishable from a Gaussian distribution, which is symmetric (Emmanouilidis et al., 2021). In particular, we have studied this NTD in a denatured state in the presence of 3 M urea, in a dispersed state at low concentration in the presence of 0.6 M urea, and in a bulk condensed phase that is obtained at 0.6 M urea concentration and higher protein concentration by liquid-liquid phase separation and isolation of the condensed phase by centrifugation. In order to rationalize the finding that distributions are very well approximated in all three cases, we first computed the distributions according to the SAW-ν model for poor- and good-solvent conditions and compared them to a Gaussian distribution (**Figure 1A**). Whereas in poor-solvent conditions, where the chain maximally compacts ($\nu = 1/3$, red line), asymmetry of $P(r)$ is pronounced, it is only modest in good-solvent conditions (green line). A fit of the good-solvent $P(r)$ by a Gaussian distribution (black line) shows rather minor deviations that may be hard to resolve by EPR distance distribution measurements. The main deviation at short distances is usually below the lower limit of the distance range of PDS techniques, whereas extension of the SAW-ν distribution to longer distances than in a Gaussian distribution may be hard to separate from intermolecular background.

In order to corroborate this hypothesis, we have generated unrestrained ensembles of the NTD of FUS (residues 1–267) and of two 100-residue sections thereof (1–100 and 67–166) by a Monte-Carlo algorithm that samples from residue-specific Ramachandran distributions of backbone torsion angles for loop regions in proteins (Hovmoller et al., 2002). Otherwise, this algorithm only avoids backbone and sidechain clashes (Jeschke, 2016). The simulation of 100-residue sections allowed to generate much larger raw ensembles than could be obtained within the same computation time for the complete NTD. The ranges were selected such that both studied subsections 10–29 and 105–128 were flanked to both sides with the maximum number of residues possible for a 100-residue construct. We consider the ensuing ensembles as good approximations for the maximum-entropy state of a protein sequence. In order to compare them with the Flory random coil model, which departs from a freely jointed chain model, we

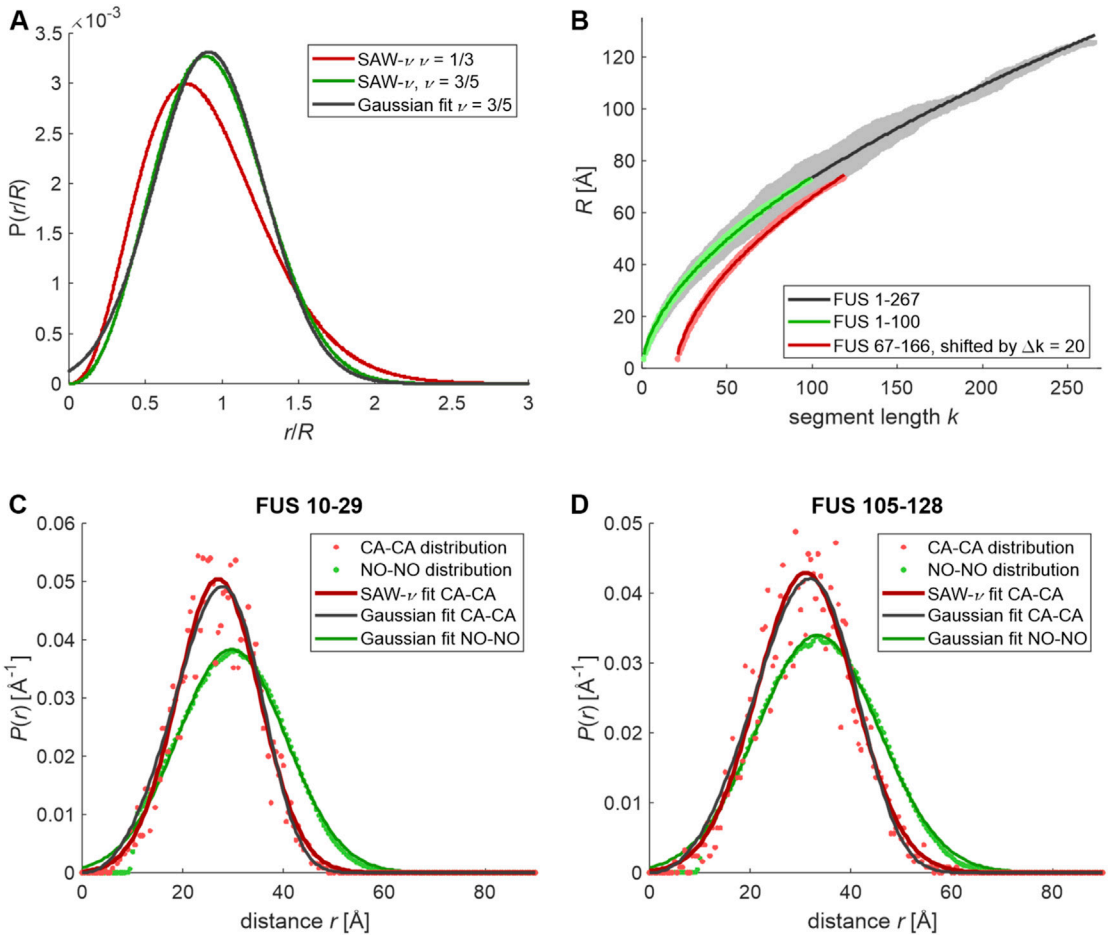


FIGURE 1 | End-to-end and label-to-label distance distributions simulated for the unrestrained NTD of FUS and sections thereof. Residue-specific Ramachandran statistics was assumed as extracted by Hovmoller et al. (2002) from PDB for residues not denoted as helix or strand by DSSP. **(A)** End-to-end distributions predicted by the self-avoid walk model with variable scaling exponent (SAW-ν) by Zheng et al. (2018) for a poor solvent (ν = 1/3, red) and for a good solvent (ν = 3/5, green) and fit of a normal distribution to the good-solvent case (black). **(B)** Segment-wise root mean square CA-CA distances extracted from 90 simulated conformers of FUS 1–267 (gray), and 2,500 conformers of FUS 1–100 (light green) and FUS 66–167 (light red, horizontally shifted by 20 segments for clarity) and fits by a power law $b{\cdot}k^{\nu}$ (solid black, green, and red lines, respectively). **(C)** CA-CA distance distribution for residues 10–29 in 5,000 conformers of FUS 1–100 (light red dots) and fits by the SAW-ν model (red line) and a normal distribution (black line) as well as the MTSL-MTSL distance distribution (pale green dots) with fit by a normal distribution (green line). **(D)** CA-CA distance distribution for residues 105–128 in 5,000 conformers of FUS 67–166 (light red dots) and fits by the SAW-ν model (red line) and a normal distribution (black line) as well as MTSL-MTSL distance distribution (pale green dots) with fit by a normal distribution (green line).

TABLE 1 | Fit parameters of experimental distance distributions for MTS labels attached at the ends of segments 10–29 and 105–128 in FUS NTD 1–267. The mean value $\langle r_{DEER}\rangle$ and standard deviation $\sigma_{r,DEER}$ are specified for Gaussian fits, while the RMS end-to-end distance R_{DEER} and scaling exponent ν_{DEER} are specified for the SAW-ν distribution.

Site1	Site 2	Conditions	$\langle r_{DEER}\rangle$ [Å]	$\sigma_{r,DEER}$ [Å]	R_{DEER} [Å]	ν_{DEER}
10	29	3 M urea	36.7	9.4	37.4	0.85
10	29	0.6 M urea	31.8	9.9	33.2	0.75
10	29	Bulk condensed	29.7	10.9	31.8	0.64
105	128	3 M urea	39.7	11.2	40.9	0.82
105	128	0.6 M urea	35.9	11.5	37.3	0.76
105	128	Bulk condensed	31.6	11.7	33.7	0.64

analyzed the scaling of the RMS end-to-end distance R_k for all segments with their length k, assigning the CA-CA distance between residues i and j to $k = j\text{-}i$ (Jeschke, 2021). For an ensemble of 90 conformers of FUS 1–267 (gray points in **Figure 1B**), we find a best fit $R_k = 5.23{\cdot}k^{0.57}$ Å (dark gray line) in rather good agreement with values inferred by FRET and MD

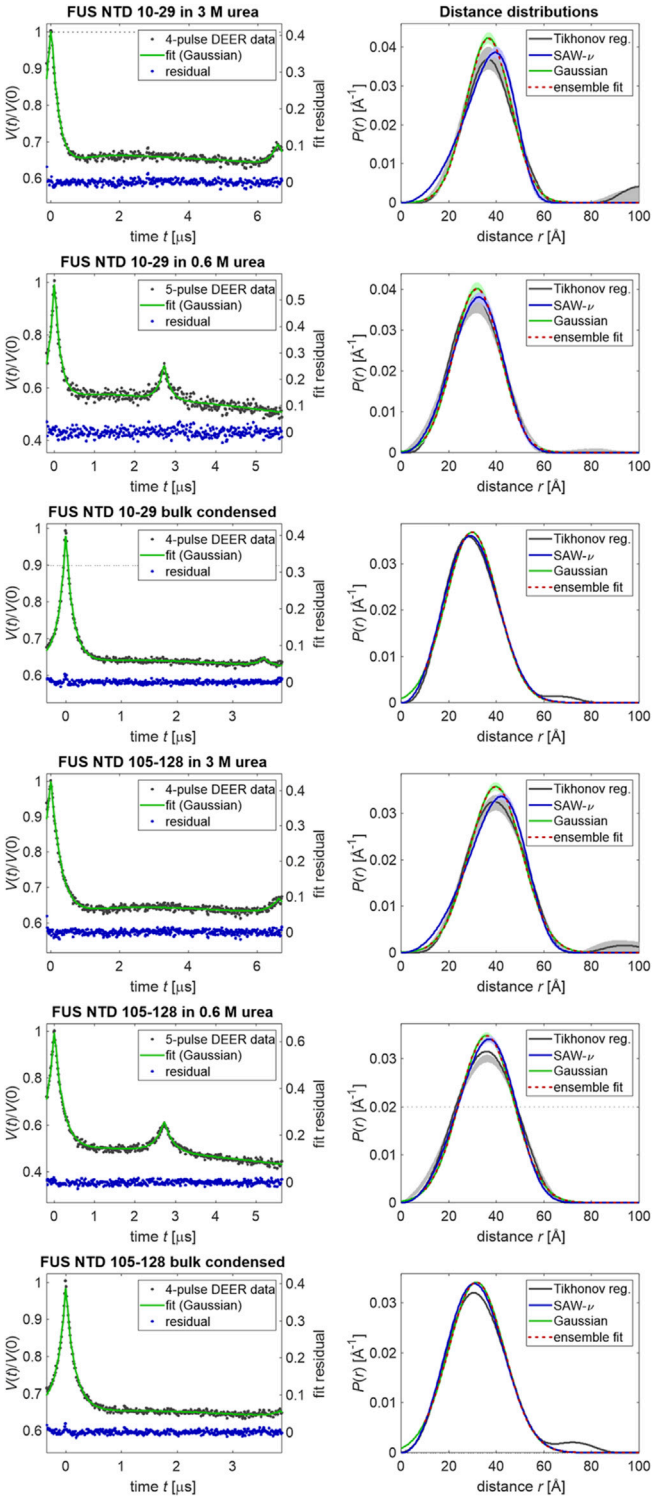


FIGURE 2 | Distance distribution measurements of sections delimited by spin labeling sites (10–29 or 105–128) measured with FUS NTD 1–267. The left column shows primary data (black dots), their DeerLab fit by a single Gaussian distribution (green, left ordinate) and the fit residual (blue dots, right ordinate). Signals from minor pathways at the end of 4-pulse DEER data or at the middle of 5-pulse DEER data were included in the kernel (Fabregas Ibáñez et al., 2020). The right column shows distance distributions obtained by Tikhonov regularization with choice of the regularization parameter by the Bayesian information criterion or residual method (black with gray uncertainty band), by single-Gaussian fitting (green with pale green uncertainty bands) and ensemble fits of the Gaussian distribution (red) from raw ensembles of 2,500 conformers each.

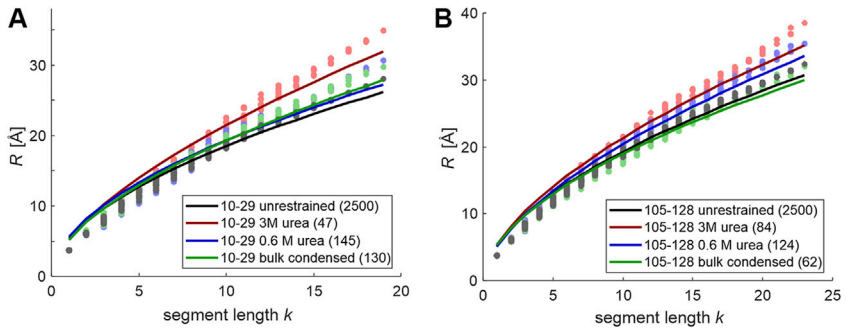


FIGURE 3 | Segment-wise root mean square CA-CA distances (dots) for chain sections 10–29 in FUS 1–100 **(A)** and 105–128 in FUS 67–166 **(B)** and attempted fits by a scaling law $b{\cdot}k^{\nu}$. For unrestrained ensembles, the average is over 2,500 conformers with uniform population. For restrained ensembles, the average population-weighted and the number of conformers in the fitted ensembles is given in the legend in parentheses.

simulations for IDPs (Hofmann et al., 2012; Zheng et al., 2018). Variation of the R_k for different pairs (i,j) at the same k appears to be substantial. However, we could trace this back to the low number of conformers used in this analysis. When we repeated the analysis with ensembles of 2,500 conformers each for FUS 1–100 (green) and FUS 67–166 (red), we found a much narrower distribution of the R_k around the best-fit scaling laws $5.25{\cdot}k^{0.57}$ Å and $5.17{\cdot}k^{0.58}$ Å, respectively. We note, however, that the R_k are not symmetrically distributed around the scaling law at all k. The R_k appear to increase somewhat more steeply than the mean scaling law for small k and somewhat less steeply for large k. This point is extended below.

We have then inquired whether the distributions of CA-CA distances in the unrestrained ensembles (red dots in **Figures 1C,D**) conform to the SAW-ν distribution (red lines). We find R = 27.7 Å and ν = 0.78 with a RMS deviation (RMSD) of 0.163 for section 10–29 in FUS 1–100 and R = 31.9 Å, ν = 0.77 with RMSD 0.171 for section 105–128 in FUS 67–166. The scaling exponents ν appear to be too large, as they should not exceed 0.6 for a random coil in a good solvent. Moreover, Gaussian fits (black lines) of the same distributions result in similar RMSD. We find a mean value $\langle r\rangle$ = 27.0 Å and standard deviation σ_r = 8.2 Å with RMSD 0.155 for section 10–29 in FUS 1–100 and $\langle r\rangle$ = 31.0 Å and standard deviation σ_r = 9.7 Å with RMSD 0.170 for section 10–29 in FUS 105–128.

Measurement of the distribution between spin labels further complicates detection of the asymmetry, as the label-to-label distribution is a convolution of the CA-CA distribution with a contribution from the sidechain rotamer distribution of the label. In order to assess this effect, we have simulated the label-to-label distribution by a rotamer library approach (Polyhach et al., 2011) in MMM (Jeschke, 2018). Since on the order of 100 × 100 rotamers at the two sites are populated for each backbone conformer, the resulting distributions (green points in **Figures 1C,D**) are much smoother. They are fitted quite well by Gaussian distributions (dark green lines) with $\langle r_{\mathrm{label}}\rangle$ = 29.7 Å and standard deviation $\sigma_{r,\mathrm{label}}$ = 10.8 Å for section 10–29 in FUS 1–100 and $\langle r_{\mathrm{label}}\rangle$ = 33.4 Å and standard deviation $\sigma_{r,\mathrm{label}}$ = 12.1 Å for section 105–128 in FUS 67–166.

Figure 2, with original data taken from (Emmanouilidis et al., 2021), demonstrates that the quality of Gaussian distribution fits to the primary DEER data is high for FUS 1–267 denatured in 3 M urea, fully dispersed at 5 μM protein concentration in 0.6 M urea, and in bulk condensed phase after LLPS. In the left column, primary data that were only phase and zero-time corrected (black dots) are superimposed by fits in DeerLab (Fabregas Ibáñez et al., 2020) in which a single Gaussian distance distribution and an exponential background function are simultaneously fitted (green line, left ordinate). The exponentially decaying background accounts for interaction with spin labels in other FUS NTD molecules that are assumed to be homogeneously distributed in three-dimensional space. The residuals are shown as blue dots (right ordinate). The only apparent deviations of the residual from the expected white noise occur near zero time in the two bulk condensed phase measurements and correspond to an underestimate of the short distance contribution by the fit. Note that an overestimate would be expected if the data were better represented by an SAW-ν distribution. The likely cause of the underestimate is neglect of the contribution at very short distances in the DeerLab simulations, as it is expected to be suppressed by insufficient excitation bandwidth. This suppression appears to be weaker than assumed.

In the right column of **Figure 2**, distance distributions obtained by Tikhonov regularization (black lines with gray uncertainty bands, 95% confidence interval) are compared with the fitted Gaussian distributions (dark green lines with pale green uncertainty bands). Tikhonov regularization does not make an assumption about the shape of the distribution, except for a certain degree of smoothness. Although it can be argued that the Tikhonov distributions deviate from the Gaussian distributions toward the expected more asymmetric shape with lower contributions at short distances and higher contributions at the longest distances, these deviations are minor and are clearly seen only in the bulk condensed phase data. Nevertheless, we have attempted fits of the DEER data by the SAW-ν model. The fitted distributions are shown as blue lines with pale blue uncertainty bands in **Figure 2**, the fit parameters are reported in **Table 1** alongside the parameters of the Gaussian fits, and the fits of the primary data and fit residuals are shown in **Supplementary**

Figure S1. Only for FUS 1–267 in 3 M urea do the SAW-ν fits differ significantly from the distributions obtained by Tikhonov regularization or Gaussian fits. However, in these cases the asymmetry is opposite to the expected one, with more gradual increase of $P(r)$ at short distances and steeper decrease at longer distances. This can be traced back to the unphysical scaling exponents of ν = 0.85 for segment 10–29 and 0.82 for segment 105–128 that we find by these fits. These findings reinforce our conclusion that parameters of the SAW-ν model cannot reliably be extracted by fitting DEER distance distributions for chain segments containing about 20 residues. The maximum-entropy reference state is well approximated by Gaussian label-to-label distance distributions. On the one hand, this finding justifies the use of Gaussian restraints in generating raw ensembles, as introduced in (Jeschke, 2016). On the other hand, it suggests that longer segments would need to be studied in order to extract parameters of the SAW-ν model. Such an approach would, however, be limited by uncertainty of the shape of DEER distance distributions at distances longer than 50–80 Å.

We also note that measurements for the characterization of LLPS are more sensible in the biphasic system containing both the dispersed and condensed phase under otherwise identical conditions, as such samples match the surface-to-volume ratio of biological membraneless organelles and such measurements require much less protein (Emmanouilidis et al., 2021). In such systems, however, the separation of the distance distributions corresponding to the dispersed and condensed phase introduces further uncertainty. We have therefore limited our analysis in the present study to the distance distributions of monophasic protein samples (either in solution or a large phase-separated compartment).

A Hybrid Experimental and Computational Approach

We have then turned to the question whether a hybrid experimental/computational approach can shed more light on the reference state. To that end we have generated six raw ensembles of 2,500 conformers each of FUS 1–100 and 67–166 by the Monte Carlo approach introduced in (Jeschke, 2016) by imposing the Gaussian restraints for sections 10–29 and 105–128, respectively, in the denatured, dispersed, and bulk condensed states. We have then refined these ensembles by fitting populations of the conformers and discarding all conformers with less than 1% of the population of the most populated conformer, as described in (Jeschke, 2021). These computations have been performed with an implementation of the EnsembleFit module of MMM into the successor program MMMx[2]. With this size of the conformer basis set, we find that the Gaussian distance distributions can be fitted virtually perfectly for both sections in all conditions tested (red dashed lines in the right column of **Figure 2**). The overlap of the normalized experimental and ensemble-simulated distance distributions varies between 98.2% (105–128, 0.6 M urea, dispersed) and 99.1% (10–29 0.6M urea, dispersed) with sizes of the refined ensemble between 47 (10–29, 3 M urea, denatured) and 145 (10–29, 0.6 M urea, dispersed) conformers. We found that fit quality was converged with respect to adding further conformers from the raw ensemble after evaluating about 1,000 conformers. We conclude that the methodology of generating raw ensembles of conformers by a Monte Carlo approach and refining and reducing these ensembles by population fitting can provide relatively small ensembles that fully represent the information contained in EPR-measured label-to-label distance distributions.

We have then tested whether these ensembles allow for inference on the scaling exponent ν of random-coil models. To that end, we have performed the analysis of the scaling of R_k for all segments of length k for FUS sections 10–29 and 105–128 for all conditions tested. The data is displayed in **Figure 3** together with data from the unrestrained ensembles (gray). In all cases we find that fits by a scaling law $b{\cdot}k^{\nu}$ are mediocre. Fitted scaling exponents ν are in a reasonable range between 0.538 (FUS 10–29 unrestrained) and 0.619 (FUS 10–29, 3M urea), as are the Kuhn lengths b between 5.14 Å (FUS 105–128, 0.6 M urea) and 5.58 Å (FUS 10–29, 0.6 M urea). However, deviations from the fits are systematic, with longer RMS end-to-end distances for the longest segments than predicted. This corresponds to chains that are stiffer than predicted. Comparison of the unrestrained case with **Figure 1B** reveals that such deviations are much smaller for longer chains than for the short sections considered here. Such behavior of shorter chain sections being stiffer than longer ones we have encountered before in the context of semi-rigid organic polymers (Godt et al., 2006; Jeschke et al., 2010). We note that the sections studied here are oligomers in the sense of the Flory random-coil model, as their length is less than ten times the Kuhn length b. As a consequence, we caution against assuming random-coil behavior for the fully disordered reference state. Instead, we advocate unrestrained Monte-Carlo simulations of large ensembles of conformers based on Ramachandran statistics for loop regions. These raw ensembles can then be fitted to experimental distance distributions and the resulting refined ensembles can be analyzed in terms of the distributions of R_k for segment lengths k. If these distributions are narrow and scaling of the mean values is monotonous and smooth, we can assume the chain to be in a fully disordered state. This need not be the case in general, as we demonstrate in the following on the example of the glycine-rich domain of hnRNP A1. This domain has high composition similarity to the FUS NTD, but was here studied in the context of the full length protein including the folded RNA-recognition motifs.

Computation of Distance Distributions for Ensemble Fitting

Choice of DEER Data Processing Approach

Accuracy of DEER distance distributions is limited by experimental and computational aspects (Jeschke, 2012). From an experimental point of view, excitation bandwidth limitations cause partial suppression of contributions below about 18 Å. Since the assumptions of weak coupling of the two spins and of negligible exchange coupling break down below about 15 Å, few efforts have been made to overcome this excitation

[2]http://www.github.com/gjeschke/MMMx.

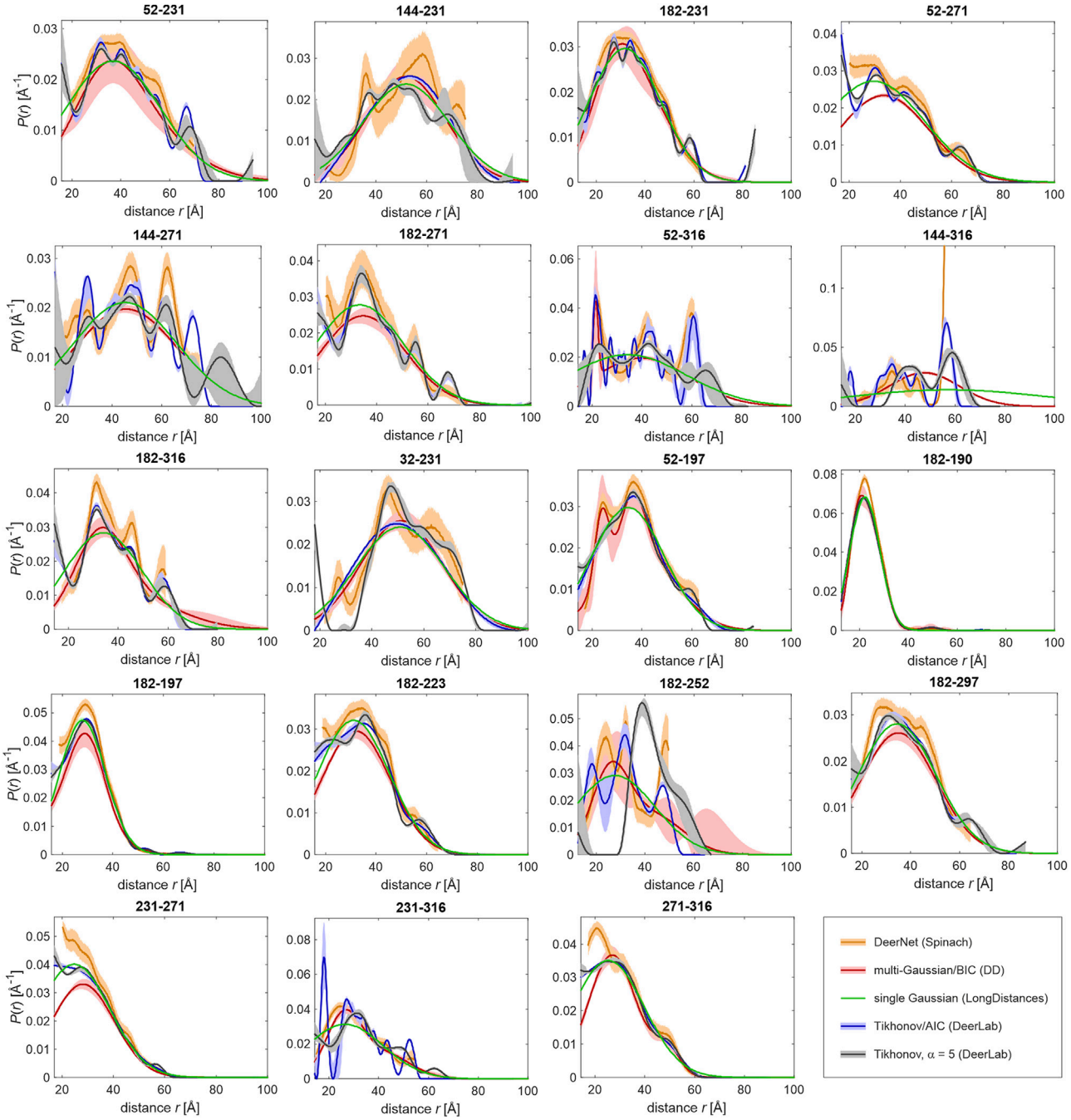


FIGURE 4 | Distance distributions computed by different approaches from primary DEER data for 19 pairs of MTS-labelled sites. The approaches are DeerNet in Spinach 2.5.5446 (orange), multi-Gaussian fit with the number of Gaussians determined by the Bayesian information criterion (BIC) in DD (red), single Gaussian fit including background in LongDistances (green), Tikhonov regularization with regularization parameter α determined by the Akaike information criterion (AIC) in DeerLab 0.8b (blue), and Tikhonov regularization with fixed α = 5 in DeerLab 0.8b (gray/black). Pale areas denote uncertainty bands (50% confidence interval for DeerLab, 95% confidence interval for DeerNet and DD).

bandwidth limitation. At long distances, the shape or even width of the distribution become uncertain, as this information is encoded in the decay of the dipolar oscillations and can thus be recovered only up to distances where several dipolar oscillation periods are still within the maximum observation time t_{max}. The accessible t_{max} depends on electron spin decoherence time and can vary strongly between proteins. For soluble proteins measured with high-power Q-band DEER, shape information is typically reliable up to 50 to 80 Å, depending on whether the solvent and protein can be deuterated. The reliable distance range

TABLE 2 | Single-Gaussian restraint sets (mean distance $\langle r \rangle$ and standard deviation σ_r) for hnRNP A1 obtained with three different program packages (DeerAnalysis, DA; DeerLab, DL; LongDistances, LD).

Site1	Site2	$\langle r \rangle_{DA}$ [Å]	σ_{DA} [Å]	$\langle r \rangle_{DL}$ [Å]	σ_{DL} [Å]	$\langle r \rangle_{LD}$ [Å]	σ_{LD} [Å]	$\langle r \rangle_{DL}/\langle r \rangle_{DA}$	$\langle r \rangle_{LD}/\langle r \rangle_{DA}$
52	231	37.5	18.1	36.7	19.5	36.9	19.4	0.979	0.984
144	231	49.7	15.4	52.7	14.9	51.8	17.2	1.060	1.042
182	231	33.3	14.8	31.7	14.8	31.7	14.8	0.952	0.952
52	271	28.1	19.7	29.4	19.9	29.5	19.9	1.046	1.050
144	271	45.7	18.6	44.7	21.1	45.2	20.8	0.978	0.989
182	271	30.5	18.8	33.2	17.1	33.4	17.2	1.088	1.095
52	316	37.5	18.2	36.3	21.2	34.2	23.6	0.968	0.912
144	316	44.1	11.7	45.6	16.9	57.3	38.5	1.034	0.873
182	316	34.8	15.7	33.8	15.6	34.0	15.6	0.971	0.977
32	231	51.2	15.5	50.7	15.8	50.9	17.1	0.990	0.994
52	197	35.6	13.8	34.9	14.2	34.4	14.5	0.980	0.966
182	190	22.5	6.2	22.0	6.3	22.2	6.1	0.978	0.987
182	197	27.4	9.5	27.9	9.2	28.1	9.1	1.018	1.026
182	223	31.4	14.2	31.0	14.5	31.1	14.2	0.987	0.990
182	252	36.6	12.7	29.9	15.1	27.6	16.5	0.817	0.754
182	297	34.1	15.8	34.2	16.1	34.5	15.9	1.003	1.012
231	271	20.1	15.3	24.5	14.1	24.7	13.9	1.219	1.229
231	316	27.1	16.5	28.5	14.1	26.5	15.4	1.052	0.978
271	316	22.6	15.8	24.9	14.6	25.3	14.5	1.102	1.120

for a given sample also depends on the signal-to-noize ratio and on the presence of instrumental or experimental artifacts in the data, as conversion of the primary data to a distance distribution is an ill-posed problem where a least-squares solution is strongly affected by slight deviations of the data from the forward model (Jeschke et al., 2002; Chiang et al., 2005). During the past two decades, several approaches were suggested for tackling this problem, such as Tikhonov regularization (Bowman et al., 2004; Jeschke et al., 2004; Chiang et al., 2005), fitting by multiple Gaussians, whose number is determined by a statistical criterion (Stein et al., 2015), and training an ensemble of neural networks with large sets of forward-modelled data that cover the expected men values, widths, and shapes of DEER signals (Worswick et al., 2018). All recent implementations provide, both, the most probable distance distribution according to the computational approach and confidence intervals. We note, however, that estimates of confidence intervals for ill-posed problems may not cover model bias, which for regularization approaches includes bias arising from the choice of the regularization parameter. The latter aspect has been studied recently (Edwards and Stoll, 2018) and different criteria have been proposed for the choice of the regularization parameter. With such a large variety of approaches existing, it is of interest whether ensemble modeling is robust with respect to the choice of the data processing approach. The problem is further complicated by the necessity to separate the dipolar evolution function from intermolecular background (Jeschke et al., 2006). Analysis of the DEER data for FUS NTD sections above has revealed that different fit approaches may lead to different distance distributions even for data sets of rather high quality.

In our approach, we first generate a raw ensemble of conformers with restraints that specify only the mean and standard deviation of a single Gaussian per site pair (Jeschke, 2016). These parameters are more robust than the shape of the distribution (Jeschke et al., 2004; Jeschke, 2012). In a second step, we reweight and contract the raw ensemble by varying conformer populations and discarding conformers with less than 1% of the population of the most populated conformer (Jeschke, 2021). In this ensemble fitting step, we use full distance distributions rather than only mean distance and width. However, even if the distribution used in the second step is the same Gaussian used in the first step, ensemble reweighting can improve the fit quality with the same or with an even smaller number of conformers. Thus, the question arises which approach should be used for generating distance distributions for the use as restraints in ensemble reweighting, which depends on the selected fit criterion. Here, we consider maximization of overlap of experimental and predicted distance distributions (Jeschke, 2021), defined by

$$o_d = \sum \min\left\{\boldsymbol{P}_{pred}, \boldsymbol{P}_{DDR}\right\}, \tag{1}$$

where $\boldsymbol{P}_{\mathrm{pred}}$ is the forward-modelled distance distribution for the reweighted ensemble and $\boldsymbol{P}_{\mathrm{DDR}}$ is the distance distribution obtained from experimental data, both given as vectors with non-negative elements whose sum is unity.

In particular, we maximize the geometric mean $\overline{o} = (\prod_{m=1}^{M} o_m)^{1/\mathrm{M}}$ of the overlaps o_m for all M restraints, which strongly penalizes large overlap deficiency 1- o_m of individual restraints. Even if the ensemble of conformers is consistent with the primary time-domain data, large overlap deficiency can arise if conversion to the distance distribution uses a poorly suited approach. In principle, the problem could be solved by directly fitting to the primary data (Hilger et al., 2007). However, this makes fitting computationally expensive and would preclude global optimization of populations (Jeschke, 2021) for large sets of conformers. An alternative ensemble reweighting approach based on the maximum entropy principle (Köfinger et al., 2019) could be used as well in this second step.

TABLE 3 | Geometric mean overlap $\bar{o}$, number of conformers, radius of gyration R_g, and ensemble width Γ in ensembles of hnRNP A1 LCD (188–320) by using distance distributions computed with different approaches (see also **Figure 4**) in ensemble fitting from the same raw ensemble.

Analysis approach	Mean overlap $\bar{o}$	Number of conformers	R_g [Å]	Γ [Å]
DeerNet	0.836	57	24.0	46.6
Multi-Gaussian (AIC)	0.880	52	23.3	47.6
Multi-Gaussian (BIC)	0.882	56	23.4	47.6
Single Gaussian (LongDistances)	0.863	49	23.7	48.2
Single Gaussian (DeerLab)	0.883	52	23.7	47.4
Tikhonov (α by AIC criterion)	0.846	50	23.8	48.4
Tikhonov (α = 5)	0.876	50	23.7	48.2

In order to assess current approaches for conversion of dipolar signals to distance distributions in this context, we have analyzed 19 DEER data sets obtained on dispersed hnRNP A1. Given that there is no prior structural information available for a priori screening of suitable labeling in the LCD, we primarily mutated available Ser residues with approximately uniform sequence separations to Cys. The labeling sites in the folded domains of hnRNP A1 were selected to be spatially well separated and solvent accessible. Note that label site selection in folded domains typically requires case-by-case considerations with general guidance provided by rotamer modeling (Polyhach and Jeschke, 2010). Specifically, for the generation of distribution restraints labeling sites in folded domains with low local backbone flexibility can provide more precise distance information (Jeschke, 2016), but may be less tolerant toward site-directed mutagenesis and successive spin labeling. The selected sites 52 and 144 are located in partially flexible loops of the RRMs of hnRNP A1, which provides a good trade-off. The dataset consists of 16 restraints between RRM-LCD sites, and three between LCD-LCD sites.

The data were analyzed by fitting with a single Gaussian distribution, multiple Gaussian distributions, whose number is determined by a statistical criterion (Stein et al., 2015), the neural network approach DeerNet (Worswick et al., 2018), and Tikhonov regularization (Bowman et al., 2004; Jeschke et al., 2004; Chiang et al., 2005) with the regularization parameter either determined by a statistical criterion (Edwards and Stoll, 2018) or fixed at a value that is judged from appearance of the distributions (**Figure 4**). Multi-Gaussian fitting was performed in DD (Stein et al., 2015) using either the Akaike or Bayesian information criterion for determining the optimal number of Gaussian components. Analysis in terms of a single Gaussian distribution was performed with three different software packages, DeerAnalysis (Jeschke et al., 2006) version 2019, DeerLab (Fabregas Ibáñez et al., 2020) version 0.8b, and LongDistances[3]. The latter two programs simultaneously fit the distance distribution and background, whereas in DeerAnalysis we first performed background separation and then fitted the dipolar evolution function. **Figure 4** shows the results obtained with LongDistances. The mean distances and standard deviations obtained with all three approaches are listed in **Table 2**. For most data sets, agreement between the results of all three approaches is rather good, with the exceptions of site pairs 144–316 and 182–252. In these two cases, the single-Gaussian model is too simplistic for separating the pair contribution from background with sufficient certainty. As an example for fit quality of the primary data, the results obtained with Tikhonov AIC fitting are shown in **Supplementary Figure S2**.

Turning to different distribution models, for site pair 144–316 with a very broad distance distribution extending to rather long distances, all approaches run into difficulties, although agreement between Tikhonov regularization, DeerNet and multi-Gaussian fitting is reasonable. Splitting of a broad distribution into many moderately broadened peaks, such as encountered for site pair 52–316 when using the Akaike information criterion for determining the regularization parameter for Tikhonov regularization, may be detrimental to fitting with the overlap criterion as it substantially reduces overlap (**Supplementary Figure S3** left). For this reason, we have resorted to Tikhonov regularization with a fixed regularization parameter α = 5 (**Supplementary Figure S3**, middle), which is not expected to oversmooth distributions for the case at hand. In contrast, a single narrow Gaussian component, as it is sometimes found with the multi-Gaussian parametrized model, is less detrimental, since it contributes only a small fraction of the distribution (**Supplementary Figure S3**, right).

For most of the site pairs, agreement of the distance distributions is good between all approaches. This applies, in particular, to comparatively narrow distributions, which do not extend beyond 60 Å. For data sets where the approaches disagree more strongly, we also find broader uncertainty bands with the individual approaches. Note, however, that not all distributions agree within their specified uncertainties. Differences between the distance distributions obtained with different approaches indicate limited quality of the data sets, mainly because of mediocre signal-to-noize ratio or a maximum observation time that is insufficient for high confidence in background separation. We note that these features are not a sign of poor experimentation. Depending on width of the distribution and maximum distance, obtaining higher-quality data may be unrealistic.

Ensemble Fitting to Distance Distributions With Moderate Shape Uncertainty

The question then arises, whether distance distributions of such site pairs can still be used in ensemble fitting. The answer is not necessarily negative, since the overlap metric is less affected by the

[3]http://www.biochemistry.ucla.edu/Faculty/Hubbell/.

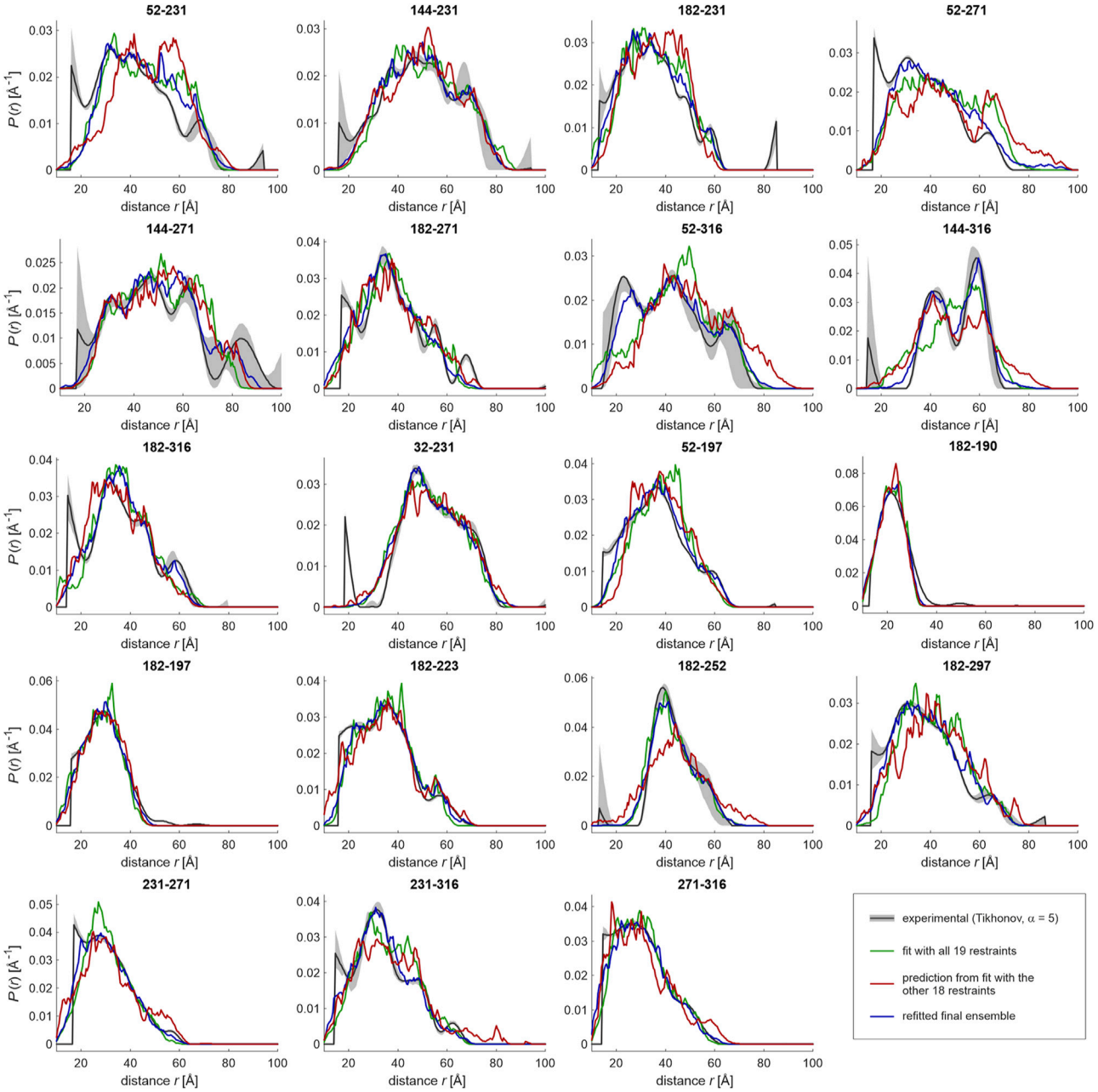


FIGURE 5 | Ensemble fitting of the LCD (residues 188–320) of hnRNP A1 with jack-knife resampling. A raw ensemble of 331 conformers was generated based on Gaussian restraints. Conformer populations were fitted by maximizing geometric mean overlap between simulated (green) and experimental (black) distance distributions. Conformers with less than 1% of the population of the most populated conformer were discarded, resulting in a refined ensemble with 60 conformers. The procedure was repeated with all possible sets of 18 out of the 19 restraints, giving slightly different sizes of the raw and refined ensembles. The unused restraint was predicted from the "leave-one-out" ensemble (red). All conformers found in individual ensembles during jack-knifing were combined with the 60 conformers in the initial ensemble and used as the basis set for the final ensemble fit (blue lines).

differences than the appearance of the distributions. Therefore, we have studied this question in detail. To that end, we have performed ensemble fitting from the same raw ensemble for all individual sets of distance distributions displayed in **Figure 4**, as well as for the Gaussian distribution obtained by DeerLab (parameters listed in **Table 2**). As a result, we have obtained refined ensembles with a reduced number of conformers and with populations assigned to these conformers.

The overview of results in **Table 3** demonstrates that fit quality and size of the population-fitted ensembles vary only moderately between the different restraint sets, despite the apparent differences in the fitted distance distributions (i.e., the

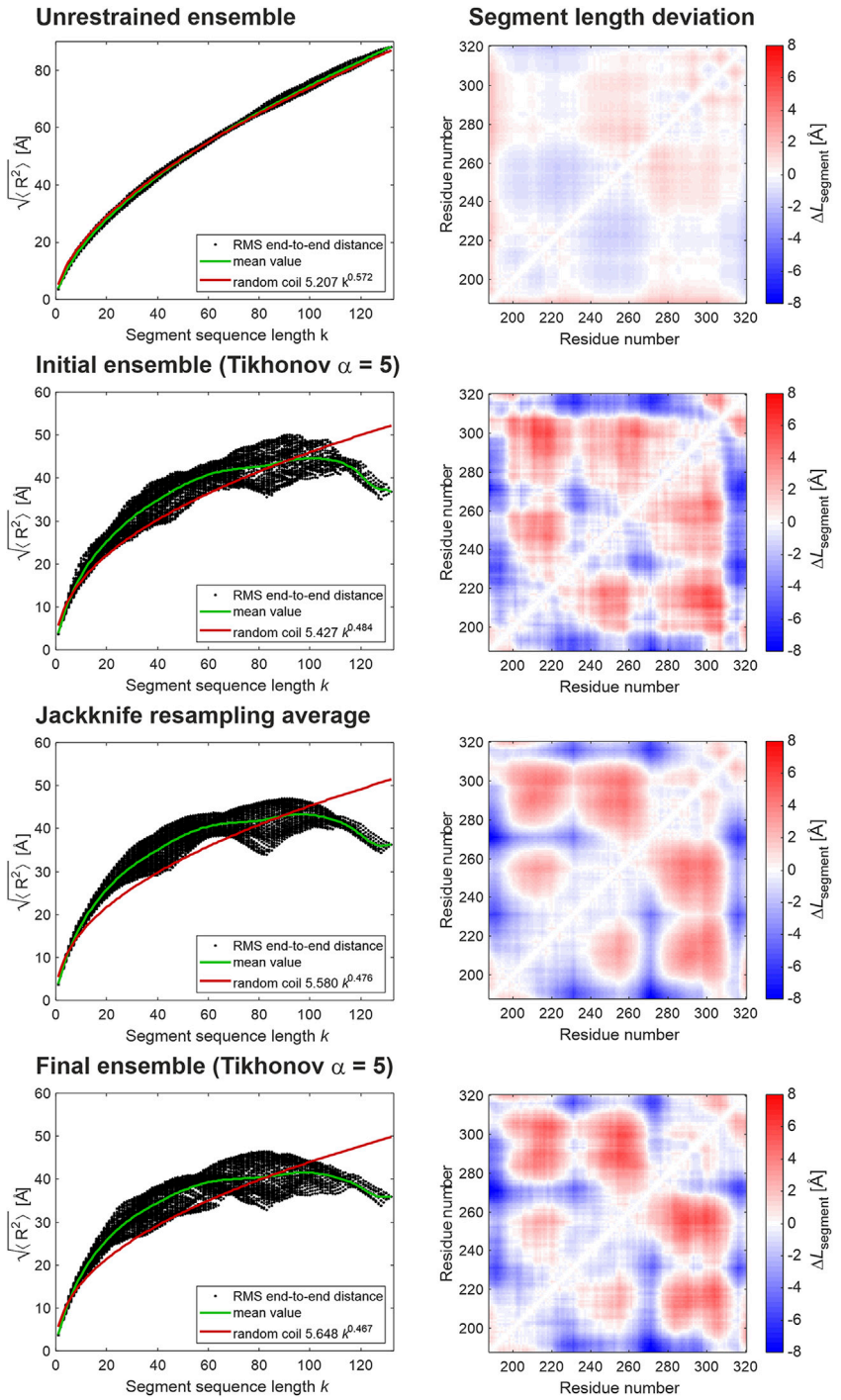


FIGURE 6 | Ensemble analysis for the LCD of hnRNP A1. The left column shows root mean square CA-CA distances for all possible segments of the LCD (residues 188–320) as a function of segment sequence length (black dots), their mean values per segment sequence length (green line), and a random coil fit (red line). The right column shows the deviation of the RMS CA-CA distances from the mean value for this segment sequence length (green line in the left column). Red hues correspond to segments more extended than the average and blue hues to segments more compact than the average.

resolution of spikes and peaks) obtained with the different approaches (see **Figure 4**). As an example for fit quality, results are shown in **Figure 5** for the case of Tikhonov regularization with fixed regularization parameter α = 5, with the experimental distributions as black lines with gray uncertainty bands and the prediction from the fitted ensemble as green lines. We note that in some cases, features of the distribution shape are fitted that cannot be reproduced by a Gaussian, most notably for

site pairs 32–231, 182–223, and 182–252. Thus, reweighting of conformers after generation of the raw ensemble can partially recover underfitting of structural information in the input single Gaussian model. Still, geometric mean overlap $\overline{o}$ is maximum for the ensemble fit to the single Gaussian distributions obtained in DeerLab.

We have also compared the ensembles obtained with restraints sets from the six data analysis approaches. Such comparison is complicated by the fact that the sets of selected conformers are distinct and the domain is only moderately structured. Therefore, we have opted for ensemble analysis approaches that can reveal weak structure. First, we considered segment-wise RMS end-to-end distances (Jeschke, 2020) as displayed for FUS in **Figures 1B**, **3**. The corresponding plots for hnRNP A1 LCD are shown in the left column of **Figure 6**. Whereas the unrestrained ensemble (top) exhibits similar behavior as the FUS NTD ensembles, for the restrained ensemble model of hnRNP A1 (second from top), scaling of segment end-to-end distances is less regular and the distributions are broader. While the patterns obtained with different data analysis approaches (**Supplementary Figures S4, S5**, left column) are not identical, they do exhibit very similar features. In particular, with all approaches the LCD of hnRNP A1 exhibits scaling behavior that decidedly differs from the one of the unrestrained ensemble. In the restrained ensembles, segment-wise end-to-end distances do not increase monotonously, but rather decline at segment sequence lengths $k > 100$. Furthermore, segment compactness varies considerably at the same sequence length in the range $40 < k < 100$. In order to analyze this variation in more detail, we use a measure related to the proximity matrix introduced in (Jeschke, 2021). The proximity matrix is related to, but distinct from, the contact matrix or map that can be obtained by NMR techniques, such as paramagnetic relaxation enhancement (Salmon et al., 2010; Clore, 2013). Unlike the contact map, which reveals close approach of residues with large sequence distance, the proximity matrix quantifies the deviation of the RMS CA-CA distance of residue pairs from the one predicted for that segment length k from the random-coil scaling law $b{\cdot}k^{\nu}$, where parameters b and ν are fitted to all segments of the chain. This deviation is normalized to the predicted RMS CA-CA distance. Thus, the proximity matrix also reveals compaction of chain segments that does not lead to contact. For the case at hand, we consider a more intuitively interpretable matrix $\boldsymbol{\Delta L}$, whose elements are the absolute deviation of the RMS CA-CA distance of residue pairs (i,j) from the mean RMS CA-CA distance for all segments of the same length k (green line in the right column in **Figure 6**). The matrix elements ΔL_{ij} for residue pairs (i,j) are then defined as

$$\Delta L_{ij} = \sqrt{\langle R_{ij}^2 \rangle} - \frac{1}{n_k} \sum_{m-l=k} \sqrt{\langle R_{lm}^2 \rangle} \tag{2}$$

where n_k is the number of residue pairs (l,m) with $k = j–i = m–l$ and $\langle \cdots \rangle$ denotes the ensemble average. This matrix is visualized in **Figure 6** (right column) with blue color encoding segments shorter than average and red color encoding those longer than average. The color scale is the same for all ensembles.

In the restrained ensembles (second row from top), we find a pattern of locally shortened and lengthened segments that is broadly replicated with all distance analysis approaches (right column in **Supplementary Figures S4, S5**), although details differ. This pattern strongly differs from the unrestrained reference state (top row), which exhibits behavior that is very similar to the one of the reference state for FUS NTD.

Likewise, radii of gyration of the LCD are very similar (minimum 23.3 Å for multi-Gaussian distributions, maximum 24.0 Å for DeerNet) as are ensemble widths

$$\Gamma = \sqrt{\frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_i p_j D_{ij}^2}{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_i p_j}} \tag{3}$$

where indices i and j run over conformers in the ensemble, the p_i and p_j are conformer populations, and the D_{ij} are RMS coordinate deviations between conformers upon optimal superposition. We find that Γ varies between 46.6 Å and 48.4 Å (**Table 3**). In conclusion, ensemble fitting to distance distributions by maximization of overlap is rather robust with respect to moderate variation in distribution shape.

Validation of the Restraint Set by Jack-Knife Resampling

For atomic-resolution structures, it is relatively well understood what constitutes a good or at least a sufficient restraint set. The same cannot be said for ensemble modeling of weakly structured proteins. On the one hand, the problem appears hopelessly underdetermined. Backbone conformation of each individual conformer with n_{res} residues is determined by $2(n_{\mathrm{res}}\text{-}1)$ torsion angles, which would suggest 262 free parameters per conformer for the glycine-rich domain of hnRNP A1 (188–320), whereas we have obtained only 19 distance distribution restraints. On the other hand, the unstructured reference state is characterized by only one (radius of gyration R_{g} or RMS end-to-end distance R) or at most two (b, ν) parameters. As we do not know beforehand, how strongly a particular protein or domain is structured, we need to estimate the number of required restraints during modeling. Moreover, the restraint set might not be internally consistent. With labeling approaches, this may happen if a label biases conformation. We have encountered such a case for the FnIII-3,4 domains of integrin α6β4, were some spin-labelled mutants had to be discarded, as they caused strong changes in relative domain orientations as seen by changes in small-angle scattering (SAXS) curves (Alonso-García et al., 2015). In a computational study on amyloid-beta, it was found that attachment of a spin label biased the conformer distribution of this intrinsically disordered peptide to some extent (Sasmal et al., 2017). It is not always possible to safely exclude such bias by additional experiments. Hence, size and internal consistency of a set of distance distribution restraints need to be validated alongside modeling.

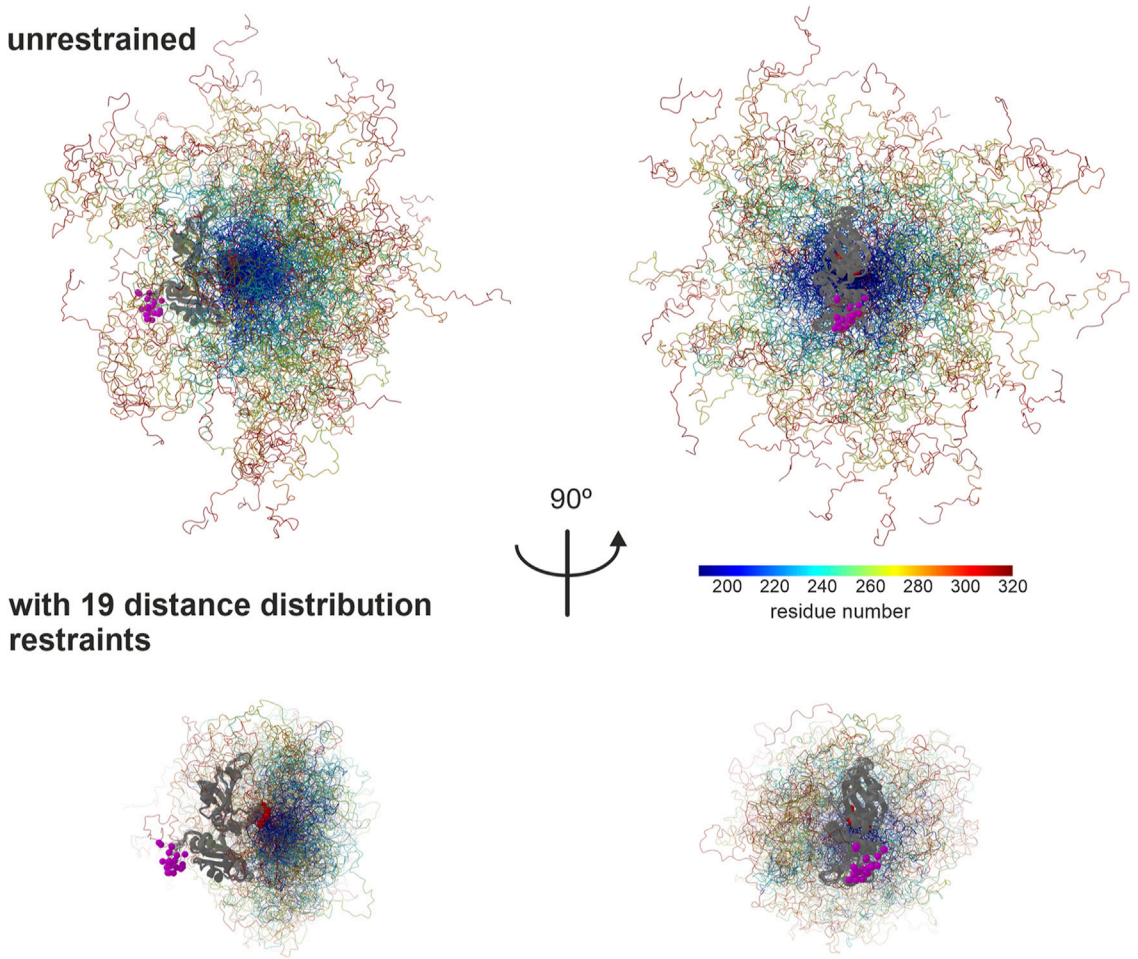


FIGURE 7 | Ensemble model for the LCD of hnRNP A1. The two RRMs (PDB 2lyv, Barraud and Allain, 2013) are shown in gray, with the N terminus to the left in the left panel and to the front in the right panel. RRM 1 is on the bottom and RRM 2 on the top. Residues in the LCD are rainbow color coded from blue (residue 188) to red (320). Population of conformers is encoded by coil thickness, with the most populated conformer having a thickness of 0.25 Å. CA atoms of the N-terminus of RRM 1 are shown as purple spheres with radius 1.5 Å and CA atoms of the C-terminus of RRM 2 are shown as maroon spheres with the same radius. Visualization by ChimeraX (Goddard et al., 2018) via an MMMx script.

Robustness of an ensemble can be estimated by resampling approaches, such as bootstrapping or jack-knifing (Berman et al., 2019). In jack-knifing, as many additional modeling runs are performed as there are restraints. In each run, one of the restraints is left out without replacement. It is tested, how strongly the ensemble changes and how well the left out restraint is predicted. Jack-knife resampling has been used before in the context of distance distribution restraints for estimating uncertainty of a high-resolution model of the dimer of Na^+/H^+ antiporter NhaA (Hilger et al., 2007). Bootstrapping approaches often remove more than one restraint and they replace left-out restraints by some remaining restraints, effectively increasing the weight of these doubly selected restraints.

Here, we implement jack-knife resampling in our ensemble modeling pipeline. To that end, we have generated raw ensembles of about 400 conformers each for all 19 restraint sets where one of the restraints is left out. We have then performed ensemble fitting with the same restraint left out and predicted the distance distribution for this unused restraint. These 19 predictions from 19 ensemble fitting runs are displayed as red lines in **Figure 5**. In general, they agree quite well with the experimental distributions, indicating that the set is, both, internally consistent and sufficiently large. Interestingly, while several restraints were measured as permutations of label combinations of major reporter sites, a few sites appeared only in a single restraint (folded: 32; LCD: 190, 197, 223, 252, and 297), and might thus be more critical for overall ensemble convergence. Indeed, the case of 182–252 is such a case of an 'isolated' restraint for which the deletion appears to lead to a slightly broader ensemble compared to the fit with all restraints. However, the predicted restraint fulfilment upon deletion of a restraint was not in general worse for isolated restraints than for restraints involving multiply restrained sites. Therefore, at least in this particular case of hnRNP A1 it appears that the coverage of labeling sites in the LCD sequence was sufficient to avoid modeling bias. However, it is clear that such consideration

must be made case-by-case, since the problem is strongly dependent on geometry and extent of ordering of domains in a given protein (Jeschke, 2016).

Returning to hnRNP A1, in a few cases, notably for site pairs 52–271, 52–316, 144–316, and 182–252, background separation may have removed genuine contributions at long distances. Except for pair 52–271, these distributions were flagged as being less reliable by disagreement between results from different data analyses approaches. In a few other cases, such as 182–271 and 32–231, the leave-one-out predictions appear to confirm distance distribution shapes that deviate from a single Gaussian.

We have combined all 19 leave-one-restraint-out ensembles and have renormalized populations to unity sum. From this "super-ensemble", we have computed segment-wize RMS end-to-end distances and the segment length deviation matrix (middle row in **Figure 6**). Because of the large size of the super-ensemble, these data are smoother. In general, they exhibit the same features as the initial ensemble obtained with all restraints. Unsurprisingly, the radius of gyration (23.7 Å) and width Γ = 48.2 Å, are unchanged.

While jack-knife resampling provides smooth estimates for segment-wise RMS end-to-end distances and the segment length deviation matrix derived from them, it does not directly provide an improved ensemble. However, during jack-knife resampling the modeling pipeline generates a large number of conformers that are consistent with at least n_r-1 of n_r restraints. During ensemble refinement and reduction, the approach selects conformers that best fit n_r-1 of n_r restraints. This provides a much improved basis set of conformers for fitting with all restraints. In our case, the 19 leave-one-restraint-out ensembles contain 1,119 conformers. Together with the 60 conformers in the initial ensemble, we have a basis that better samples conformational space and should thus allow for an improved fit. This expectation is indeed borne out, as can be seen by comparing the blue lines (fit with 1,179 pre-selected conformers) with the green lines (initial fit with 331 conformers) in **Figure 5**. In fact, the large raw ensemble may allow for some overfitting, i.e., the final ensemble may reproduce some features of the distance distributions that are uncertain, as can be seen by comparison with **Figure 4**. Overfitting of distance distributions is not easily quantified, as the distributions are solutions of an ill-posed problem. In future work, we will address this problem by considering fit quality of the primary data. In any case, the final ensemble fitted with the improved basis set provides smoother segment-wise RMS end-to-end distances and a smoother segment length deviation matrix than the initial ensemble. These characteristics are closer to the jack-knife mean estimates (bottom row in **Figure 6**). Therefore, we consider this ensemble with 129 conformers and mean overlap $\overline{o}$ = 0.919 as a better representation of the structure of the glycine-rich domain of hnRNP A1 than the initial fit. We note that jack-knife resampling generally provides a large number of conformers that are consistent with at least n_r-1 of n_r distance distribution restraints. Since finding such conformers is computationally more expensive than ensemble reweighting, final computation of a "super-ensemble" is advantageous.

This final ensemble is visualized in **Figure 7** together with an ensemble with the same number of 129 conformers randomly selected from an unrestrained raw ensemble with 2,146 conformers. For the unrestrained ensemble, we assumed uniform populations of the conformers. The restrained ensemble is much more compact. Residues 188–240 are located on the side of the two RRMs away from the N terminus, to a larger extent than in the unrestrained ensemble. Beyond residue 240, in the restrained ensemble the chain tends to backtrack toward the RRM, leaving the N-terminal side of the RRMs, but not the opposite side exposed. This exposure is also apparent in the unrestrained ensemble, where it is a purely geometrical effect. However, it is more pronounced in the restrained ensemble. Since the function of hnRNP A1 involves RNA binding, biological interpretation of this result requires further experiments including RNA binding, which are beyond the scope of this work and will be reported elsewhere. We may note here that, although a random-coil model obviously does not fit the weakly structured LCD of hnRNP A1 very well (**Figure 6**), it does predict scaling exponents $\nu \approx 0.48$, as they have earlier been observed for foldable proteins (Hofmann et al., 2012). Yet, the LCD is obviously not folded, as the distance distributions shown in **Figures 4**, **5** can be fitted only with a broad distribution of conformers.

DISCUSSION

Model for the Unstructured Reference State and Detection of Weak Structure

For peptide chains of 100 or more residues, residue-specific Ramachandran statistics for loop regions is nicely consistent with Flory random coil scaling of segment RMS end-to-end distances (**Figure 1B**). Unfortunately, this approximation is not very good for short sections of 20–25 residues (**Figure 3**), which appear to be stiffer than predicted by a self-avoiding random walk of a freely jointed chain model, irrespective of the solvent conditions and condensation state. Attempts to extract random-coil parameters from DEER distance distributions are further confounded by convolution of the backbone end-to-end distance distribution with the rotamer distribution of the spin label side chain. Both problems are expected to lessen if the section length is increased. However, such a strategy shifts the mean of the distribution from the most favored range for DEER (25–40 Å) toward longer distances. This in turn shifts contributions by the longest conformers to a range where separation of the single-chain contribution from the intermolecular background becomes uncertain.

Therefore, our approach for recognizing deviation from an unstructured reference state is hybrid. First, we generate a large raw ensemble of 1,000–5,000 conformers that is in broad agreement with experimental restraints that are approximated by single Gaussian distributions in this step. The raw ensemble is based on residue-specific Ramachandran statistics for loop regions that is biased only by the experimental restraints. No residue-residue interaction potential is assumed. Second, we refine and contract this raw ensemble by fitting populations of

conformers and by discarding conformers with very low population. In this step, we fit to full distance distributions that can have any shape. Third, we analyze scaling of segment-wise RMS end-to-end distances R_k with segment sequence length k and distribution of the R_k for given k. Broad distributions of the R_k and non-monotonic scaling of the mean R_k with k reveal deviations from random-coil behavior. These deviations can be heterogeneous along the sequence. We can map them to residue pairs by computing a segment length deviation matrix, whose elements are defined by **Eq. 2**. Indeed, based on our analysis of the full length hnRNP A1 construct, which contains both a strongly and a weakly ordered domain, we demonstrate that the segment analysis approach cannot only globally, but also locally uncover LCD structural deviations from random coil behavior.

This approach does not make any assumptions on the LCD beyond Ramachandran statistics. Even that assumption can, to some extent, be altered by the experimental restraints. With the numbers of conformers indicated above, it appears to be possible to fit experimental DEER distance distributions within their uncertainty for unstructured and weakly structured domains of 100–150 residues length on a current desktop computer. For longer domains or constructs, the number of required conformers in the raw ensemble may need to be reassessed or larger computational resources are required. It is also expected that the more structured a domain is, the more conformers need to be sampled in order to obtain a representative ensemble. In the limit of highly structured domains, the approach is not expected to be competitive with strategies that estimate variation from atomic-resolution structure.

Robustness of Distance Distribution Restraints

The mapping of dipolar signals to distance distributions is not continuous in a mathematical sense of the term. Accordingly, computation of distance distributions corresponds to solving an ill-posed problem. The solution must be stabilized in some way, for instance, by regularization (Bowman et al., 2004; Jeschke et al., 2004; Chiang et al., 2005), by restricting it to a space of parametrized distribution functions (Stein et al., 2015), or by training a neural network for a restrained set of distance distributions (Worswick et al., 2018). Especially for weakly ordered or unstructured domains, fits by a single Gaussian function appear to work surprisingly well, as demonstrated here. In general, different analysis approaches provide distributions that may differ beyond their own uncertainty estimates (**Figure 4**). This indicates that model bias is a matter of concern. We find that differences between results from different approaches are minor if distributions are only moderately broadened and well within the preferred distance range of DEER. They can be large if the width is several tens of Ångström or if background separation becomes uncertain.

Ensemble fitting by distance distribution restraints is stabilized by the overlap criterion, as overlap of an ensemble-predicted distribution with the various distributions computed from the same experimental data varies much less than the shape of the various experiment-derived distributions. This finding applies to minor deviations between distance distributions computed by different approaches, as they are seen in **Figure 4**. If there exist major differences between distance distributions obtained by different approaches, data is of poor quality and should not be used for ensemble fitting. If in doubt, it may be prudent to perform ensemble modeling with distance distribution restraint sets derived by alternative approaches, as we have demonstrated here for the glycine-rich domain of hnRNP A1. In any case, we recommend to process experimental data sets by three different approaches: Tikhonov regularization, multi-Gaussian fitting, and neural network analysis. As seen in **Figure 4**, such comparison can reveal restraint uncertainty better than merely computation of uncertainty bands for a single data analysis approach. Comparison of the three distributions also reveals whether single-Gaussian restraints are a sensible choice for final ensemble fitting, as appears to be the case for hnRNP A1 188–320. This question is important, as fitting of a single Gaussian, on the one hand, is the most robust approach, but on the other hand, runs the largest risk of model bias.

For hnRNP A1, we have found that 19 distance distribution restraints suffice for stably characterizing substantial deviation of a 133-residue segment from random-coil behavior. Depending on how large such deviations are, the number of required distance distribution restraints may vary. Jack-knife resampling provides a general approach for testing whether a restraint set is sufficient for a given problem. If it is not, further restraints need to be added. For the case of hnRNP A1, we cannot exclude that less than 19 restraints would have sufficed. Establishing this would require a systematic analysis of how many and which restraints can be left out without substantially changing the proximity matrix. Such an analysis is computationally very expensive and beyond the scope of the current study.

Resampling for Ensemble Model Validation

The number of distance distribution restraints n_{res} that enter into an ensemble model is expected to be in the range between 10 and 50. With much less than 10 restraints, it is unlikely that structure can be revealed except, perhaps, for very short peptides. Measuring much more than 50 restraints appears to be unrealistic because of the effort required in sample preparation. In this situation, jack-knife resampling appears to be a viable approach for ensemble validation, as we have demonstrated here for n_{res} = 19 on hnRNP A1 188–320. Jack-knife resampling requires n_{res} + 1 modeling runs, one of them with all restraints and n_{res} runs where one of the restraints is left out. We have shown that these modeling runs provide an improved basis set of conformers that can be used as input for a final ensemble fit. Regarding computational effort, such jack-knife resampling is certainly feasible for domains with up to 150 residues and for up to 50 restraints on current desktop computers. Further automation in software packages is required to make jack-knife resampling convenient. This will be pursued in further development of MMMx.

Choice of Labeling Sites

For characterizing the LCD of hnRNP A1, we relied on three reference sites in the structured domain of the protein. Such a choice is advantageous in cases where a structured domain

exists and where the localization of the intrinsically disordered domain with respect to the structured domain is of interest. Could we have characterized weak order of the LCD on its own, provided that it would have been the same as in the presence of the RRM? To answer this question, we performed a computational experiment that assumed the final ensemble of our study as ground truth. In addition to the three experimental intra-LCD restraints (231–271, 231–316, and 271–316) we simulated distance distributions for 16 additional intra-LCD site pairs. These pairs were selected from LCD sites that we had also labeled in our experiments (190, 197, 223, 231, 252, 271, 297, and 316) by limiting sequence distance to no more than 85 residues, i.e., the maximum sequence separation within the LCD in our experimental restraints (**Supplementary Table S1**). With these restraints, we generated a raw ensemble of 1,290 conformers and then performed ensemble reweighting with the EnsembleFit module of MMMx. The final ensemble fits the 16 simulated and three experimental restraints with an overlap deficiency of 0.047. As seen in **Supplementary Figure S6**, site-resolved compaction of the LCD is reasonably well reproduced with this set of restraints, with notable deviations near the C terminus. In our original restraint set, residue 316 is localized with respect to the RRM by three restraints, whereas in the simulated restraint set, due to our chosen maximum cutoff segment length of 85 residues, it is not localized with respect to any site upstream of residue 231. This suggests that some site pairs with long sequence separation must be included in the set in order to avoid that structural features are missed.

CONCLUSION

Ensemble modeling and refinement with label-to-label distance distributions measured by EPR pulsed dipolar spectroscopy is a feasible approach for characterizing weak structure in protein domains. Our approach involves generation of a basis set of conformers that is consistent with Ramachandran statistics for loop residues and with restraints. This raw ensemble is refined and contracted by fitting to distance distributions restraints. It is then analyzed in terms of segment-wize RMS end-to-end distributions. For estimating uncertainty and validating restraints, we recommend computation of distance distributions from primary data by several alternative approaches and jack-knife resampling of the restraints in ensemble modeling. The current implementation is viable up to about 150 residues and up to about 50 distance distribution restraints on current desktop computers. Extension of these limits appears to be feasible. Our approach could be extended to integrative modeling that uses restraints from further experimental techniques in ensemble refinement.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

IR prepared hnRNP A1 samples, measured most hnRNP A1 DEER data, performed most of the hnRNP A1 DEER data analysis, and contributed to conception of the modeling and data analysis approaches. LE-H prepared FUS samples, measured FUS DEER data, performed all FUS DEER data analysis, contributed to conception of the data analysis approaches and implemented the SAW-ν model in DeerLab. EL performed preparatory work and provided guidance on hnRNP A1 sample preparation. LE optimized conditions for liquid-liquid phase separation of FUS and provided guidance. MY provided guidance and measured some of the hnRNP A1 DEER data. FA was involved in conception of the whole project and provided guidance. GJ was involved in conception of the whole project, implemented modeling and data analysis approaches into MMM and MMMx, performed part of hnRNP A1 DEER data analysis, ensemble modeling, jack-knife resampling, and ensemble analysis, provided guidance and wrote the first version of the manuscript. All authors contributed to the final version of the manuscript.

FUNDING

This work was funded by Swiss National Science Foundation grants 200020_188467 (hnRNP A1 and modeling aspects) and CRSII5_170976 (FUS).

ACKNOWLEDGMENTS

We thank Christian Altenbach for analyzing data with LongDistances, Eric Hustedt for analyzing data with DD, and both of them as well as Matthias Elgeti, Luis Fábregas Ibáñez, Ilya Kuprov, Daniel Nettels, and Ben Schuler for helpful discussions. This project also profited from discussions in a consortium on a white paper on measuring, analyzing, and interpreting distance distributions. In this context, we are grateful to Marina Bennati, Enrica Bordignon, Jack Freed, Daniella Goldfarb, Hassane Mchaourab, Thomas Prisner, Olav Schiemann, and Stefan Stoll. We thank a reviewer for a question that inspired the computational experiments described in *Choice of Labeling Sites*.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.636599/full#supplementary-material.

REFERENCES

Alonso-García, N., García-Rubio, I., Manso, J. A., Buey, R. M., Urien, H., Sonnenberg, A., et al. (2015). Combination of X-ray crystallography, SAXS and DEER to obtain the structure of the FnIII-3,4 domains of integrin α6β4. *Acta Cryst. D Biol. Crystallogr.* 71, 969–985. doi:10.1107/S1399004715002485

Barraud, P., and Allain, F. H.-T. (2013). Solution structure of the two RNA recognition motifs of hnRNP A1 using segmental isotope labeling: how the relative orientation between RRMs influences the nucleic acid binding topology. *J. Biomol. NMR* 55, 119–138. doi:10.1007/s10858-012-9696-4

Berman, H. M., Adams, P. D., Bonvin, A. A., Burley, S. K., Carragher, B., Chiu, W., et al. (2019). Federating structural models and data: outcomes from A workshop on archiving integrative structures. *Structure* 27, 1745–1759. doi:10.1016/j.str.2019.11.002

Bernadó, P., Mylonas, E., Petoukhov, M. V., Blackledge, M., and Svergun, D. I. (2007). Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.* 129, 5656–5664. doi:10.1021/ja069124n

Bonomi, M., Heller, G. T., Camilloni, C., and Vendruscolo, M. (2017). Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* 42, 106–116. doi:10.1016/j.sbi.2016.12.004

Bowman, M. K., Maryasov, A. G., Kim, N., and DeRose, V. J. (2004). Visualization of distance distribution from pulsed double electron-electron resonance data. *Appl. Magn. Reson.* 26, 23–39. doi:10.1007/BF03166560

Brangwynne, C. P., Tompa, P., and Pappu, R. V. (2015). Polymer physics of intracellular phase transitions. *Nat. Phys.* 11, 899–904. doi:10.1038/NPHYS3532

Cesari, A., Reisser, S., and Bussi, G. (2018). Using the maximum entropy principle to combine simulations and solution experiments. *Computation* 6, 15. doi:10.1016/j.sbi.2016.12.00410.3390/computation6010015

Chiang, Y.-W., Borbat, P. P., and Freed, J. H. (2005). The determination of pair distance distributions by pulsed ESR using Tikhonov regularization. *J. Magn. Reson.* 172, 279–295. doi:10.1016/j.jmr.2004.10.012

Clore, G. M. (2013). Generating accurate contact maps of transient long-range interactions in intrinsically disordered proteins by paramagnetic relaxation enhancement. *Biophys. J.* 104, 1635–1636. doi:10.1016/j.bpj.2013.01.060

Drescher, M. (2011). EPR in protein science. *Top. Curr. Chem.* 321, 91–119. doi:10.1007/128_2011_235

Edwards, T. H., and Stoll, S. (2018). Optimal Tikhonov regularization for DEER spectroscopy. *J. Magn. Reson.* 288, 58–68. doi:10.1016/j.jmr.2018.01.021

Emmanouilidis, L., Esteban-Hofer, L., Damberger, F. F., de Vries, T., Nguyen, C. K. X., Fábregas Ibáñez, L., et al. (2021). NMR and EPR reveal a compaction of FUS upon condensation to droplets formation. *Nat. Chem. Biol.* [Epub ahead of print]. doi:10.1038/s41589-021-00752-3

Fábregas Ibáñez, L., Jeschke, G., and Stoll, S. (2020). DeerLab: a comprehensive software package for analyzing dipolar electron paramagnetic resonance spectroscopy data. *Magn. Reson.* 1, 209–224. doi:10.5194/mr-1-209-2020

Fitzkee, N. C., and Rose, G. D. (2004). Reassessing random-coil statistics in unfolded proteins. *Proc. Natl. Acad. Sci. U.S.A.* 101, 12497–12502. doi:10.1073/pnas.0404236101

Geist, L., Henen, M. A., Haiderer, S., Schwarz, T. C., Kurzbach, D., Zawadzka-Kazimierczuk, A., et al. (2013). Protonation-dependent conformational variability of intrinsically disordered proteins. *Protein Sci.* 22, 1196–1205. doi:10.1002/pro.2304

Gigli, L., Andrałojć, W., Dalaloyan, A., Parigi, G., Ravera, E., Goldfarb, D., et al. (2018). Assessing protein conformational landscapes: integration of DEER data in Maximum Occurrence analysis. *Phys. Chem. Chem. Phys.* 20, 27429–27438. doi:10.1039/c8cp06195e

Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H., et al. (2018). UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* 27, 14–25. doi:10.1002/pro.3235

Godt, A., Schulte, M., Zimmermann, H., and Jeschke, G. (2006). How flexible are poly(para-phenyleneethynylene)s? *Angew. Chem.-Int. Edit.* 45, 7560–7564. doi:10.1002/anie.200602807

Habchi, J., Tompa, P., Longhi, S., and Uversky, V. N. (2014). Introducing protein intrinsic disorder. *Chem. Rev.* 114, 6561–6588. doi:10.1021/cr400514h

Hilger, D., Polyhach, Y., Padan, E., Jung, H., and Jeschke, G. (2007). High-resolution structure of a Na^+/H^+ antiporter dimer obtained by pulsed election paramagnetic resonance distance measurements. *Biophys. J.* 93, 3675–3683. doi:10.1529/biophysj.107.109769

Hofmann, H., Soranno, A., Borgia, A., Gast, K., Nettels, D., and Schuler, B. (2012). Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16155–16160. doi:10.1073/pnas.1207719109

Hovmöller, S., Zhou, T., and Ohlson, T. (2002). Conformations of amino acids in proteins. *Acta Crystallogr. D Biol. Cryst.* 58, 768–776. doi:10.1107/S0907444902003359

Jean-Philippe, J., Paz, S., and Caputi, M. (2013). hnRNP A1: the Swiss army knife of gene expression. *Int. J. Mol. Sci.* 14, 18999–19024. doi:10.3390/ijms140918999

Jeschke, G., Chechik, V., Ionita, P., Godt, A., Zimmermann, H., Banham, J., et al. (2006). DeerAnalysis 2006-a comprehensive software package for analyzing pulsed ELDOR data. *Appl. Magn. Reson.* 30, 473–498. doi:10.1007/BF03166213

Jeschke, G. (2012). DEER distance measurements on proteins. *Annu. Rev. Phys. Chem.* 63, 419–446. doi:10.1146/annurev-physchem-032511-143716

Jeschke, G. (2016). Ensemble models of proteins and protein domains based on distance distribution restraints. *Proteins* 84, 544–560. doi:10.1002/prot.25000

Jeschke, G., Koch, A., Jonas, U., and Godt, A. (2002). Direct conversion of EPR dipolar time evolution data to distance distributions. *J. Magn. Reson.* 155, 72–82. doi:10.1006/jmre.2001.2498

Jeschke, G. (2018). MMM: a toolbox for integrative structure modeling. *Protein Sci.* 27, 76–85. doi:10.1002/pro.3269

Jeschke, G. (2021). MMM: integrative ensemble modeling and ensemble analysis. *Protein Sci.* 30, 125–135. doi:10.1002/pro.3965

Jeschke, G., Panek, G., Godt, A., Bender, A., and Paulsen, H. (2004). Data analysis procedures for pulse ELDOR measurements of broad distance; distributions. *Appl. Magn. Reson.* 26, 223–244. doi:10.1007/BF03166574

Jeschke, G., Sajid, M., Schulte, M., Ramezanian, N., Volkov, A., Zimmermann, H., et al. (2010). Flexibility of shape-persistent molecular building blocks composed of p-phenylene and ethynylene units. *J. Am. Chem. Soc.* 132, 10107–10117. doi:10.1021/ja102983b

Kofinger, J., Stelzl, L. S., Reuter, K., Allande, C., Reichel, K., and Hummer, G. (2019). Efficient ensemble refinement by reweighting. *J. Chem. Theor. Comput.* 15, 3390–3401. doi:10.1021/acs.jctc.8b01231

Kurzbach, D., Vanas, A., Flamm, A. G., Tarnoczi, N., Kontaxis, G., Maltar-Strmecki, N., et al. (2016). Detection of correlated conformational fluctuations in intrinsically disordered proteins through paramagnetic relaxation interference. *Phys. Chem. Chem. Phys.* 18, 5753–5758. doi:10.1039/c5cp04858c

Liu, J., Perumal, N. B., Oldfield, C. J., Su, E. W., Uversky, V. N., and Dunker, A. K. (2006). Intrinsic disorder in transcription factors. *Biochemistry* 45, 6873–6888. doi:10.1021/bi0602718

Marsh, J. A., and Forman-Kay, J. D. (2012). Ensemble modeling of protein disordered states: experimental restraint contributions and validation. *Proteins* 80, 556–572. doi:10.1002/prot.23220

Polyhach, Y., Bordignon, E., and Jeschke, G. (2011). Rotamer libraries of spin labelled cysteines for protein studies. *Phys. Chem. Chem. Phys.* 13, 2356–2366. doi:10.1039/c0cp01865a

Polyhach, Y., and Jeschke, G. (2010). Prediction of favourable sites for spin labelling of proteins. *Spectr.-Int. J.* 24, 651–659. doi:10.1155/2010/706498

Ritsch, I. (2019). Distributions of molecular conformations and interactions revealed by EPR spectroscopy-methodology and application to hnRNPA1. PhD thesis. Zurich (Switzerland): ETH Zurich.

Salmon, L., Nodet, G., Ozenne, V., Yin, G., Jensen, M. R., Zweckstetter, M., et al. (2010). NMR characterization of long-range order in intrinsically disordered proteins. *J. Am. Chem. Soc.* 132, 8407–8418. doi:10.1021/ja101645g

Sasmal, S., Lincoff, J., and Head-Gordon, T. (2017). Effect of a paramagnetic spin label on the intrinsically disordered peptide ensemble of amyloid-beta. *Biophys. J.* 113, 1002–1011. doi:10.1016/j.bpj.2017.06.067

Schiemann, O., and Prisner, T. F. (2007). Long-range distance determinations in biomacromolecules by EPR spectroscopy. *Q. Rev. Biophys.* 40, 1–53. doi:10.1017/S003358350700460X

Stein, R. A., Beth, A. H., and Hustedt, E. J. (2015). A straightforward approach to the analysis of double electron-electron resonance data. *Methods Enzymol.* 563, 531–567. doi:10.1016/bs.mie.2015.07.031

Uversky, V. N. (2019). Intrinsically disordered proteins and their "mysterious" (Meta)Physics. *Front. Phys.* 7, 10. doi:10.3389/fphy.2019.00010

Uversky, V. N. (2017). Intrinsically disordered proteins in overcrowded milieu: membrane-less organelles, phase separation, and intrinsic disorder. *Curr. Opin. Struct. Biol.* 44, 18–30. doi:10.1016/j.sbi.2016.10.015

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2005). Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* 18, 343–384. doi:10.1002/jmr.747

Worswick, S. G., Spencer, J. A., Jeschke, G., and Kuprov, I. (2018). Deep neural network processing of DEER data. *Sci. Adv.* 4, eaat5218. doi:10.1126/sciadv.aat5218

Zheng, W., Zerze, G. H., Borgia, A., Mittal, J., Schuler, B., and Best, R. B. (2018). Inferring properties of disordered chains from FRET transfer efficiencies. *J. Chem. Phys.* 148, 123329. doi:10.1063/1.5006954

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 © 2021 Ritsch, Esteban-Hofer, Lehmann, Emmanouilidis, Yulikov, Allain and Jeschke. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 22 April 2021
doi: 10.3389/fmolb.2021.657946

The Effect of Cholesterol on Membrane-Bound Islet Amyloid Polypeptide

Mikkel Christensen[1,2,3], *Nils A. Berglund*[1] *and Birgit Schiøtt*[1,2]*

[1] Department of Chemistry, Aarhus University, Aarhus, Denmark, [2] Interdisciplinary Nanoscience Center (iNANO), Aarhus University, Aarhus, Denmark, [3] Sino-Danish Center for Education and Research, Beijing, China

Islet amyloid polypeptide (IAPP) is a proposed cause of the decreased beta-cell mass in patients with type-II diabetes. The molecular composition of the cell-membrane is important for regulating IAPP cytotoxicity and aggregation. Cholesterol is present at high concentrations in the pancreatic beta-cells, and in-vitro experiments have indicated that it affects the amyloid formation of IAPP either by direct interactions or by changing the properties of the membrane. In this study we apply atomistic, unbiased molecular dynamics simulations at a microsecond timescale to investigate the effect of cholesterol on membrane bound IAPP. Simulations were performed with various combinations of cholesterol, phosphatidylcholine (PC) and phosphatidylserine (PS) lipids. In all simulations, the helical structure of monomer IAPP was stabilized by the membrane. We found that cholesterol decreased the insertion depth of IAPP compared to pure phospholipid membranes, while PS lipids counteract the effect of cholesterol. The aggregation propensity has previously been proposed to correlate with the insertion depth of IAPP, which we found to decrease with the increased ordering of the lipids induced by cholesterol. Cholesterol is depleted in the vicinity of IAPP, and thus our results suggest that the effect of cholesterol is indirect.

Keywords: cholesterol, amylin, simulations, diabetes, aggregation, amyloid

OPEN ACCESS

Edited by:
Fabio Trovato,
Freie Universität Berlin, Germany

Reviewed by:
Birgit Strodel,
Julich-Forschungszentrum, Helmholtz-Verband Deutscher Forschungszentren (HZ), Germany
Carmelo La Rosa,
University of Catania, Italy
Yuguang Mu,
Nanyang Technological University, Singapore
Feng Ding,
Clemson University, United States

***Correspondence:**
Birgit Schiøtt
birgit@chem.au.dk

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *24 January 2021*
Accepted: *26 March 2021*
Published: *22 April 2021*

Citation:
Christensen M, Berglund NA and Schiøtt B (2021) The Effect of Cholesterol on Membrane-Bound Islet Amyloid Polypeptide. Front. Mol. Biosci. 8:657946. doi: 10.3389/fmolb.2021.657946

INTRODUCTION

Islet amyloid polypeptide (IAPP, also known as amylin) is co-secreted with insulin (Lukinius et al., 1989) from the beta cells of the pancreatic islets, and together with insulin it regulates the glucose metabolism (Lutz, 2010, 2012; Hay et al., 2015). In most patients with type-II diabetes mellitus (T2DM) IAPP aggregates to form islet amyloid (Maloy et al., 1981; Westermark, 1995; Westermark et al., 2011). IAPP aggregation is cytotoxic and contributes to the loss of beta cell associated with progressed T2DM (Abedini and Schmidt, 2013; Cao et al., 2013). While T2DM is initially characterized by a reduced insulin response, the loss of beta cells decreases the production of insulin and IAPP and thus further reduces the regulation of the glucose metabolism (Höppener et al., 2000). The cytotoxic mechanism of IAPP toward beta cells is unclear, but it has been hypothesized to be related to both membrane damage, inflammation, receptor interactions, oxidative stress, mitochondrial dysfunction, and inducing defects in autophagy (Abedini and Schmidt, 2013; Milardi et al., 2021).

Islet amyloid polypeptide is a 37 residue peptide with a net positive charge of +3 (N-terminal, Lys1 and Arg11) at physiological pH, as illustrated in **Figure 1A**. A histidine residue at position 18 can be protonated to carry a positive charge depending on the pH (Abedini and Raleigh, 2005). A disulfide bridge connects Cys2 and Cys7 and loops the N-terminus. IAPP is an intrinsically disordered peptide, meaning that it varies between multiple folded and unfolded conformations in solution as revealed by NMR experiment and CD spectroscopy (Williamson and Miranker, 2007; Williamson et al., 2009). At hydrophobic/hydrophilic interfaces (e.g., the interface of a phospholipid bilayer or a micelle) IAPP can stabilize in a conformation with amphiphilic α-helices (Knight et al., 2006). Based on structural ensembles solved from NMR experiments of micelle bound IAPP the α-helical structure of IAPP can be either straight (helical from residue 5 to 28) (Patil et al., 2009) or kinked (with a turn at residue Ser19 to Phe23, as sketched in **Figure 1B**) (Nanga et al., 2011). The helicity of membrane bound IAPP has also been observed on large unilamellar vesicles (LUVs) using circular dichroism (Jayasinghe and Langen, 2005; Knight et al., 2006; Engel, 2009; Khemtémourian et al., 2011; Lee et al., 2012).

Few structures of the mature fibrils are available; Wiltzius et al. (2008) has proposed a structure based on X-ray crystallography of two IAPP segments and Luca et al. (2007) has proposed a structure using constraints from solid-state nuclear magnetic resonance spectroscopy. Common for the fibril structures is the stacking of IAPP peptides along the fibril axis, with beta sheets forming between adjacent peptides. The individual peptides have a turn spanning from residue His18 to Phe23 (Luca et al., 2007).

Membrane damage induced by IAPP oligomers or amyloid is a proposed mechanism of cytotoxicity (Raleigh et al., 2017). Several studies have investigated membrane induced IAPP aggregation, since islet amyloid is located at the beta cell surfaces. The cell membrane is complex and contains many different lipids (Nicolson, 2014) and the aggregation is very dependent on the lipid composition, as indicated from observed lipid dependent changes in fibril morphology and the rate of formation (Sciacca et al., 2018). Dye-leakage and Thioflavin-T (ThT) assay fluorescence experiments are commonly used to study IAPP aggregation on model membranes to investigate the aggregation rate and membrane perforation of IAPP (Zhang et al., 2017). The aggregation rate and membrane perforation of IAPP is drastically increased in model systems with zwitterionic phosphatidylcholine (PC) lipids and anionic lipids such as phosphatidylglycerol (PG) and phosphatidylserine (PS) (Sciacca et al., 2008; Zhang et al., 2017). Working toward more realistic membrane models, several studies have recently focused on the interactions between IAPP and membrane cholesterol (CHOL) (Li et al., 2016; Zhang et al., 2017, 2018; Hao et al., 2018). CHOL is highly abundant in mammalian plasma membranes and it has been shown to be important for the function of many membrane associated peptides and proteins (Grouleff et al., 2015), transporting both ions and molecules e.g., the sarco/endoplasmic reticulum Ca^{2+}-ATPase (SERCA) (Li et al., 2004), the dopamine transporter (Zeppelin et al., 2018), aquaporins (Jacob et al., 1999), and peptides such as amyloid-β (Ji et al., 2002; Habchi et al., 2018). CHOL changes the fluidity and order of lipid membranes thereby indirectly affecting proteins that are embedded or attached to the membrane, or directly by interacting with the protein or peptide in a manner that affects structure and function (Grouleff et al., 2015; Zeppelin et al., 2018). Some experiments points toward an inhibiting role of CHOL toward IAPP aggregation, whereas other studies indicate that CHOL accelerate IAPP aggregation and cytotoxicity (Li et al., 2016; Zhang et al., 2017; Hao et al., 2018).

Zhang et al. (2017) studied the membrane interactions of IAPP in large unilamellar vesicles, probing amyloid formation using ThT fluorescence and membrane permeabilization from dye leakage experiments. They found that CHOL slows down amyloid formation and dye leakage in membranes consisting of 1,2-dioleoyl-*sn*-glycero-3-phosphatidylcholine (DOPC or PC), 10% 1,2-dioleoyl-*sn*-glycero-3-phosphatidylserine (DOPS or PS), and CHOL concentrations of 20 and 40%. They also showed that high concentrations of anionic DOPS lipids (above 25%) diminish the effect of CHOL (Zhang et al., 2017). In a follow-up study, Zhang et al. further investigated the effect of changing the type of sterol. They investigated the membrane binding, perforation, and amyloid formation on membranes with eight different types of sterols (Zhang et al., 2018). They noticed that for the fraction of vesicles that bound IAPP, the membrane leakage, and amyloid formation was inversely correlated with the ordering of the lipids in the vesicles. This points to the conclusion that the effect of CHOL is indirect, thus dictated by changing membrane properties, rather than by a direct interaction of CHOL with IAPP. In a study by Yang et al. the interaction between the 19 N-terminal residues of IAPP and small unilamellar vesicles (SUVs) consisting of DPPC and CHOL was studied. Using ^{31}P-NMR they estimated the dissociation constant of IAPP from the vesicle for different concentrations of CHOL, and found that the N-terminal fragment bound stronger to membranes with increasing concentrations of CHOL. Based on mutational studies they hypothesized a direct interaction between IAPP and CHOL by CH−π interactions between Phe15 and CHOL (Li et al., 2016). Sciacca et al. (2016) studied the effect of CHOL in

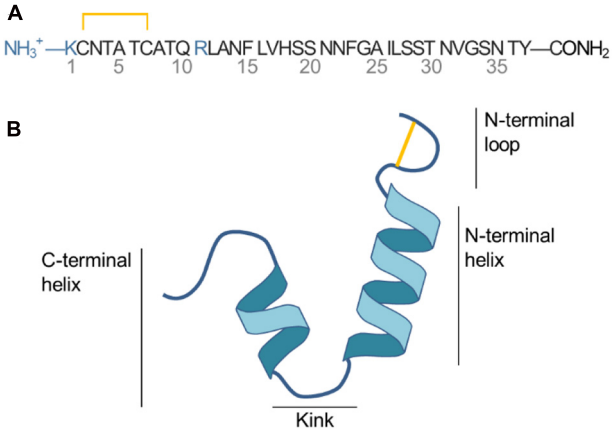


FIGURE 1 | (A) Sequence of IAPP. **(B)** Sketch of the secondary structure of IAPP in the helical kinked conformation.

LUVs with DOPC, DPPC, and CHOL in a ratio known to form lipid nanodomains with the formation of liquid ordered domains with saturated lipids and CHOL and liquid disordered domains containing unsaturated lipids. A DOPC:DPPC ratio of 1:2, CHOL accelerated amyloid formation and leakage with a CHOL content of 20 or 40%, compared to pure DOPC vesicles, as shown from ThT fluorescence and dye leakage experiments, respectively. Based on atomic force microscopy imaging they found that the disruption of the membrane occur at the boundary between the l_o and l_d domains (Sciacca et al., 2016). Several studies point to the same conclusion: Membranes with CHOL-induced domain formation accelerate IAPP aggregation and membrane damage, and IAPP will have a high affinity for the boundary between the l_o and l_d domain (Trikha and Jeremic, 2011; Caillon et al., 2014). Conversely, in membranes without domain formation, CHOL inhibits the aggregation and membrane damage induced by IAPP (Zhang et al., 2017).

Interactions between IAPP and phospholipid membranes have previously been investigated using MD simulations. Many of these studies have focused on simple membrane compositions composed of zwitterionic phospholipids and anionic lipids such as PG or PS (Dong et al., 2018). Simulation studies have previously been performed with CHOL, however, these studies have been coarse-grained, and thus not able to investigate the details of the interactions between IAPP and CHOL. In this study we apply atomistic molecular dynamics (MD) simulations to investigate the interaction between IAPP and lipid bilayers with various lipid compositions. With MD simulations, we achieve atomistic details about the interactions between IAPP and lipid bilayers with a mixture of CHOL, DOPC, and DOPS lipids. Further analysis of the effect of these lipids on membrane-bound IAPP are performed. Lipid rafts and the effect of these on membrane-bound IAPP will not be discussed further in this paper, since the focus herein is on the effect of CHOL on IAPP in membranes that do not form lipid rafts. Due to the limited timescale of MD simulations, the investigations are limited to the membrane-bound helical state of a monomer, and not the effect of the lipids on the oligomerization and resulting conformation transitions, which was recently described from simulations in membranes without CHOL (Skeby et al., 2016; Christensen et al., 2017).

MATERIALS AND METHODS

Islet amyloid polypeptide is expected to be helical in the membrane bound state (Jayasinghe and Langen, 2005; Knight et al., 2006; Engel, 2009; Khemtémourian et al., 2011; Lee et al., 2012). As for most other amphiphilic helices, crossing the lipid headgroup region involves an energy-barrier, and

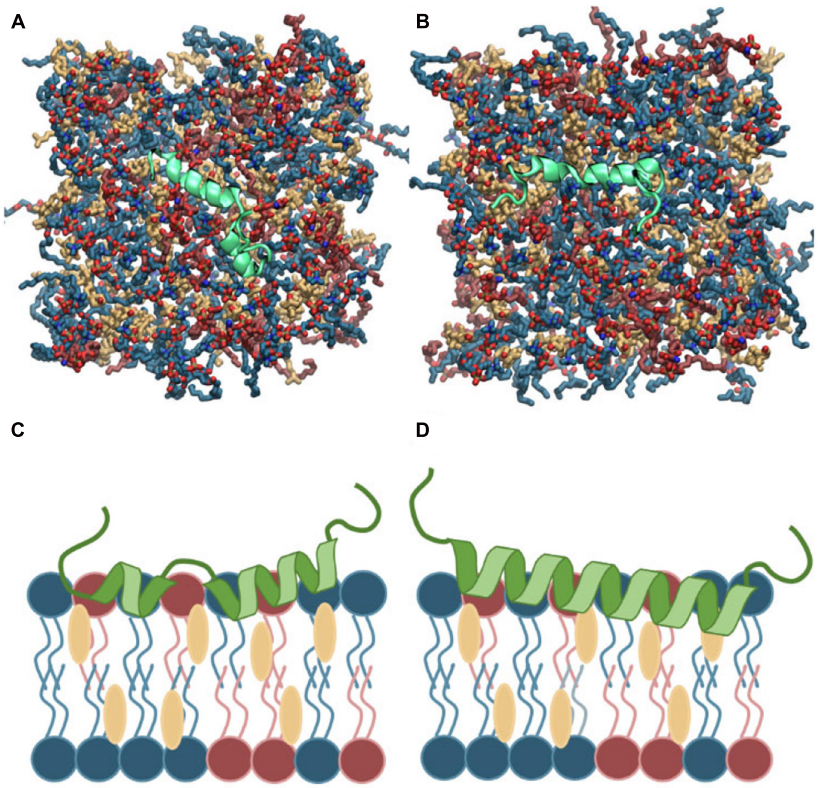


FIGURE 2 | (A,B) Top view of the atomistic starting structures of kinked and straight IAPP peptides in the membrane bound helical conformation. Water and ions are not shown. **(C,D)** Sketches of the membrane bound state of kinked and straight IAPP peptides shown from the side. The peptide is illustrated in green and DOPC, DOPS, and CHOL are shown in blue, red, and yellow, respectively.

simulating the binding is therefore computationally challenging (Ben-Tal et al., 1996). Coarse-grained MD simulations of a single IAPP peptide placed above lipid bilayers of each of the chosen bilayer compositions were performed initially to position IAPP at the membrane surface, while maintaining the helical conformation (**Supplementary Figure 1A**). Within 0.5 μs the peptides were bound to the bilayer, with the amphiphilic helices in the hydrophobic/hydrophilic interface of the bilayer (**Supplementary Figure 1B**). Crossing the headgroup region is a complex process, but in the coarse grained simulation the process is simplified and the amphiphilic helices can quickly cross this barrier and be positioned at the hydrophobic/hydrophilic interface. The membrane bound structures were subsequently converted to atomistic structures using the Martini Backward script (Wassenaar et al., 2014). To increase the sampling, two starting conformations of the peptides were applied; a straight conformation with a helical conformation spanning from residue 7 to 28, and a kinked conformation with residues 7 to 18 and residues 23 to 28, respectively, forming each an α-helix, with a turn between the two helical regions (as sketched in **Figures 2**, **3**). Both of these starting conformations are based on micelle bound structural ensembles of IAPP from NMR experiments (Patil et al., 2009; Nanga et al., 2011). This procedure was executed for different combinations of the lipids: DOPC, DOPS, and CHOL. Four repeats of 1 μs were produced for each of four membrane compositions: DOPC, DOPC/CHOL, DOPC/DOPS, and DOPC/DOPS/CHOL; and for each of the two initial peptide conformations (Straight/Kinked). An overview of the atomistic simulations can be found in **Table 1**. The membrane compositions are chosen to compare well with experimental data from the literature (Zhang et al., 2017). Examples of the DOPC/DOPS/CHOL setup is illustrated in **Figure 2**.

Coarse-Grained Simulations

The coarse-grained simulations were performed in Gromacs 2016.3 (Abraham et al., 2015) with the polarizable Martini 2.2P forcefield for protein and water (Yesylevskyy et al., 2010) and the lipids and ions modeled using the MARTINI 2.0 forcefield (Rzepiela et al., 2011; de Jong et al., 2013). The systems were built using the *Insane* script (Wassenaar et al., 2015), which can build lipid bilayers with the lipids randomly placed in each leaflet in the specified ratio and placing a peptide relative to the bilayer. The peptides were placed with the center of mass 40 Å from the membrane center and in a random orientation in each simulation repeat. The systems were minimized using steepest descent optimization, followed by a 50 ns equilibration with the number of particles, pressure, and temperature (NPT ensemble) kept constant, and with the protein constrained. The protein was

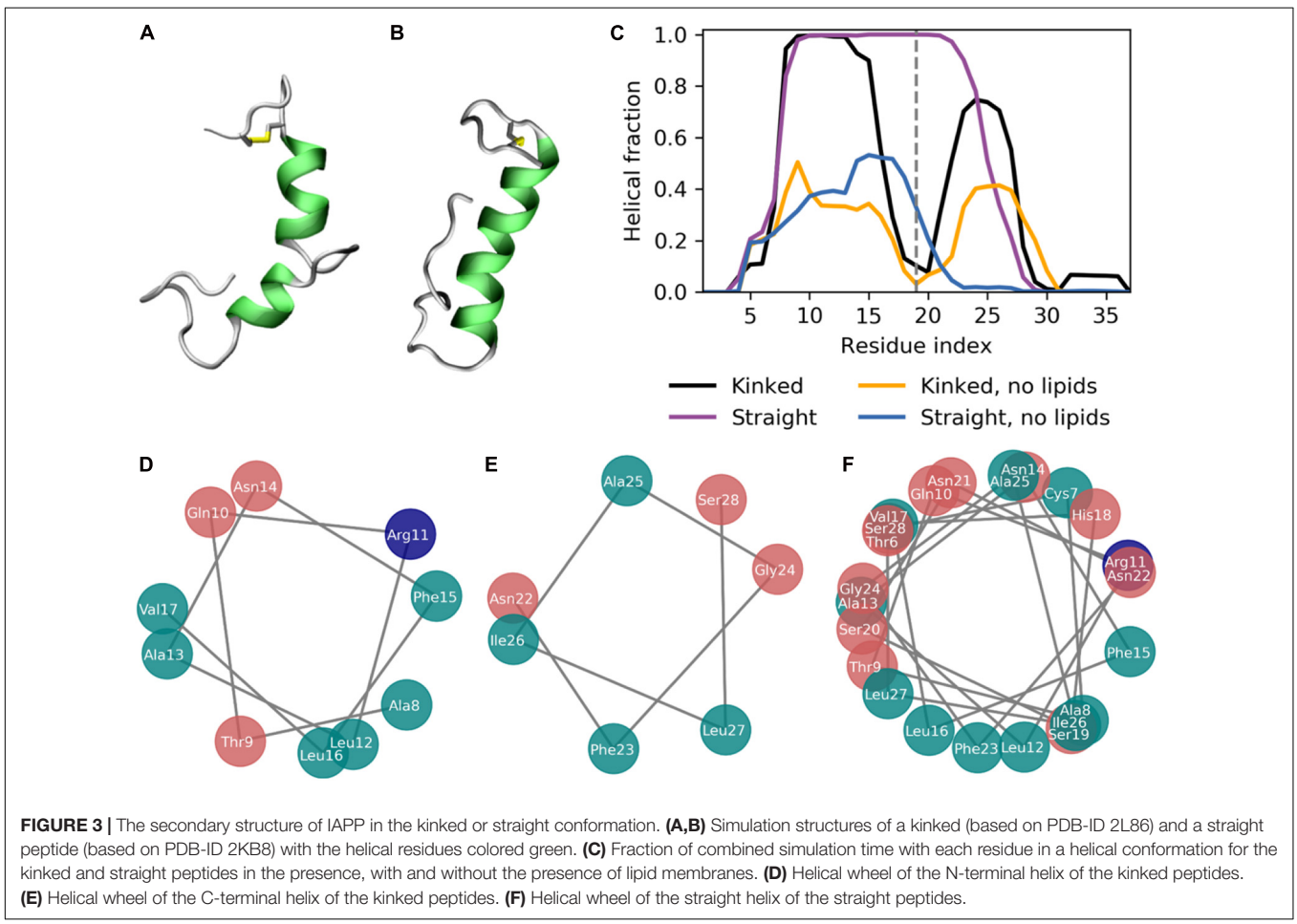


FIGURE 3 | The secondary structure of IAPP in the kinked or straight conformation. **(A,B)** Simulation structures of a kinked (based on PDB-ID 2L86) and a straight peptide (based on PDB-ID 2KB8) with the helical residues colored green. **(C)** Fraction of combined simulation time with each residue in a helical conformation for the kinked and straight peptides in the presence, with and without the presence of lipid membranes. **(D)** Helical wheel of the N-terminal helix of the kinked peptides. **(E)** Helical wheel of the C-terminal helix of the kinked peptides. **(F)** Helical wheel of the straight helix of the straight peptides.

released and a production run of 0.5 μs was performed in the NPT ensemble. The secondary structure of IAPP was restrained to the secondary structure of the starting structure, as is common practice in the MARTINI 2.0 forcefield (Rzepiela et al., 2011; de Jong et al., 2013). The flexible ends and kink region of IAPP were not constrained during the membrane binding simulations.

The leap-frog integrator (Hockney et al., 1974) was used with a timestep of 20 fs. During the NPT equilibration the Berendsen pressure coupling (Berendsen et al., 1984) was used with a time-constant of τ_p of 12 ps and reference pressure of 1 bar. The pressure coupling was applied semi-isotropically (x-y and z decoupled), and with an isothermal compressibility of $3\ \ 10^{-4}\text{bar}^{-1}$ for the system. In the production run the pressure was controlled with the Parrinello-Rahman barostat (Parrinello and Rahman, 1981) with a τ_p of 24 ps, a compressibility of $3\ \ 10^{-4}\text{bar}^{-1}$ and a reference pressure of 1 atm. The velocity rescaling algorithm was used to keep the temperature constant with a coupling constant τ_t of 1 ps. The coulomb interactions were treated using a reaction-field,(Tironi et al., 1995) with a cutoff of 1.1 nm and a relative electrostatic screening of 2.5 (de Jong et al., 2016). The van der Waals interactions are cut-off at 1.1 nm using a potential shift cutoff-scheme.

Atomistic Simulations

The atomistic simulations were performed in Gromacs 2016.3 (Abraham et al., 2015) which uses the leap-frog integrator (Hockney et al., 1974). The lipids, water, and ions were modeled using the CHARMM36 force field (Klauda et al., 2010), and proteins were modeled using the CHARMM36m force field (Huang et al., 2016), a recently developed force field with improved parameters for simulating intrinsically disordered peptides and the recommended force field for simulations of proteins and lipids from the CHARMM force field developers (Huang et al., 2016). To obtain a 2 fs timestep, all bonds to hydrogens are constraint using the LINCS algorithm (Hess et al., 1997). All histidine was modeled as the neutral δ-tautomer, which is known to allow binding of IAPP to the bilayer surface (Skeby et al., 2016).

The membrane bound structures returned from the Martini *backward* script (Wassenaar et al., 2014), were first minimized using the steepest descent algorithm with a maximum of 5,000 steps. The initial velocities were assigned randomly from a Maxwell-Boltzmann distribution. Then a 0.1 ns equilibration was performed in a canonical ensemble with number of particles, volume, and temperature constant (NVT ensemble). Next, a 3 ns equilibration was performed in the NPT ensemble. Finally, a 1 μs simulation production run was performed, also in the NPT ensemble.

The electrostatic interactions were calculated using Particle Mesh Ewald (Essmann et al., 1995) with a grid spacing of 0.1 nm and a short-range cutoff of 1.2 nm, using the Verlet cutoff-scheme. The van der Waals interactions were calculated with a force-switch modification (Steinbach and Brooks, 1994) that switch off the interaction between 1.0 and 1.2 nm. The temperature was controlled using the Nose-Hoover thermostat (Hoover, 1985), with a time constant of 1 ps, to keep the temperature at a constant temperature of 310K. A pressure of 1 atm was held using the Parrinello-Rahman pressure coupling (Parrinello and Rahman, 1981), with a time-constant of 5 ps, applied uniformly in the x-y dimension and separate in the z-dimension.

RESULTS

A series of simulations of membranes with DOPC, DOPS, and CHOL was performed to investigate the interactions between IAPP and the lipid components of the membrane and to study the membrane bound structure of IAPP and how it depends on the bilayer composition. The simulations are summarized in **Table 1** of the methods section.

The Helical Conformations Are Stabilized at the Membrane Interface

The simulations were initiated in the α-helical membrane bound state. The initial structure of the peptides in this study is helical peptides with a kink (**Figure 3A**) or with a single straight helix (**Figure 3B**). It can be seen from **Figure 3C**, that the helicity is stabilized by the membrane, likely due to the amphiphilic nature of the helices, with the hydrophobic residues buried in the membrane and the hydrophilic residues exposed to the solvent. The amphiphilicity of the peptide can be seen from the orientation of the hydrophobic and hydrophilic residues in the helical wheels in **Figures 3D–F** generated from a projection of the Cα atoms on the plane perpendicular to the helix axes. The residues Ala8, Leu12, Phe15, Leu16, Phe23, Ile26, and Leu27 are the hydrophobic residues of the helices, that are oriented toward the interior of the membrane. Without the presence of a lipid membrane, the α-helix is not stable, and from the helical fractions in **Figure 3C** it can be seen that the helices are unfolded most of the time.

TABLE 1 | Simulation overview of atomistic simulations with various membrane composition including DOPC, DOPS, and CHOL.

Membrane composition	Lipid ratio	Helix conformation	Number of repeats	Simulation time
–	–	Kinked	3	1 μs
		Straight	3	1 μs
DOPC	100%	–	1	1 μs
		Kinked	4	1 μs
		Straight	4	1 μs
DOPC:DOPS	70 – 30%	–	1	1 μs
		Kinked	4	1 μs
		Straight	4	1 μs
DOPC:CHOL	60 – 40%	–	1	1 μs
		Kinked	4	1 μs
		Straight	4	1 μs
DOPC:DOPS:CHOL	45 – 15 – 40%	–	1	1 μs
		Kinked	4	1 μs
		Straight	4	1 μs

The helicity fraction is described as the fraction of the accumulated simulation time each residue is helical (**Figure 3B**). For membrane inserted peptides, the energy barrier for conformational changes is very high, and is thus an event that occurs at a relatively long timescale (Ulmschneider and Ulmschneider, 2018), and therefore the helicity of IAPP starting in the kinked or straight will be analyzed separately. For the peptides starting in the kinked conformation, the helical fraction in the N-terminus peaks from Cys7 to Leu16, with decreasing fractions toward each end of the helix. In the C-terminus the helicity peaks from residue Asn21 to Ile28 (**Figure 2C**). The helical fraction of the N-terminus peaks at 1, meaning that some central residues are not seen to unfold at any point during the simulations. The C-terminal peaks at 0.8, indicating that all residues are unfolded at some point in the simulation series. The helicity varies less for the straight peptides than the kinked peptides. From the helicity fraction of the straight peptides in **Figure 3B** it can be seen that the helicity peaks from residue Asn3 to Ser28, with residues Thr9 to Phe23 helical throughout all simulation repeats. In the kinked peptides, the turn connecting the two helical segments consist of the 18-HSSNN-22 fragment, these are all hydrophilic residues and can therefore not contribute to the hydrophobic side of the helix. However, the turn region has been proposed to be important for the oligomerization of IAPP. Especially His18 and the charge state of this residue is found to be important for interactions between the IAPPs (Abedini and Raleigh, 2005; Christensen et al., 2017). In the straight peptides the kink region is tackled by having Ser20 and Ser19 on the hydrophobic side of the helix, oriented toward the sides of the helix, possibly allowing the hydrophilic sidechains to interact with the hydrophilic headgroups (**Figure 3F**).

The number of helical residues in the course of the simulations is shown in **Figure 4A**, separated for the N-terminal residues Lys1 to Ser19 and the C-terminal residues Ser20 to Tyr37, and when a lipid membrane is present. **Figure 4B** shows mean and standard deviation of the number of helical residues for the last 0.5 μs of the simulations. For the peptides initiated in a kinked conformations, the number of helical residues in the N-terminus remains around 10 and shows very little fluctuation in the simulations. The largest fluctuations in the helicity observed is in the C-terminus, as seen from **Figure 4A** only around five residues remain helical, and in some simulations the C-terminal helix completely unfolds. The peptides starting in the straight conformation have around 13 and 5 helical residues in the N-terminus and C-terminus, respectively. Both parts of the straight helix seems to be more stable than the two helical segments of the kinked peptides, as seen from the small error bars for the straight peptides in **Figure 4B**. It can be speculated that the stability of the straight peptides is most likely due to the fact that it has fewer helix ends where unfolding can occur, and an increase in intramolecular interactions arising from the secondary structure. There are no obvious differences in the stability of the helicity depending on the membrane composition. In the simulations of IAPP in solvent, with no lipid bilayer in the simulations, the helicity is less stable as seen from **Figures 4A,B**. The N-terminal helix is unfolded in all repeats and only in a single repeat does the C-terminal part of the straight helix remain helical.

Insertion of Peptide on the Membrane Is Affected by the Membrane Lipids and the Peptide Conformation

Figure 5 shows the average distance between each residue and the x/y-plane as measured as the average z-coordinate of the phospholipid phosphates, with z being the direction of the bilayer normal; the plane is used as an indication of the hydrophobic/hydrophilic interface. The simulation series are compared in pairs to show the effect of adding additional lipid types to the simulation system. In all simulations the hydrophilic residues of the helix are pointing away from the membrane while the hydrophobic residues are mostly pointing toward the membrane, giving rise to a zigzag-shaped curve. Most of the helical residues are placed below the hydrophobic/hydrophilic interface, indicating that the peptides are inserted in the membrane. Generally the highest variation in the membrane distance is seen in the non-helical region (residue 1 to 7 and residue 27 to 37), since these parts of the peptide are not usually inserted into the membrane.

The difference in the mean residue-membrane distances (Δmembrane distance) between the two simulation series in question is shown in **Figure 6**. The background color of the plots is an indication of whether the Δmembrane distance is significantly different from 0, using the *T*-test for comparing the means of two simulation series. The difference between pure PC and PC/PS membranes has a negligible effect on the membrane distance in both the kinked and the straight conformation (**Figures 6A,B**), as seen from a Δmembrane distance close to 0. This indicates that the membrane bound state of IAPP is well-defined in these simple phospholipid membranes. Addition of CHOL to a DOPC membrane increases the distance between the peptides and the membrane interface, and most significantly in the helical regions, as seen from an increase in Δmembrane distance of up to 4 Å in **Figures 6A,B**. The same is observed when comparing the membrane distance of a PC/PS membrane to a PC/CHOL membrane (**Figures 6A,B**). Interestingly, the peptides starting in the kinked conformation bound to a PC/PS and a PC/PS/CHOL membrane have the similar distance to the membrane interface (**Figure 6A**), this indicates that PS lipids keep the helices closer to the membrane, and thus counteracts the effect of CHOL. This effect is, however, not seen for the straight peptides, where the addition of CHOL still significantly increases the membrane distances compared to PC/PS membranes (**Figure 6B**).

There are general differences between the position of the kinked and the straight peptides. Comparing **Figures 5A,B**, it can be observed that the top of the helical segment of the straight peptides is positioned about 2 Å below the hydrophobic/hydrophilic interface in a pure phospholipid membrane, and the top of kinked helices in corresponding membranes are positioned at the level of the hydrophobic/hydrophilic interface. The membrane positions and the insertion for the straight and the kinked peptides

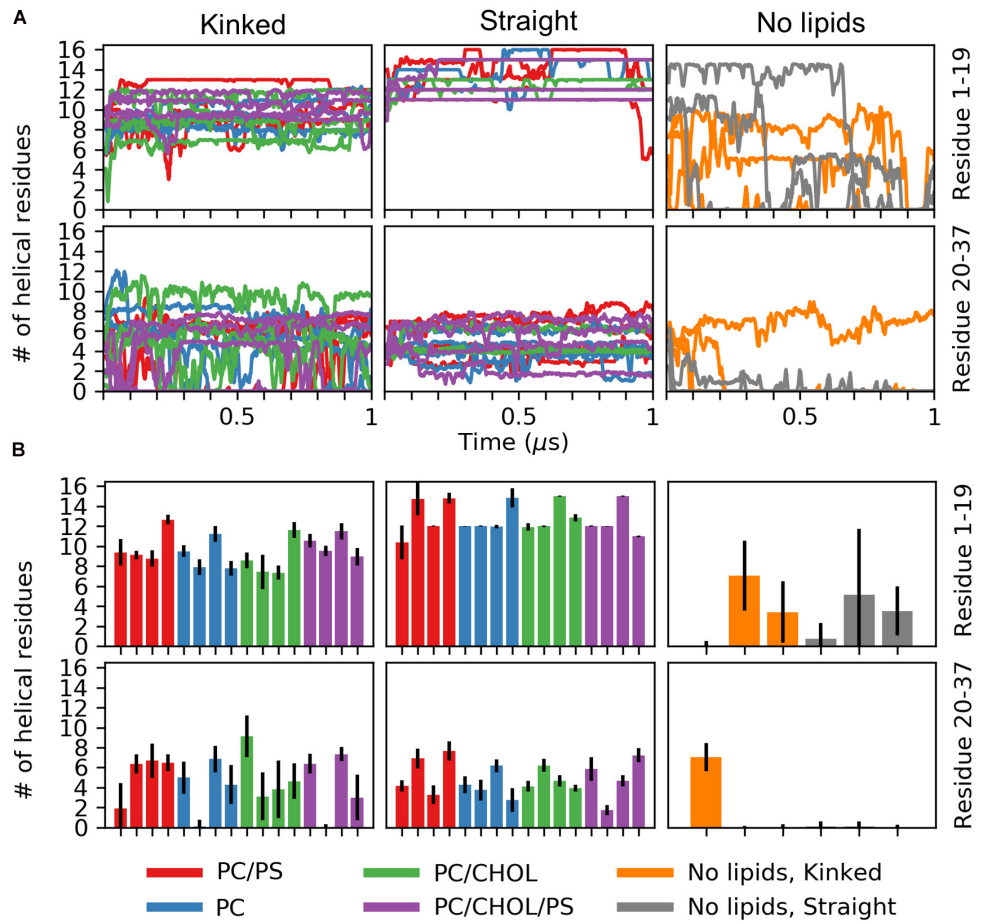


FIGURE 4 | Number of helical residues for each series of simulation (PC in blue, PC/CHOL in green, PC/PS in red, and PC/PS/CHOL in purple). The number of helical residues is plotted separately for the N-terminus (residue 1–19, first row) and the C-terminal (residue 20–37, second row) and separated for the peptides starting in the kinked conformation and the straight conformation. **(A)** The development of the number of residues in a helical conformation in time. **(B)** Mean and standard deviation of the number of helical residues in each of the simulation repeat.

are compared in **Figures 7A,B**. Not surprisingly, the turn region (residue 19–22) is positioned significantly deeper in all simulations with straight peptides when comparing to the kinked peptides. In the pure phospholipid membranes (pure DOPC or DOPC/DOPS), the conformation affects the position of the helical segments. In the DOPC membrane the N-terminal helix is inserted about 2 Å deeper for the peptide in the straight conformation, and most significantly for residues Gln10 and Asn14 (**Figure 7B**). In the DOPC/DOPS membranes both parts of the helix are inserted significantly deeper in the membrane by about 2 Å (**Figure 7B**). In the CHOL containing membranes there is no significant difference in the membrane distance, other than in the kink region, as seen in **Figure 7B**.

Cholesterol Increases the Order and Thickness of Lipid Bilayers

The difference in the position of IAPP on the membrane upon the addition of CHOL and/or DOPS can either be due to changes in the membrane fluidity or arise from direct interactions with the lipids. With the addition of CHOL to the lipid bilayers the order and rigidity is expected to increase (Gracià et al., 2010). The area per lipid was calculated using the *Membrainy* tool (Carr and MacPhee, 2015). From **Figure 8A** it can be seen that in the simulations with CHOL, the area per lipid is decreased from around 70Å^2 to ∼53Å^2. Changing from a composition of DOPC to DOPC/DOPS decreases the area per lipid slightly from around 70Å^2 to around 67Å^2. The lipid diffusion is also affected by the addition of CHOL (**Figure 8B**), in the DOPC bilayers the lipid diffusion is about $7 \cdot 10^{-8}$cm^2/s, whereas in a DOPC/CHOL bilayer the diffusion has dropped to half, to about $3.5 \cdot 10^{-8}$cm^2/s. The diffusion in a DOPC/DOPS bilayer is about $5.5 \cdot 10^{-8}$cm^2/s, slightly less than for a pure DOPC bilayer. The diffusion coefficients were calculated using the *gmx msd* tool of GROMACS 2018.1 (Abraham et al., 2015). The same pattern is observed for the order parameter (**Figure 8C**): CHOL increases the order parameter, and DOPC/DOPS membranes are also slightly more ordered than DOPC membranes. The order parameters were also calculated using the *Membrainy* tool (Carr and MacPhee, 2015). Another sign of the membranes

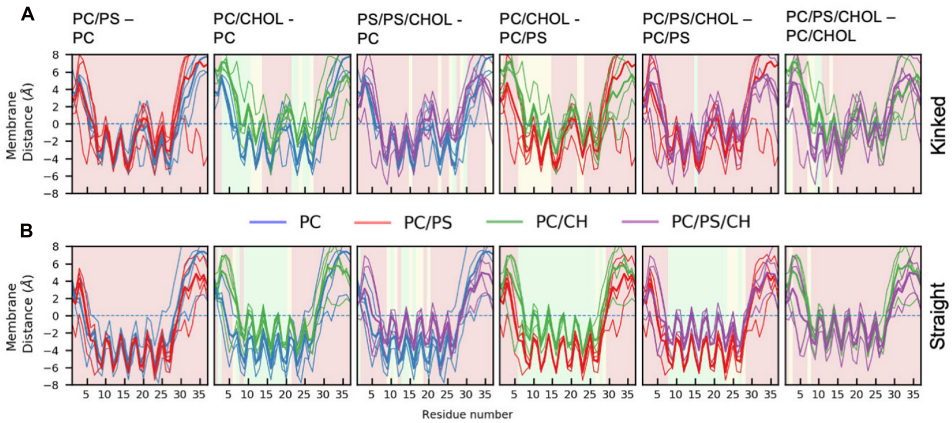


FIGURE 5 | Position of Cα of the peptide relative to the phosphate plane for. **(A)** The kinked conformation **(B)** straight conformation. Each panel shows the average distance between the IAPP Cα and the x-y plane spanned at the average z-coordinate of the phospholipid phosphates (indicated with a dashed line). The lines are colored based on the composition of the lipid membrane in the simulations. The average for the individual simulations are shown in thin lines and the common average in a thicker line. For the sake of comparison, two systems are shown simultaneously. The background color of the plots indicates whether the position on the membrane is significantly different between the two simulation series (green: $p \leq 0.05$, yellow: $0.05 < p < 0.10$, red: $p > 0.10$, as explained below).

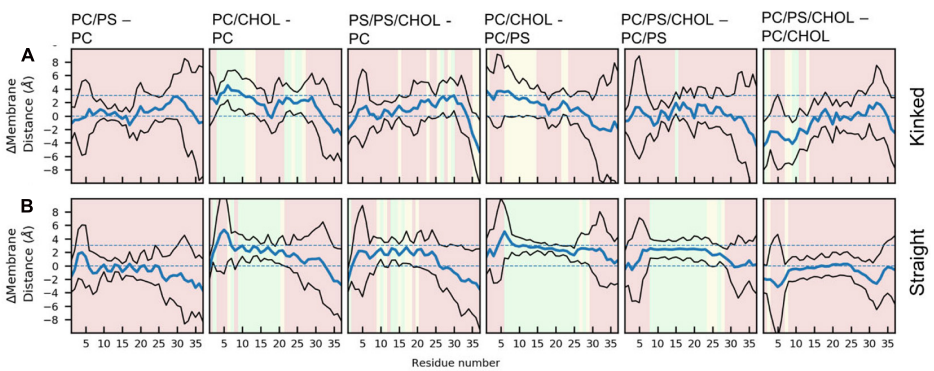


FIGURE 6 | Each panel shows the differences between the peptide insertion in the **(A)** kinked and the **(B)** straight IAPP. The mean and standard deviation is indicated with a blue and black lines, respectively. The background color of the plots indicates the significance of the difference (green: $p \leq 0.05$, yellow: $0.05 < p < 0.10$, red: $p > 0.10$).

with CHOL bring more ordered is reflected in the membrane thickness (**Figure 8D**), which is increased from around 3.8 nm to around 4.1 nm when adding CHOL. The membrane thickness was calculated from the distance between the peaks of the phospholipid phosphate density of each leaflet.

For the simulations with the straight peptides, the increased membrane distance in the presence of CHOL correlates with the increased membrane thickness and order. This effect is, however, not seen in the simulations with the kinked peptide, where the presence of DOPS lipids seems to diminish the effect of CHOL on the positioning of the helix on the membrane.

Anionic Lipids Are Enriched in the Vicinity of IAPP and Cholesterol Is Depleted

The lipid depletion-enrichment index (D-E index) is calculated to investigate local concentration variations of the lipids surrounding the peptide (Corradi et al., 2018). The analysis is based on the method developed by Corradi et al. (2018) and (Corradi et al., 2018). The first panel in **Figure 9A** shows the D-E index of lipids within the first lipid shell surrounding the membrane bound peptide (lipids within 7 Å of the $IAPP_{1-19}$). For the systems with DOPC and DOPS lipid there is a significant depletion of DOPC and a significant enrichment DOPS within the at all the measured distances (7, 14, and 21 Å), the effect is, however, decreasing with the distance. The distances 7, 14, and 21 Å, were used in the original paper by Corradi et al. (2018) and represent the three nearest lipid shells around $IAPP_{1-19}$, as illustrated in **Figure 9B**. The enrichment is most likely due to favorable interactions between the cationic residues of IAPP and the anionic groups of the PS headgroup. In the systems with DOPC and CHOL, there is a slight depletion of CHOL within the first shell around the peptide. The depletion of CHOL is independent of DOPS, and therefore indicates a preference for interactions between the helix and phospholipids

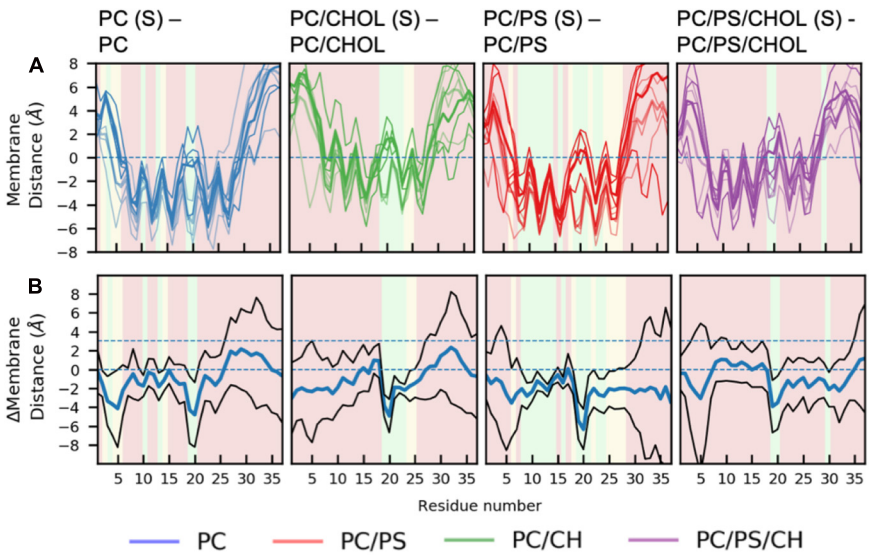


FIGURE 7 | Comparison of membrane positions of straight and kinked IAPP relative to the membrane. **(A)** Position of Cα of the peptide relative to the phosphate plane, colored as in **Figure 5**. **(B)** Differences between the peptide insertion in the kinked and the straight IAPP, colored as in **Figure 6**.

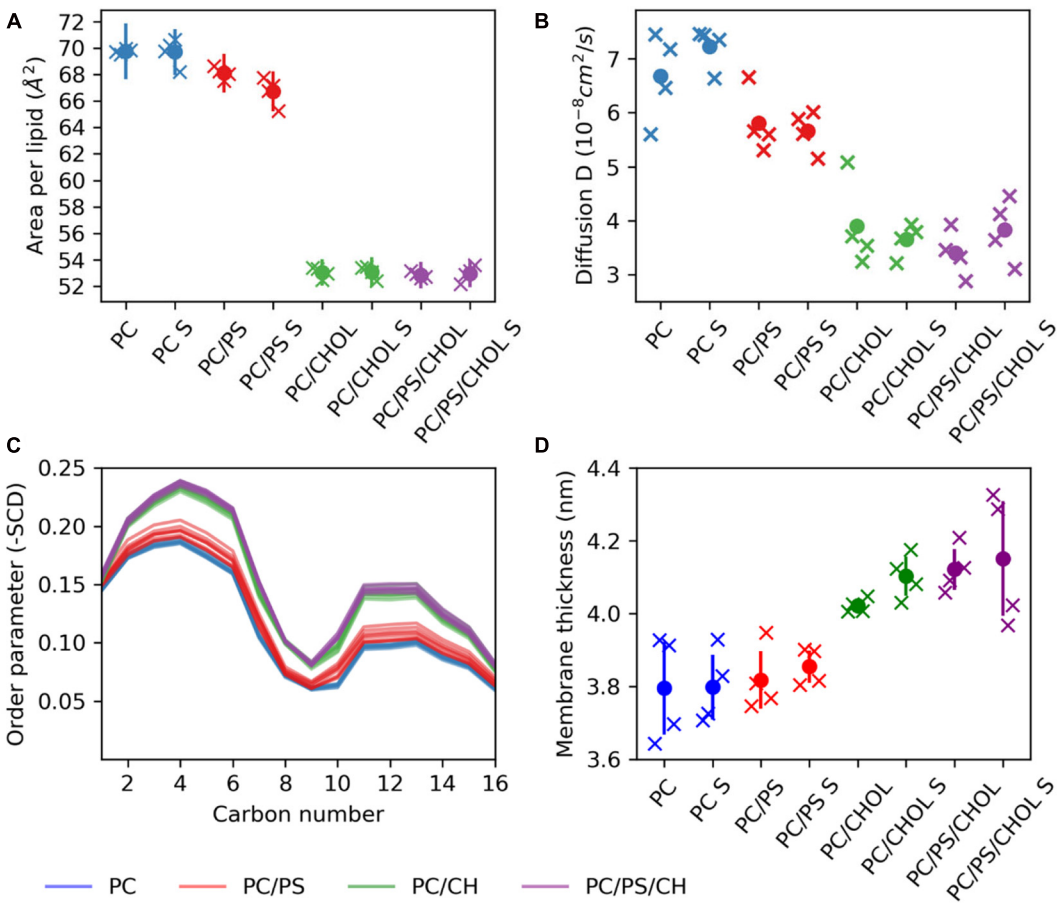


FIGURE 8 | Membrane properties. **(A)** Area per lipid. **(B)** Diffusion coefficient. **(C)** Order parameter **(D)** Membrane thickness. The properties are shown for each simulation series (each individual simulation marked by an X), with S indicating the simulation series with IAPP starting in the straight conformation.

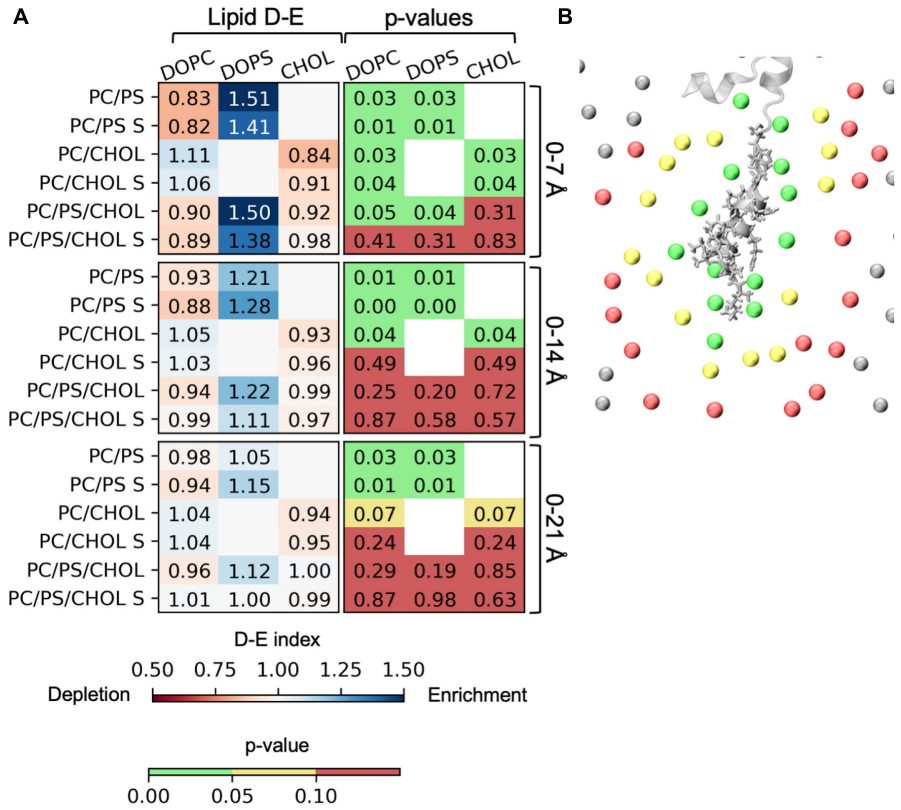


FIGURE 9 | Depletion enrichment index (D-E index) of lipids around the peptide. **(A)** The panels represent the D-E index of the lipids within 7, 14, and 21 Å. The first column shows the values of the D-E index, colored on a scale from 0.5 to 1.5. The second column shows the *p*-values. The statistical significance was evaluated using the single sample *T*-test, with 1 as the null-hypothesis, indicating that there is no enrichment or depletion of a given lipid type around the peptide. **(B)** The phospholipid phosphates within 7 Å (green), 14 Å (yellow), and 21 Å (red) of $IAPP_{1-19}$. The D-E index is calculated for each simulation series, and the simulation series with IAPP starting in the straight conformation are denoted with an "S."

rather than CHOL, as seen by the values below 1 in the CHOL column for the 0–7 Å range. More variation is seen in the three component system with both DOPC, DOPS, and CHOL, the DE-indices are less significant, probably due to the increased possibility of variation, and thus more sampling is required. However, the enrichment of DOPS in the inner shell is significant for the systems starting with the peptide in a kinked conformation.

Flexibility in the N-terminal Loop Shows Two Dominant Orientations

Since the effect of the lipids on the position of IAPP can not be explained completely by the physical properties of the membrane, a structural clustering is performed to investigate the specific interactions between the lipids and IAPP. In order to perform a structural clustering of IAPP and the membrane lipids, the conformational flexibility of IAPP has to be taken into account. The largest flexibility is in (a) the N-terminal loop (residue Lys1 to Cys7), (b) in the C-terminus after the C-terminal helix from residue Ser28 to Tyr37, and (c) in the kink region of the peptides between the N-terminal and C-terminal helix; the kink usually includes residue His18 to Phe23, and allows for variation in the angle between the two helices.

A geometric clustering was performed to better understand the flexibility in the N-terminal loop (More details are available in the **Supplementary Material**). It was found to be oriented in one of two ways for most of the simulation time. Either the N-terminal loop is pointing to the left or the right side compared to the helix, when viewing from the N-terminus, as illustrated in **Figure 10**: These two orientations collect 72.8% of the conformations sampled by the peptides (38.0 and 34.8%, for orientation 1 and 2, respectively) the additional clusters contained far fewer structures (**Supplementary Figure 2**). In the individual simulations the peptide does not sample both orientations equally. Figures of the distribution of the orientations for each simulation repeat are found in the SI along with calculations of the median structure for each repeat. It is seen from the graphs that when CHOL is present IAPP orientation 1 (O1) is slightly favored over orientation 2 (O2).

The effect of the orientation of the N-terminal loop might seem insignificant, however, the N-terminal loop contains two of the charges and is therefore essential for the interaction with the phospholipids. In addition, Cys2-Cys7 disulfide bridge in the N-terminal loop has also been proposed to decrease the

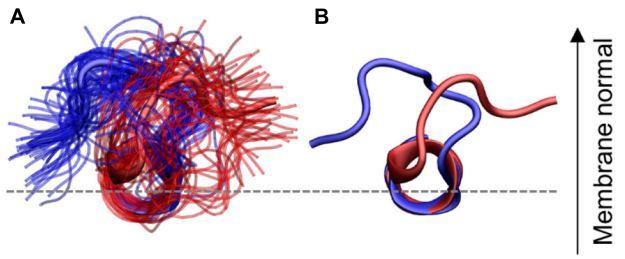


FIGURE 10 | N-terminal loop orientations, structures with loop orientation 1 (O1) and 2 (O2) are shown in blue and red, respectively. **(A)** Ensemble of structures from the two largest clusters **(B)** central structure from each of the two orientations. The C-terminal region is hidden for clarity.

propensity of IAPP to aggregate (Milardi et al., 2008). The importance of the N-terminal orientation will become apparent in the following analysis, about the specific interactions between IAPP and the membrane lipids.

Certain Regions of the N-terminal Helix Favor Lipid Interactions

The 3D occupancy maps in **Figure 11** reveals the volumes around $IAPP_{1-19}$ with a high density of DOPC, DOPS and CHOL in the simulations (with the peptide aligned in the helical segment $IAPP_{7-16}$). The analysis is repeated for IAPP with the N-terminal loop in each of the two orientations. The occupancy maps can be seen from other angles in **Supplementary Figure 4**.

DOPC is the lipid present at the highest concentration in the bilayer in all simulations. The density map in **Figure 11A** indicates that DOPC preferentially interacts with $IAPP_{O1}$ in two regions around the peptide; at the N-terminal of the peptide, where DOPC interacts with the loop region (Lys1 – Cys7). Seen from above (**Figure 11A**) DOPC also preferentially interacts with the right side of the N-terminal helix, on this side it can interact with Arg11, Phe15, and Asn14. At the kink region, DOPC has a preferential interaction near Phe15, His18, and Ser19. The interaction volumes of DOPC are very dependent on the orientation of the N-terminal loop; DOPC interacts mainly with $IAPP_{O2}$ in the loop region around Lys1, between Lys1 and Arg11, between Arg11 and Phe15, and near His18 and Phe15 (**Figure 11B**). In both of the orientations of the N-terminal loop, DOPC shows a preference toward binding to the right side of the helix (**Figures 11A,B**).

DOPS is negatively charged, with two negative charges and one positive charge in the head group, and therefore preferentially interacts with the cationic groups of IAPP. In $IAPP_{O1}$, DOPS interacts near the N-terminus and Lys1 or between Arg11, Phe15, and Asn14 (**Figure 11C**). $IAPP_{O2}$ has the positive charges of Lys1 and the N-terminus on the same side as Arg11, which makes it favorable for DOPS to be positioned in the cleft between these residues, to the right side of the helix (**Figure 11D**).

Cholesterol is much more hydrophobic than DOPC and DOPS, and is therefore less prone to interactions with the hydrophilic face of the amphipathic helix. However, CHOL has a hydroxyl group that preferentially interacts with the hydrophilic groups near the interface, such as exposed backbone atoms and residue sidechains at the helix ends (**Figures 11E,F**). The interaction volumes of CHOL are mostly positioned below the helix, in the vicinity of the hydrophobic residues Leu12, Leu16, and Phe15. The interactions are unspecific, more evenly distributed around the IAPP than the phospholipids, and seem to be less dependent on the orientation of the N-terminal loop and Lys1 (**Figures 11E,F**).

The Binding Modes of the Lipids May Be Related to the Insertion of the Peptide

For a more detailed description of preferred binding sites of the membrane lipids to IAPP we have performed a distance based clustering of the helix-lipid pairs (details available in **Supplementary Material**). It was observed from the 3D Occupancy maps, that CHOL interacts differently with $IAPP_{1-19}$ than the phospholipids, and these will therefore be presented separately. The following analysis is based on a collection of all the IAPP-lipid pairs in all structures with the N-terminal loop in either O1 or O2.

For the phospholipids, the dominant interactions are very dependent on hydrogen bonding and salt bridge interactions between the lipids headgroup and the hydrophilic residues. The most frequently occurring binding modes (BMs) of DOPC and DOPS are illustrated with representative structure in **Figure 12**, and the relative frequencies are shown in **Supplementary Table 1**. The five largest cluster will be discussed here, as they constitute the vast majority of the interactions.

Binding mode 1 (BM1) This is the most prominent lipid binding mode placed on the right side of the helix, with the lipid binding in a pocket between Arg11 and Asn14 and the tails between Leu12 and Phe15. This binding mode is independent of the orientation of the N-terminal loop and the binding mode is observed for both DOPC and DOPS lipids. *Binding mode 2* (BM2) is another prominent "binding mode." It entails the unspecific salt-bridge interactions with Lys1, both DOPC and DOPS has a negative charge at the phospholipid phosphate which can interact with Lys1, the N-terminus, or bridging between the two. In the case of DOPS, the two negative charges can interact with Lys1 and the N-terminus simultaneously, thus making this interaction very favorable for DOPS lipids, as seen from the relatively high occurrence of this binding mode for DOPS lipids compared to DOPC lipids. Independent of the orientation of the N-terminal loop, it is a common interaction point for the phospholipids. In *Binding mode 3* (BM3) the N-terminal loop between Cys2 and Cys7 consists of several hydrophilic residues (Asn3, Thr4, and Thr6), and it has exposed backbone atoms, making it a good interaction point for the hydrophilic headgroups of both DOPC and DOPS. In O2 of *Binding mode 4* (BM4) the N-terminus and Lys1 is pointing toward the right side of the helix. In this orientation the phospholipid can interact with Lys1 (or the N-terminus) and Arg11 simultaneously. *Binding mode 5* (BM5) only occurs for the loop in O1, where the phospholipids can bind between Gln10 and Lys1 (or the N-terminus).

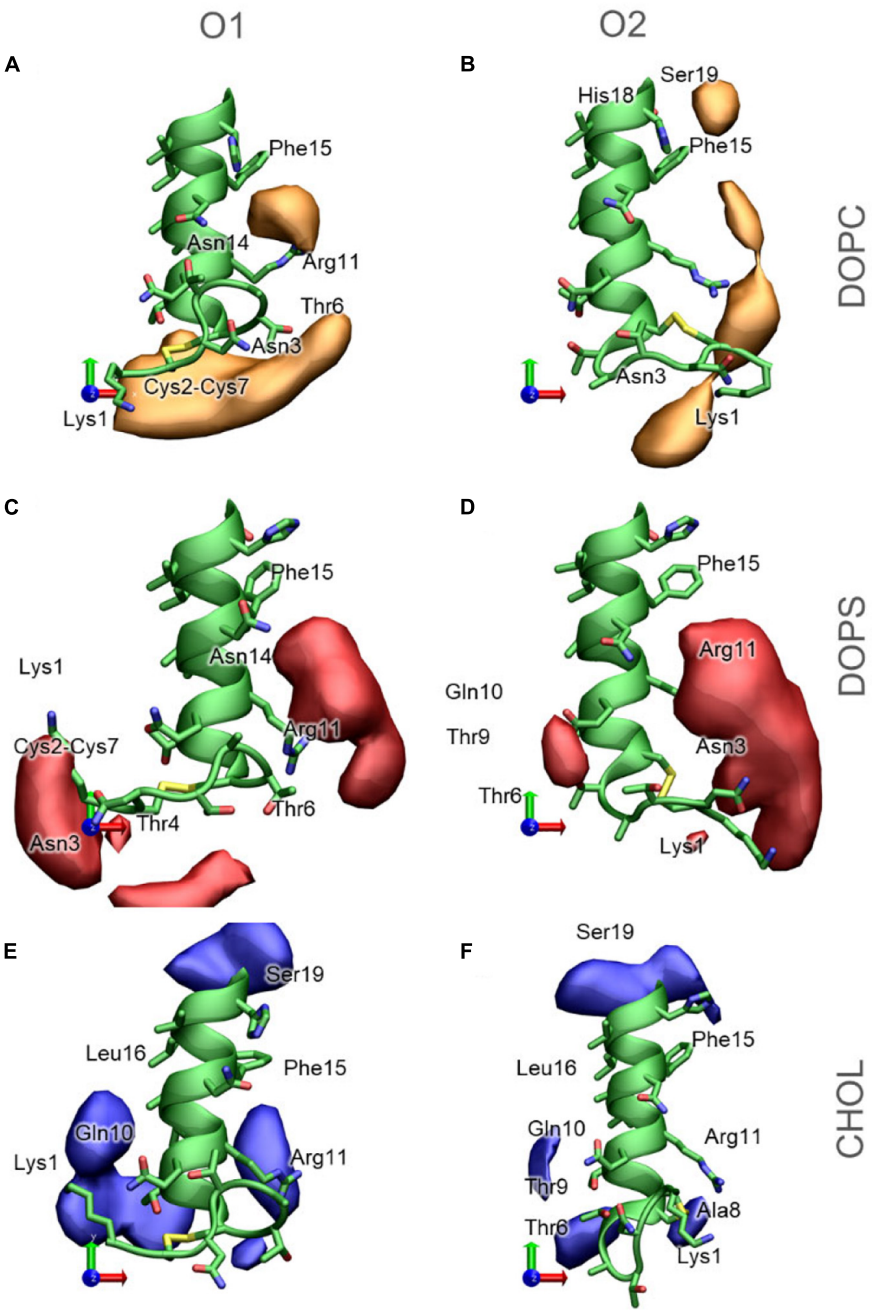


FIGURE 11 | 3D Occupancy maps of the lipids around $IAPP_{1-19}$, depending on the orientation of the N-terminal loop **(A)** DOPC around $IAPP_{O1}$ **(B)** DOPC around $IAPP_{O2}$ **(C)** DOPS around $IAPP_{O1}$ **(D)** DOPS around $IAPP_{O2}$ **(E)** CHOL around $IAPP_{O1}$ **(F)** CHOL around $IAPP_{O2}$. $IAPP_{1-19}$ is shown in green, and selected residues are labeled.

In addition to the binding modes presented here, less frequently occurring binding encounters are found, e.g., some binding modes interacting with the kinked region, more details about these binding modes are available in **Supplementary Material**. The major binding modes presented here compares well with the 3D occupancy maps in **Figure 8**, however, the binding modes that involve the more flexible loop region and Lys1 might be under-represented in the occupancy maps.

It would be expected that the membrane composition affects the distribution and frequency of the lipid binding modes. The lipids could potentially be competing for the binding site, and the change in positioning of the helix in the presence of CHOL could affect the frequency of experiencing a given binding modes. The binding frequency of the binding modes could also be dependent on whether IAPP is straight or kinked. The fraction of each simulations series contributing to the binding

modes of phospholipids and CHOL is shown in **Supplementary Figures 7, 8**, respectively. Few trends can be observed: For DOPC the contribution from each simulation is very heterogeneous across the simulation series, this indicates that the other lipids are not significantly competing with DOPC for any of the binding modes. In the interactions between DOPS and IAPP in O1, the straight peptides contributes most to BM1 and BM2, while the kinked peptides seem to provide better lipid interactions as in BM3 and BM5. None of these binding modes are in the kink-region, which indicates that the difference might arise from the difference in membrane insertion, which depends on the conformation of IAPP (as shown in **Figure 7**). There is a smaller fraction of the DOPS lipids interacting in BM4 in the presence of CHOL, which could indicate competing interactions.

Moving on to cholesterol, the 3D Occupancy maps show the interactions between the N-terminal helix and CHOL is less dependent on the orientation of the N-terminal loop. The interactions between CHOL and the N-terminal helix are primarily defined by hydrophobic interactions and hydrogen bonding of the hydroxyl group. The binding modes of CHOL are depicted in **Figure 13**, and their relative frequencies are listed in **Supplementary Table 2**.

Binding mode 1 (BM1) of CHOL to the N-terminus of IAPP is at the kink region, this region has several hydrophilic residues (His18 and Ser19) and exposed backbone in the kinked peptides. At the kink region the hydrophobic residues Phe15, Leu16 and Val17 can interact with the hydrophobic rings of CHOL. Even for the straight peptides the kink region is an important interaction point (**Supplementary Figure 8**), due to interactions with these residues. *Binding mode 2* (BM2) is at the left side of the helix, near the N-terminal. In BM2, CHOL hydrogen bond to Thr9 or Gln10, and interact with the hydrophobic residues Leu12 and Ala13. In *binding mode 3* (BM3) CHOL interacts with the N-terminal helix end, most frequently interacting with Ala8, Thr9, and Leu12, and the exposed backbone atoms. *Binding mode 4* (BM4) is very unspecific, in this binding position CHOL interacts with the hydrophobic residues below the helix, most frequently Leu12 and Phe15. In *Binding mode 5* (BM5), CHOL interacts with the right side of the helix between the loop and Arg11, with hydrogen bonding to Arg11 and hydrophobic interactions with Ala8.

The relative occurrence of the CHOL binding modes can also be affected by the orientation of the N-terminal loop, the conformation of IAPP (straight or kinked), and the membrane composition. The relative contribution of each simulations series to the binding modes is shown in **Supplementary Figure 8**.

Binding mode 1 involves binding in the kink region, it is therefore expected to be affected by whether IAPP is in a straight or a kinked conformation. In the PC/CHOL membranes, BM1

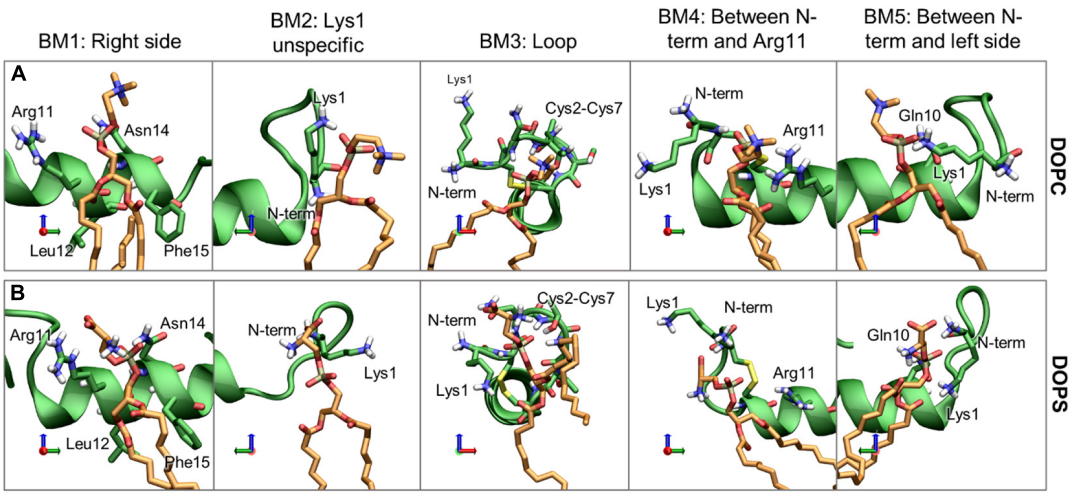


FIGURE 12 | Representative structures of the phospholipid binding modes (BM1-5), for **(A)** DOPC and **(B)** DOPS. IAPP is shown in green, with green carbons, and the lipids are shown with orange carbons.

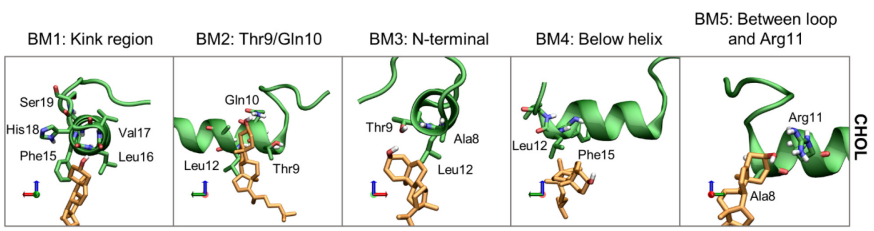


FIGURE 13 | Representative structures of the CHOL binding modes (BM1-5). IAPP is shown in green, with green carbons, and the CHOL is shown with orange carbons.

is observed more frequently for the CHOL binding to IAPP in the kinked conformation than in the straight conformation, independent of the orientation of the N-terminal loop. In the PC/PS/CHOL membrane, this trend is not observed, and their relative binding frequency is independent of the conformation of IAPP, as seen in **Supplementary Figure 8**. This also indicates that the binding frequency of CHOL in BM1 on IAPP in the kinked conformation is reduced in the presence of DOPS (**Supplementary Figure 8**). Since DOPS does not have a strong tendency to interact with the kink region (As seen in **Figure 11C**), it is not likely that the effect is due to competitive interactions. In the system PC/CHOL S (straight conformation), PC/PS/CHOL, and PC/PS/CHOL S, IAPP is inserted deeper in the membrane than in the PC/CHOL systems, as shown in **Figure 7**. It could therefore be, that the change in the binding frequency of CHOL in BM1 is affected by the positioning of IAPP in the membrane, which could make binding in the kinked region less favorable.

Other than the variations in the binding frequency in BM1 there are no major variations across the systems. The binding modes are very evenly distributed around $IAPP_{1-19}$, which was also observed from the 3D occupancy maps in **Figure 11**.

CONCLUSION

Cholesterol is an important element of the lipid membrane, and it is often present as a sizeable fraction of the membrane lipids. CHOL affect the aggregation propensity and membrane effects of IAPP, however, the specific role of CHOL is unclear. This study investigates the effect of CHOL on the membrane bound state of IAPP.

The results presented here, confirm the stabilizing effect of lipid bilayers toward the helical structures of IAPP. Both the straight and the kinked conformations of IAPP were found to be stable on the membrane but unfolded in solution. The C-terminus was found to be less structured than the core membrane binding region in the N-terminal half of the peptide, which is in good agreement with previous results (Skeby et al., 2016). No significant variation of the helix stability was observed between the membrane compositions, but the straight helices were found to be more stable than the kinked helices, as seen from the secondary structure calculation shown in **Figure 4**.

It was found that the depth of membrane insertion of the amphipathic helices was affected by CHOL and whether IAPP was in a straight or a kinked conformation. CHOL generally results in a less deep insertion of the IAPP, than in pure phospholipid membranes, which can be explained by the increased ordering of lipids. A deep binding indicates a strong binding, and IAPP is observed to bind weaker to ordered membranes in experiments (Zhang et al., 2018). The present study has revealed that a membrane compositions of PC/PS/CHOL and a kinked conformation of IAPP, the PS lipids diminishes the effect of CHOL, and IAPP binds as deeply as in the pure phospholipid membranes, as seen from the membrane distance calculation in **Figure 6**. This concurs with the experimental results from Zhang et al. (2017) that CHOL in simple phospholipid membranes generally inhibits membrane induced amyloid formation of IAPP, and that the presence of PS lipids diminishes the inhibiting effect of CHOL. The insertion depth of IAPP was also affected by the conformation of IAPP, the straight conformation inserted significantly deeper than the kinked conformation, especially in the pure phospholipid membranes.

It was found that the membrane patch around $IAPP_{1-19}$ was enriched with DOPS lipids and slightly depleted of CHOL, as seen from the D-E indices in **Figure 9**. The interactions between $IAPP_{1-19}$ and the phospholipids is very dependent on the positively charged groups (N-terminal, Lys1, and Arg11). CHOL has a less well-defined interaction pattern, and it mostly interacts underneath the helix and with the kink region, as can be observed from the most frequent CHOL binding modes and 3D occupancy maps. It has been proposed from experiments that CHOL interacts directly with Phe15, which increases the binding affinity of IAPP to the membrane (Hao et al., 2018). It was proposed that CHOL interacted with a specific CHOL binding segment that involved a CARC motif that included Arg11, Phe15, and Val17, but this is not supported by our results. Phe15 is, however, found to be important for CHOL interaction, since Phe15 is centrally positioned in the N-terminal helix, and thus involved in many of the binding modes.

Based on these results we propose, that the effect of CHOL on the membrane bound state of IAPP and the membrane induced amyloid formation, is based on changes in insertion of IAPP due to changes in the physical properties of the membrane, rather than through direct interactions. We managed to observe significant changes in the insertion depth of IAPP in the presence of CHOL, and we found that PS can diminish the effect of CHOL. This provides a structural interpretation of the effect of CHOL on the membrane binding and the effect of the lipids on the aggregation rate.

This study is a step toward a deeper understanding of the role of CHOL in IAPP aggregation. Further studies, which take lipid raft formation into account need to be performed, but this topic is reserved for future work.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

MC did the simulations and analysis. BS designed the project and supervised the work. All authors wrote the manuscript.

FUNDING

This work was supported by the Lundbeck Foundation and the Sino-Danish Center for Education and Research (Grant No. R191-2015-827). Computations were performed at the Abacus 2.0 of the DeIC National HPC Centre, University of Southern Denmark.

ACKNOWLEDGMENTS

We would like to acknowledge Novo Nordisk Foundation (Grant No. NNF18OC0032608) for computational resources, and the Sino-Danish Center and Lundbeck foundation for other funding.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.657946/full#supplementary-material

REFERENCES

Abedini, A., and Raleigh, D. P. (2005). The role of his-18 in amyloid formation by human islet amyloid polypeptide. *Biochemistry* 44, 16284–16291. doi: 10.1021/bi051432v

Abedini, A., and Schmidt, A. M. (2013). Mechanisms of islet amyloidosis toxicity in type 2 diabetes. *FEBS Lett.* 587, 1119–1127. doi: 10.1016/j.febslet.2013.01.017

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

Ben-Tal, N., Ben-Shaul, A., Nicholls, A., and Honig, B. (1996). Free-energy determinants of alpha-helix insertion into lipid bilayers. *Biophys. J.* 70, 1803–1812. doi: 10.1016/s0006-3495(96)79744-8

Berendsen, H. J. C., Postma, J. P., Van Gunsteren, W. F., Dinola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J.Phys. Chem.* 81, 3684–3690. doi: 10.1063/1.448118

Caillon, L., Duma, L., Lequin, O., and Khemtemourian, L. (2014). Cholesterol modulates the interaction of the islet amyloid polypeptide with membranes. *Mol. Membr. Biol.* 31, 239–249. doi: 10.3109/09687688.2014.987182

Cao, P., Marek, P., Noor, H., Patsalo, V., Tu, L.-H., Wang, H., et al. (2013). Islet amyloid: from fundamental biophysics to mechanisms of cytotoxicity. *FEBS Lett.* 587, 1106–1118. doi: 10.1016/j.febslet.2013.01.046

Carr, M., and MacPhee, C. E. (2015). Membrainy: a 'smart', unified membrane analysis tool. *Source Code Biol. Med.* 10:3.

Christensen, M., Skeby, K. K., and Schiøtt, B. (2017). Identification of key interactions in the initial self-assembly of amylin in a membrane environment. *Biochemistry* 56, 4884–4894. doi: 10.1021/acs.biochem.7b00344

Corradi, V., Mendez-Villuendas, E., Ingólfsson, H. I., Gu, R.-X., Siuda, I., Melo, M. N., et al. (2018). Lipid-protein interactions are unique fingerprints for membrane proteins. *ACS Cent. Sci.* 4, 709–717. doi: 10.1021/acscentsci.8b00143

de Jong, D. H., Baoukina, S., Ingólfsson, H. I., and Marrink, S. J. (2016). Martini straight: boosting performance using a shorter cutoff and GPUs. *Comput. Phys. Commun.* 199, 1–7. doi: 10.1016/j.cpc.2015.09.014

de Jong, D. H., Singh, G., Bennett, W. F. D., Arnarez, C., Wassenaar, T. A., Schäfer, L. V., et al. (2013). Improved parameters for the martini coarse-grained protein force field. *J. Chem. Theory Comput.* 9, 687–697.

Dong, X., Qiao, Q., Qian, Z., and Wei, G. (2018). Recent computational studies of membrane interaction and disruption of human islet amyloid polypeptide: monomers, oligomers and protofibrils. *Biochim.Biophys. Acta (BBA) Biomembr.* 1860, 1826–1839. doi: 10.1016/j.bbamem.2018.03.006

Engel, M. F. M. (2009). Membrane permeabilization by islet amyloid polypeptide. *Chem. Phys. Lipids* 160, 1–10. doi: 10.1016/j.chemphyslip.2009.03.008

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

Gracià, R. S., Bezlyepkina, N., Knorr, R. L., Lipowsky, R., and Dimova, R. (2010). Effect of cholesterol on the rigidity of saturated and unsaturated membranes: fluctuation and electrodeformation analysis of giant vesicles. *Soft Matter* 6, 1472–1482. doi: 10.1039/b920629a

Grouleff, J., Irudayam, S. J., Skeby, K. K., and Schiøtt, B. (2015). the influence of cholesterol on membrane protein structure, function, and dynamics studied by molecular dynamics simulations. *Biochim.Biophys. Acta (BBA) Biomembr.* 1848, 1783–1795. doi: 10.1016/j.bbamem.2015.03.029

Habchi, J., Chia, S., Galvagnion, C., Michaels, T. C. T., Bellaiche, M. M. J., Ruggeri, F. S., et al. (2018). Cholesterol catalyses Aβ42 aggregation through a heterogeneous nucleation pathway in the presence of lipid membranes. *Nat. Chem.* 10, 673–683. doi: 10.1038/s41557-018-0031-x

Hao, R., Li, Y., Guan, L., Lu, T., Meng, F., Wang, C., et al. (2018). Cholesterol-sensing role of phenylalanine in the interaction of human islet amyloid polypeptide with lipid bilayers. *RSC Adv.* 8, 40581–40588. doi: 10.1039/c8ra07310d

Hay, D. L., Chen, S., Lutz, T. A., Parkes, D. G., and Roth, J. D. (2015). Amylin: pharmacology, physiology, and clinical potential. *Pharmacol. Rev.* 67, 564–600.

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. doi: 10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h

Hockney, R. W., Goel, S. P., and Eastwood, J. W. (1974). Quiet high-resolution computer models of a plasma. *J. Comput. Phys.* 14, 148–158. doi: 10.1016/0021-9991(74)90010-2

Hoover, W. G. (1985). Canonical dynamics: equilibrium phase-space sistributions. *Phys. Rev. A* 31, 1695–1697. doi: 10.1103/physreva.31.1695

Höppener, J. W. M., Ahrén, B., and Lips, C. J. M. (2000). Islet amyloid and type 2 diabetes mellitus. *New Enl. J. Med.* 343, 411–419.

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., De Groot, B. L., et al. (2016). CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* 14:71. doi: 10.1038/nmeth.4067

Jacob, R. F., Cenedella, R. J., and Mason, R. P. (1999). Direct evidence for immiscible cholesterol domains in human ocular lens fiber cell plasma membranes. *J. Biol. Chem.* 274, 31613–31618. doi: 10.1074/jbc.274.44.31613

Jayasinghe, S. A., and Langen, R. (2005). Lipid membranes modulate the structure of islet amyloid polypeptide. *Biochemistry* 44, 12113–12119. doi: 10.1021/bi050840w

Ji, S.-R., Wu, Y., and Sui, S.-F. (2002). Cholesterol is an important factor affecting the membrane insertion of β-amyloid peptide (Aβ1–40), which may potentially inhibit the fibril formation. *J Biol. Chem.* 277, 6273–6279. doi: 10.1074/jbc.m104146200

Khemtémourian, L., Doménech, E., Doux, J. P. F., Koorengevel, M. C., and Killian, J. A. (2011). Low pH acts as inhibitor of membrane damage induced by human islet amyloid polypeptide. *J. Am. Chem. Soc.* 133, 15598–15604. doi: 10.1021/ja205007j

Klauda, J. B., Venable, R. M., Freites, J. A., O'connor, J. W., Tobias, D. J., Mondragon-Ramirez, C., et al. (2010). Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J. Phys. Chem. B* 114, 7830–7843. doi: 10.1021/jp101759q

Knight, J. D., Hebda, J. A., and Miranker, A. D. (2006). Conserved and cooperative assembly of membrane-bound α-helical states of islet amyloid polypeptide. *Biochemistry* 45, 9496–9508. doi: 10.1021/bi060579z

Lee, C.-C., Sun, Y., and Huang, H. W. (2012). How type II diabetes-related islet amyloid polypeptide damages lipid bilayers. *Biophys. J.* 102, 1059–1068. doi: 10.1016/j.bpj.2012.01.039

Li, Y., Ge, M., Ciani, L., Kuriakose, G., Westover, E. J., Dura, M., et al. (2004). Enrichment of endoplasmic reticulum with cholesterol inhibits sarcoplasmic-endoplasmic reticulum calcium aTPase-2b activity in parallel with increased order of membrane lipids implications for depletion of endoplasmic reticulum calcium stores and apoptosis in cholesterol-loaded macrophages. *J. Biol. Chem.* 279, 37030–37039. doi: 10.1074/jbc.m405195200

Li, Y., Guan, L., Lu, T., Li, H., Li, Z., and Li, F. (2016). Interactions of the N-terminal domain of human islet amyloid polypeptide with lipid membranes: the effect of cholesterol. *RSC Adv.* 6, 96837–96846. doi: 10.1039/c6ra19714k

Luca, S., Yau, W.-M., Leapman, R., and Tycko, R. (2007). Peptide conformation and supramolecular organization in amylin fibrils: constraints from solid-state NMR. *Biochemistry* 46, 13505–13522. doi: 10.1021/bi701427q

Lukinius, A., Wilander, E., Westermark, G. T., Engstrom, U., and Westermark, P. (1989). Co-localization of islet amyloid polypeptide and insulin in the b cell secretory granules of the human pancreatic-islets. *Diabetologia* 32, 240–244. doi: 10.1007/bf00285291

Lutz, T. A. (2010). The role of amylin in the control of energy homeostasis. *Am. J. Physiol. Regul. Integ. Comp. Physiol.* 298, R1475–R1484.

Lutz, T. A. (2012). Control of energy homeostasis by amylin. *Cell. Mol. Life Sci.* 69, 1947–1965. doi: 10.1007/s00018-011-0905-1

Maloy, A. L., Longnecker, D. S., and Robert Greenberg, E. (1981). The relation of islet amyloid to the clinical type of diabetes. *Hum. Pathol.* 12, 917–922. doi: 10.1016/s0046-8177(81)80197-9

Milardi, D., Gazit, E., Radford, S. E., Xu, Y., Gallardo, R. U., Caflisch, A., et al. (2021). Proteostasis of islet amyloid polypeptide: a molecular perspective of risk factors and protective strategies for type II diabetes. *Chem. Rev.* 121, 1845–1893. doi: 10.1021/acs.chemrev.0c00981

Milardi, D., Pappalardo, M., Pannuzzo, M., Grasso, D. M., and Rosa, C. L. (2008). The role of the Cys2-Cys7 disulfide bridge in the early steps of Islet amyloid polypeptide aggregation: a molecular dynamics study. *Chem. Phys. Lett.* 463, 396–399. doi: 10.1016/j.cplett.2008.07.110

Nanga, R. P. R., Brender, J. R., Vivekanandan, S., and Ramamoorthy, A. (2011). Structure and membrane orientation of IAPP in its Natively amidated form at physiological pH in a membrane environment. *Biochim. Biophys. Acta BBA Biomembr.* 1808, 2337–2342. doi: 10.1016/j.bbamem.2011.06.012

Nicolson, G. L. (2014). The fluid—mosaic model of membrane structure: still relevant to understanding the structure, function and dynamics of biological membranes after more than 40years. *Biochim. Biophys. Acta BBA Biomembr.* 1838, 1451–1466. doi: 10.1016/j.bbamem.2013.10.019

Parrinello, M., and Rahman, A. (1981). Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* 52, 7182–7190. doi: 10.1063/1.328693

Patil, S. M., Xu, S., Sheftic, S. R., and Alexandrescu, A. T. (2009). Dynamic α-helix structure of micelle-bound human amylin. *J. Biol. Chem.* 284, 11982–11991. doi: 10.1074/jbc.m809085200

Raleigh, D., Zhang, X., Hastoy, B., and Clark, A. (2017). The β-cell assassin: IAPP cytotoxicity. *J. Mol. Endocrinol.* 59, R121–R140.

Rzepiela, A. J., Louhivuori, M., Peter, C., and Marrink, S. J. (2011). Hybrid simulations: combining atomistic and coarse-grained force fields using virtual sites. *Phys. Chem. Chem. Phys.* 13, 10437–10448. doi: 10.1039/c0cp02981e

Sciacca, M. F., Pappalardo, M., Milardi, D., Grasso, D. M., and La Rosa, C. (2008). Calcium-activated membrane interaction of the islet amyloid polypeptide: implications in the pathogenesis of type II diabetes mellitus. *Arch. Biochem. Biophys.* 477, 291–298. doi: 10.1016/j.abb.2008.06.018

Sciacca, M. F., Tempra, C., Scollo, F., Milardi, D., and La Rosa, C. (2018). Amyloid growth and membrane damage: current themes and emerging perspectives from theory and experiment son Aβ and hIAPP. *Biochim. Biophys. Acta BBA Biomembr.* 1860, 1625–1638. doi: 10.1016/j.bbamem.2018.02.022

Sciacca, M. F. M., Lolicato, F., Di Mauro, G., Milardi, D., D'urso, L., Satriano, C., et al. (2016). The role of cholesterol in driving IAPP-membrane interactions. *Biophys. J.* 111, 140–151. doi: 10.1016/j.bpj.2016.05.050

Skeby, K. K., Andersen, O. J., Pogorelov, T. V., Tajkhorshid, E., and Schiøtt, B. (2016). Conformational dynamics of the human islet amyloid polypeptide in a membrane environment: toward the aggregation prone form. *Biochemistry* 55, 2031–2042. doi: 10.1021/acs.biochem.5b00507

Steinbach, P. J., and Brooks, B. R. (1994). New spherical-cutoff methods for long-range forces in macromolecular simulation. *J. Comput. Chem.* 15, 667–683. doi: 10.1002/jcc.540150702

Tironi, I. G., Sperb, R., Smith, P. E., and Van Gunsteren, W. F. (1995). A generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.* 102, 5451–5459. doi: 10.1063/1.469273

Trikha, S., and Jeremic, A. M. (2011). Clustering and internalization of toxic amylin oligomers in pancreatic cells require plasma membrane cholesterol. *J. Biol. Chem.* 286, 36086–36097. doi: 10.1074/jbc.m111.240762

Ulmschneider, J. P., and Ulmschneider, M. B. (2018). Molecular dynamics simulations are redefining our view of peptides interacting with biological membranes. *Acc. Chem. Res.* 51, 1106–1116. doi: 10.1021/acs.accounts.7b00613

Wassenaar, T. A., IngóLfsson, H. I., BöCkmann, R. A., Tieleman, D. P., and Marrink, S. J. (2015). Computational lipidomics with insane: a versatile tool for generating custom membranes for molecular simulations. *J. Chem. Theory Comput.* 11, 2144–2155. doi: 10.1021/acs.jctc.5b00209

Wassenaar, T. A., Pluhackova, K., Böckmann, R. A., Marrink, S. J., and Tieleman, D. P. (2014). Going backward: a flexible geometric approach to reverse transformation from coarse grained to atomistic models. *J. Chem. Theory Comput.* 10, 676–690. doi: 10.1021/ct400617g

Westermark, P. (1995). "Islet amyloid polypeptide and amyloid in the islets of langerhans," in *Diabetes: Clinical Science in Practice*, eds R. D. G. Leslie and D. Robbins (Cambridge: Cambridge Univ. Press), 189–199.

Westermark, P., Andersson, A., and Westermark, G. T. (2011). Islet amyloid polypeptide, islet amyloid, and diabetes mellitus. *Physiol. Rev.* 91, 795–826. doi: 10.1152/physrev.00042.2009

Williamson, J. A., Loria, J. P., and Miranker, A. D. (2009). Helix stabilization precedes aqueous and bilayer-catalyzed fiber formation in islet amyloid polypeptide. *J. Mol. Biol.* 393, 383–396. doi: 10.1016/j.jmb.2009.07.077

Williamson, J. A., and Miranker, A. D. (2007). Direct detection of transient α-helical states in islet amyloid polypeptide. *Protein Sci.* 16, 110–117. doi: 10.1110/ps.062486907

Wiltzius, J. J. W., Sievers, S. A., Sawaya, M. R., Cascio, D., Popov, D., Riekel, C., et al. (2008). Atomic structure of the cross-β spine of islet amyloid polypeptide (Amylin). *Protein Sci.* 17, 1467–1474. doi: 10.1110/ps.036509.108

Yesylevskyy, S. O., Schäfer, L. V., Sengupta, D., and Marrink, S. J. (2010). Polarizable water model for the coarse-grained MARTINI force field. *PLoS Comput. Biol.* 6:e1000810. doi: 10.1371/journal.pcbi.1000810

Zeppelin, T., Ladefoged, L. K., Sinning, S., Periole, X., and Schiøtt, B. (2018). A direct interaction of cholesterol with the dopamine transporter prevents its out-to-inward transition. *PLoS Comput. Biol.* 14:e1005907. doi: 10.1371/journal.pcbi.1005907

Zhang, X., London, E., and Raleigh, D. P. (2018). Sterol structure strongly modulates membrane–islet amyloid polypeptide interactions. *Biochemistry* 57, 1868–1879. doi: 10.1021/acs.biochem.7b01190

Zhang, X., St Clair, J. R., London, E., and Raleigh, D. P. (2017). Islet amyloid polypeptide membrane interactions: effects of membrane composition. *Biochemistry* 56, 376–390. doi: 10.1021/acs.biochem.6b01016

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 © 2021 Christensen, Berglund and Schiøtt. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 10 May 2021
doi: 10.3389/fmolb.2021.659687

Molecular Mechanics Study of Flow and Surface Influence in Ligand–Protein Association

*Shivansh Kaushik and Chia-en A. Chang**

Department of Chemistry, University of Chemistry, Riverside, CA, United States

Ligand–protein association is the first and critical step for many biological and chemical processes. This study investigated the molecular association processes under different environments. In biology, cells have different compartments where ligand–protein binding may occur on a membrane. In experiments involving ligand–protein binding, such as the surface plasmon resonance and continuous flow biosynthesis, a substrate flow and surface are required in experimental settings. As compared with a simple binding condition, which includes only the ligand, protein, and solvent, the association rate and processes may be affected by additional ligand transporting forces and other intermolecular interactions between the ligand and environmental objects. We evaluated these environmental factors by using a ligand xk263 binding to HIV protease (HIVp) with atomistic details. Using Brownian dynamics simulations, we modeled xk263 and HIVp association time and probability when a system has xk263 diffusion flux and a non-polar self-assembled monolayer surface. We also examined different protein orientations and accessible surfaces for xk263. To allow xk263 to access to the dimer interface of immobilized HIVp, we simulated the system by placing the protein 20Å above the surface because immobilizing HIVp on a surface prevented xk263 from contacting with the interface. The non-specific interactions increased the binding probability while the association time remained unchanged. When the xk263 diffusion flux increased, the effective xk263 concentration around HIVp, xk263–HIVp association time and binding probability decreased non-linearly regardless of interacting with the self-assembled monolayer surface or not. The work sheds light on the effects of the solvent flow and surface environment on ligand–protein associations and provides a perspective on experimental design.

Keywords: molecular modeling, molecular recognition, drug design, GeomBD, ligand-receptor binding

OPEN ACCESS

Edited by:
Joanna Trylska,
University of Warsaw, Poland

Reviewed by:
Nanjie Deng,
Pace University, United States
Peter M. Kekenes-Huskey,
University of Kentucky, United States
Sanjib Senapati,
Indian Institute of Technology Madras, India

***Correspondence:**
Chia-en A. Chang
chiaenc@ucr.edu

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *28 January 2021*
Accepted: *06 April 2021*
Published: *10 May 2021*

Citation:
Kaushik S and Chang C-eA (2021) Molecular Mechanics Study of Flow and Surface Influence in Ligand–Protein Association. Front. Mol. Biosci. 8:659687. doi: 10.3389/fmolb.2021.659687

INTRODUCTION

Molecular association is the first critical step in all chemical and biological processes such as the immune response, signal transduction, drug-protein binding, and chemical catalysis (Ozbabacan et al., 2010; Dill and Bromberg, 2012; Baron and McCammon, 2013; Lin et al., 2020). Simulation techniques play a crucial role in investigating the environment that may affect ligand–protein binding, which provides a fundamental understanding of molecular recognition and reduces the cost and time in molecular design. Although diffusion-controlled association rate constants may

be approximated analytically, most ligand–protein systems have slower association rates than the diffusion-limited rate because the association event involves multiple steps (Di Cera, 2017; Pang and Zhou, 2017). Conformational rearrangement of both molecules largely determines their binding kinetics, but the two molecules must have an initial encounter first. This first step may be greatly affected by the environment, which can result in different measured association-rate constants.

Although many ligand–protein bindings take place in a closed system without water flow such as in an experimental beaker or cell, ligand–protein association can also occur in a more dynamic environment. Different compartments within a cell also create various membrane environments when ligands and proteins associate and function (Zotter et al., 2017). For example, techniques using continuous flow biocatalysis have been developed recently to improve the efficiency of chemical synthesis, such as improved mixing, mass transfer, and automation (Planchestainer et al., 2017; Britton et al., 2018). Surface plasmon resonance (SPR), a powerful technique to measure molecular binding kinetics and thermodynamics, also utilizes flow chemistry and a continuous flow environment (Hinman et al., 2018; Prabowo et al., 2018; Ershov et al., 2020). However, how the flow may affect the ligand–protein association processes is unclear. Moreover, these methods need to immobilize one molecule on a surface that may have intermolecular interactions with the molecules to be tested. Therefore, the choice of surface and flow rate usually need to be optimized for various systems.

In addition to experimentally measured values such as association rate constants, molecular simulation can reveal atomistic details of molecular association to further interpret experimental results and understand binding mechanisms. Molecular encounter usually involves two molecules searching for each other in a vast space and for longer than a nanosecond search time. Brownian dynamics (BD) simulations have been used for computational assessment of the encounter processes for several decades (Northrup et al., 1984; Huber and McCammon, 2019). Many software packages are available for studying a variety of problems related to bimolecular association. For example, MacroDox, UHBD, BrownDye Simulation of Diffusional Association (SDA), BD_BOX, BROMOC, ReaDDy, Smoldyn, SEEKR, and BDpack are used to probe ligand–protein associations with a flexible or rigid biomolecular model (Madura et al., 1995; Northrup et al., 1999; Huber and McCammon, 2010; Długosz et al., 2011; Schöneberg and Noé, 2013; De Biase et al., 2015; Martinez et al., 2015; Saadat and Khomami, 2015; Andrews, 2017; Votapka et al., 2017). Our group has been developing the GeomBD program, which focuses on using BD to investigate inter-enzyme intermediate transfer and substrate association on surface environments in biosensor and nanoenzyme structures (Roberts and Chang, 2015, 2016). Because of the diverse biosystems and binding environments, modifying an existing BD package to answer various questions is common practice.

Here we used the HIV protease (HIVp) and xk263 as a model system to study the effect of flow and surface on ligand–protein associations. HIVp belongs to the class of aspartyl proteases and is essential for maturation and assembly of infectious virions (Kohl et al., 1988). The protein cleaves the large polyprotein precursors to mature viral proteins (Huang et al., 2014), an essential function for viral replication, and is a major drug target for AIDS treatment. This is a well-studied system with rich experimental binding data for many inhibitors and FDA approved drugs (Ghosh et al., 2016). The flexible flaps of HIVp can serve as a gate: its opening and closing affects ligand binding. In proteins, an open/closed rate of a gate related to the diffusion of their binding partners determines a fast or slow gating for ligand binding (Szabo et al., 1982; McCammon, 2011). Because of the complex gating behavior of HIVp and its importance in therapeutics, earlier work applied BD to study drug–HIVp binding and also developed a specialized coarse-grained model for HIVp to model the large-scale motions of the flaps (Tozzini and McCammon, 2005; Chang et al., 2006; Kang et al., 2011; Li et al., 2012; Bernetti et al., 2019).

In this work, we used BD simulations to examine the xk263–HIVp association under the influence of ligand diffusion flux and a surface environment. The simulation applied rigid-body BD movements and considered atomistic details using pre-computed grids when computing intermolecular interactions and driving forces. The HIVp was immobilized on a surface or artificially placed 20 Å above the surface. We introduced a self-assembled monolayer (SAM) with a CH_3 terminal group to model a hydrophobic surface environment (Cholko et al., 2019). We analyzed ligand association time and binding probability with different diffusion fluxes of xk263, surface and HIVp orientations. The non-polar SAM provided only weak intermolecular attractions with xk263, but the surface did not accelerate the xk263 encounter processes. The x-direction diffusion flux of xk263 significantly reduced weak intermolecular interactions and searching near the surface which shortened the association time but also significantly reduced the probability of successful binding. The concentration gradient of xk263 was affected by the xk263 diffusion flux as well. Our studies suggest that the flow may have noticeable effects on molecular encounter processes and provide insights into the differences in the ligand–protein association in diverse conditions, such as in static cells or a continuous flow biocatalysis environment.

METHODS

Model System

The model system is a rectangular prism (400 × 400 × 220 $Å^3$), closed at the top, and consisting of HIVp, xk263, and a SAM surface (**Figure 1A**). The crystal structure of HIVp and the xk263 inhibitor were obtained from the Protein Data Bank (codes 1HHP and 1HVR, respectively) (Spinelli et al., 1991; Lam et al., 1994). The HIVp flaps, two polypeptides that cover the active site, are in the open position. The semi-open form crystal structure of HIVp was simulated to derive the open form (Supplementary Methods) (Huang et al., 2017). HIVp was kept fixed in the model (**Figure 1B**). The SAM surface with HIVp at the center consisted of undecanethiol chains on a gold sheet of 1 atom thickness (**Supplementary Methods**) (Cholko et al., 2019). The size of the SAM was 400 × 400 $Å^2$ having a hexagonal packing pattern with a packing density in the order of 10^{14} cm^{-2}, which

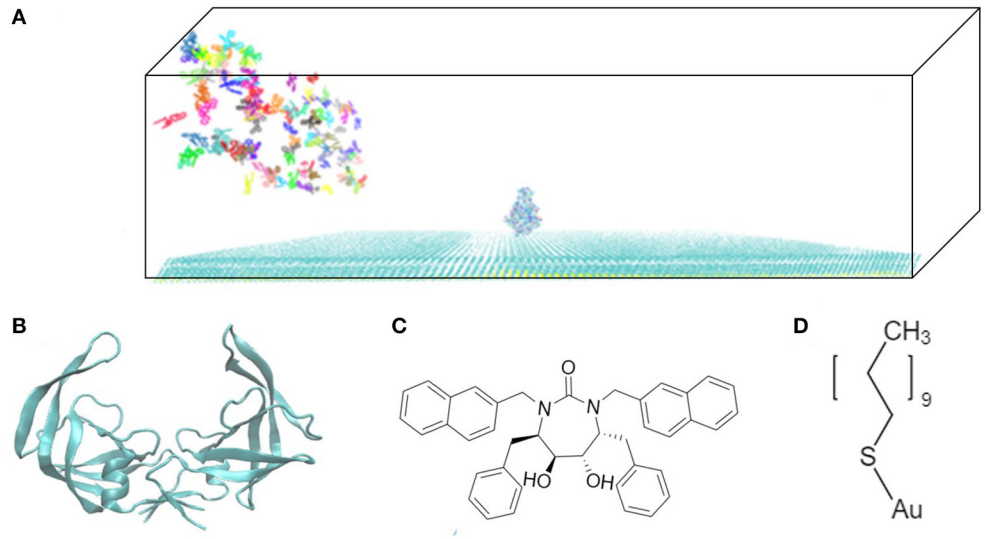


FIGURE 1 | Model system. **(A)** HIVp is placed at the center of the CH_3-SAM surface with xk263 ligands in a yz-plane when a simulation starts. **(B)** Structure of an open-form HIVp in cartoon representation. **(C)** xk263 ligand, and **(D)** undecanethiol molecule used to build the SAM surface. Notably, in each BD run, the molecular system has only one xk263 and one HIVp. To speed up the calculations, multiple replicas of xk263 are simulated simultaneously, but the replicas do not interact with each other.

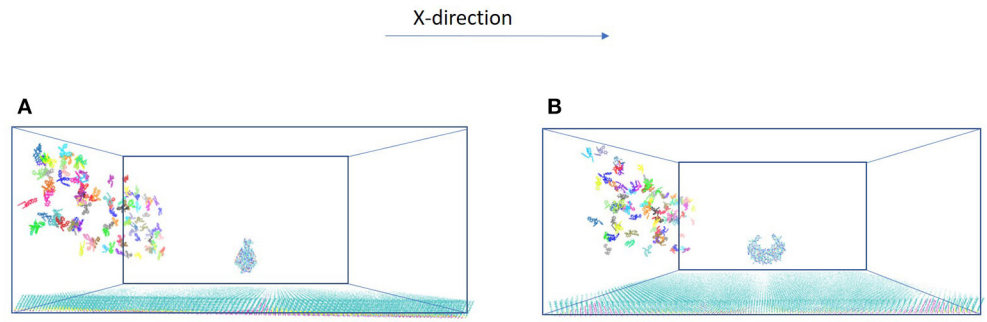


FIGURE 2 | Two orientations of HIVp. The protein is placed 20 Å above the SAM **(A)** parallel to the yz-plane and **(B)** perpendicular to the yz-plane.

gives an average chain separation of 4.98 Å. The system had a periodicity in the y-direction and termination boundary in the +x-direction as the ligand flows in the +x-direction. HIVp had two orientations (**Figure 2**).

Simulation Details

The in-house modified GeomBD2 program was used for all BD runs (Roberts and Chang, 2016). The program first creates three grid files: the exclusion volume, screened Coulumbic potential, and 12-6 Lennard-Jones potential for HIV and SAM. Intermolecular interactions between ligand xk263 and HIVp or SAM are grid-based calculations, and the forces are estimated numerically (Roberts and Chang, 2016). To speed up the calculations, these precalculated grids are extended up to 40 and 15 Å around HIVp or SAM for screened Coulumbic and 12-6 LJ potentials, respectively. The force field parameters were obtained from Amber-ff14SB for HIVp and CH_3-SAM. The partial charges of xk263 and SAM were computed by using the AM1-BCC charge method with the antechamber program (Cholko et al., 2019). The motion of the ligand is a Brownian motion, governed by an overdamped Langevin equation, in implicit water (Northrup et al., 1984).

$$r_i\,(t+\Delta t) = r_i\,(t) - \frac{D_i}{k_B T}\frac{\partial E}{\partial r_i}\Delta t + \sqrt{2D_i\Delta t}R \qquad (1)$$

where D_i is the translational or rotational diffusion coefficient, k_B is Boltzmann's constant, T is temperature, $\delta E/\delta r_i$ is the potential gradient computed numerically based on the potential grids, Δt is the time step, and R is the stationary Gaussian random number with a zero mean. To obtain the flow for the ligand, the overdamped Langevin equation is modified by adding *N*, a small fractional number, in the displacement of the x-direction.

$$x_i\,(t+\Delta t) = x_i\,(t) - \frac{D_i}{k_B T}\frac{\partial E}{\partial x_i}\Delta t + \sqrt{2D_i\Delta t}R + N \qquad (2)$$

For each molecular system, we terminated the simulation when the runs generated at least 600 xk263-HIVp associates and the computed average association time was within the standard deviation of <3%. A successful xk263–HIVp association is defined as xk263 reaches within 11.5 Å ASP25, a residue in the active site of HIVp, and the association time is recorded. Notably, after generating 600 successful bindings, the computed average associating time for each system is within 3%.

In the system, xk263 starts diffusing from the yz-plane at the -x-direction boundary (**Figure 1A**). A run is ended when xk263 reaches the binding pocket or the +x-direction boundary, and another new run will be started by randomly placing xk263 in a new position on the yz-plane (**Figure 3**). The clock random number seed is used, so no two runs can be identical. In addition, we can save trajectories of our BD runs to visualize ligand diffusion and binding/unbinding events.

Calculations for System Analysis

For the calculations of diffusion flux, diffusion coefficients, CH_3-SAM interaction, and xk263 distribution, we used the trajectories with no termination at the +x-direction boundary. So, the trajectories used for the calculations are continuous and 0.3-μs long for each case.

xk263 Diffusion Flux

The diffusion flux is the rate of molecules transferred across the plane per unit time per unit area, representing the flow of xk263. It was calculated by using the continuous BD trajectories and reported in units of molecules/s.m^2. The plane is an imaginary plane at the +x- directional boundary perpendicular to the direction of flow. The calculated flux densities according to the N values (in Equation 2) are reported (**Supplementary Table 1**).

Diffusion Coefficient

The diffusion coefficient was calculated by Einstein's relation, $<r^2> = 2nDt$, where r is the displacement in time t, n is the dimensionality, and D is the diffusion coefficient.

CH_3-SAM–xk263 Interaction Energy

To calculate the interaction energetics of xk263 with CH_3-SAM, we used the selected portion of continuous trajectories with xk263 diffusions within 25 Å above the CH_3-SAM. The interaction energy includes 12-6 Lennard-Jones potential and screened Coulombic potential. The cutoffs for vdW and electrostatic potential were 12 Å and 40 Å, respectively.

The calculation of 12-6 Lennard-Jones potential is as follows:

$$E_{vdw}=4\epsilon\left(\frac{\sigma^{12}}{r_{ij}^{12}}-\frac{\sigma^{6}}{r_{ij}^{6}}\right) \quad (3)$$

where r_{ij} is the distance between atoms i and j, ϵ is the pairwise well-depth parameter and σ is the distance at which the atomic radii meet and where the potential changes sign.

$$\sigma=\frac{\sigma_i+\sigma_j}{2} \quad (4)$$

$$\epsilon=\sqrt{\epsilon_i\epsilon_j} \quad (5)$$

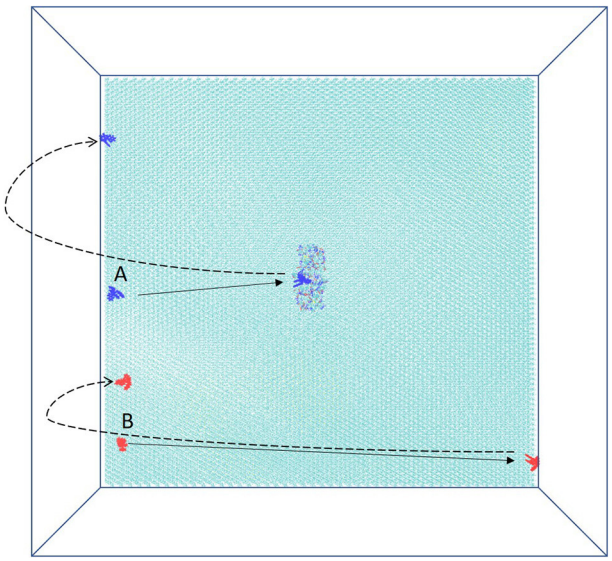


FIGURE 3 | Schematic top view of the simulation box. HIVp is placed on CH_3-SAM, and two initial starting positions of xk263, A and B, are on the yz-plane. The simulation is terminated after xk263 binding to HIVp, and another new replica will start. If xk263 reaches the +x-direction wall, the simulation is also terminated, and a new replica will start in a different position on the yz-plane.

TABLE 1 | Summary of molecular systems used in this study.

Details of model system	
HIVp orientations	2
Periodic wall	y
Flow direction	x
Top	Closed in z-direction
Ligand concentration	110 μM
Ligand initial position	As a yz-plane
Surface	with and without CH_3-SAM
Protein height	0 and 20 Å above the surface
Box dimension	400 × 400 × 220 Å^3
Termination	Reaching +x-direction wall or within 11.5 Å of ASP 25′

The calculation of screened Coulombic potential is as follows:

$$E_{SC}=\frac{k_e q_i q_j e^{-kr_{ij}}}{r_{ij}\epsilon} \quad (6)$$

where r_{ij} is the distance between two point-charges i and j from the ligand and receptor, respectively, q_i and q_j are the partial electron charges of the ligand point charges i and receptor point charges j, k is the screening parameter, k_e is the Coulomb constant, kb is the Boltzmann constant and ϵ is the solution dielectric constant.

xk263 Distribution

The distribution of xk263 in the system was calculated by using continuous BD trajectories. The volume of the system was

divided into 11 slices of 20 Å height each. The total number of appearances of xk263 replicas was computed for each slice in 0.3 μs. The calculation of *g(z)*, distribution, is according to Equation 8

$$g(z) = \frac{N}{M} = \frac{N}{n_R \times \frac{v}{V} \times f} \quad (7)$$

where N is the total number of appearances of xk263 replicas in a slice within a certain number of frames, M is the expected total number of appearances approximated from a random diffusional motion, n_R is the number of xk263 replicas in the system, v is the volume of the slice, V is the total volume of the system, and f is the number of frames to be analyzed.

RESULTS AND DISCUSSION

Understanding ligand–receptor associations in various environments is of vital importance in drug binding in cells or in different experiments. In this study, we analyzed the effects of the SAM surface, the diffusion flux of ligand xk263 and the position of HIVp from the surface and in comparison with results with a static environment (**Table 1**). HIVp has flexible flaps, which are mostly in closed conformations (Katoh et al., 2003; Kang et al., 2011). To bind with a ligand, the flaps may open spontaneously or are induced to open by the ligand. Because this study focuses on the initial molecular encounter processes, we chose a flap open conformation and used rigid-body BD simulations to model the ligand association. In addition, the orientation of HIVp with respect to the +x direction diffusion flux of xk263 can affect ligand association. Therefore, we examined two HIVp orientations, perpendicular and parallel with the ligand flux (**Figure 2**). We modeled 60 different systems (**Figure 4**). We chose the values of xk263 diffusion flux that gradually show the changes of the modeled values with the flux.

Environmental Factors Involved in Ligand Diffusion and Distribution

Intermolecular Interactions Between xk263 and CH_3-SAM

We first examined the interactions between xk263 and CH_3-SAM when the ligand diffusion flux varies. **Table 2** shows weak van der Waals (vdW) interactions between xk263 and CH_3-SAM regardless of the flux, which resulted in no significant adsorption of xk263 on SAM (**Figure 6A**). Because of the hydrophobic nature of CH_3-SAM and xk263, it is not surprising that the electrostatic interaction was close to zero.

Notably, our BD movement yielded the translational diffusion coefficient of xk263, $D_{lig} = 5.58 \pm 0.33 \times 10^{-6}$ cm^2/s when a system had no SAM and no ligand diffusion flux. The modeled ligand diffusion coefficient is in good agreement with analytical values, $D_{analytical} = 4.91 \times 10^{-6}$ cm^2/s, which also validates our simulation setting and BD algorithm (**Supplementary Methods**). Although the intermolecular vdW was weak, the attractions still affected the diffusion of xk263 molecules, and the molecule diffused a little slower in the presence of SAM with or without xk263 diffusion flux (**Figure 5**, **Supplementary Table 1**). Because the flow rate is inversely proportional to the viscosity (Pfitzner, 1976), the diffusion coefficient of xk263 increased with increasing

TABLE 2 | Interaction energies between xk263 and CH_3-SAM in 7 different xk263 diffusion fluxes.

xk263 diffusion flux (× 10^{22}/s.m^2)	E_{elec} (kcal/mol)	E_{vdW} (kcal/mol)
0	−0.045 ± 0.14	−1.126 ± 0.89
0.45	−0.056 ± 0.12	−1.098 ± 0.65
1.09	−0.010 ± 0.11	−1.547 ± 1.14
3.43	0.015 ± 0.14	−1.194 ± 0.95
5.62	0.003 ± 0.11	−1.488 ± 0.96
11.29	0.006 ± 0.10	−1.273 ± 1.13
17.20	0.005 ± 0.13	−1.069 ± 0.83

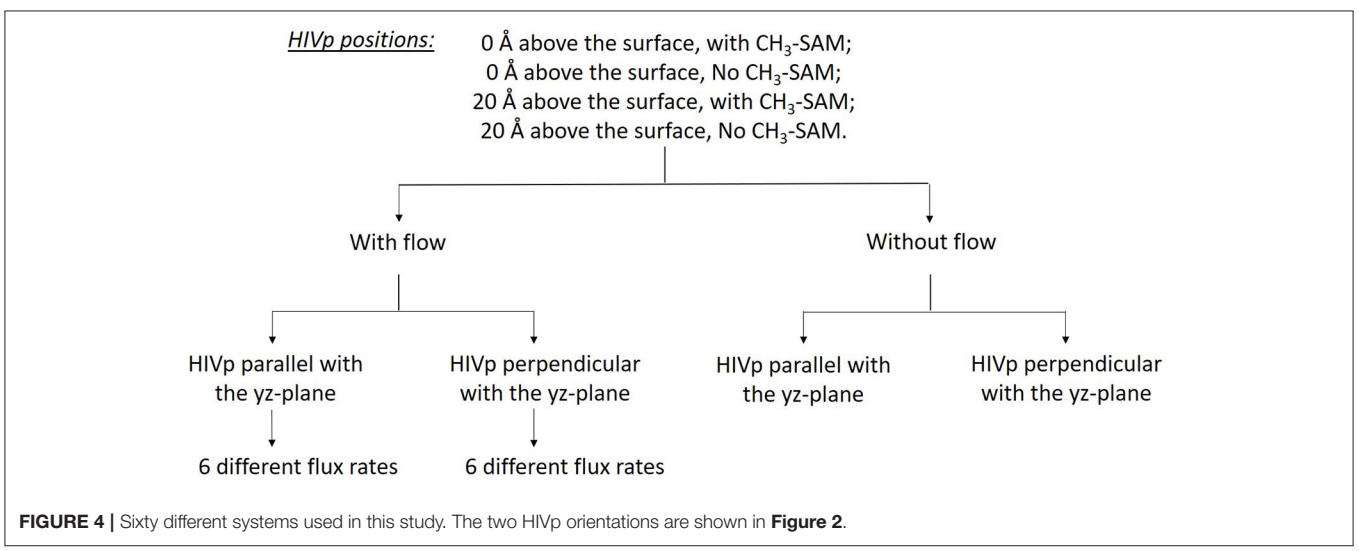


FIGURE 4 | Sixty different systems used in this study. The two HIVp orientations are shown in **Figure 2**.

xk263 diffusion flux, as anticipated. However, because the water flow rate was not equal to the xk263 diffusion flux, we did not see a simple linear relationship between the two values.

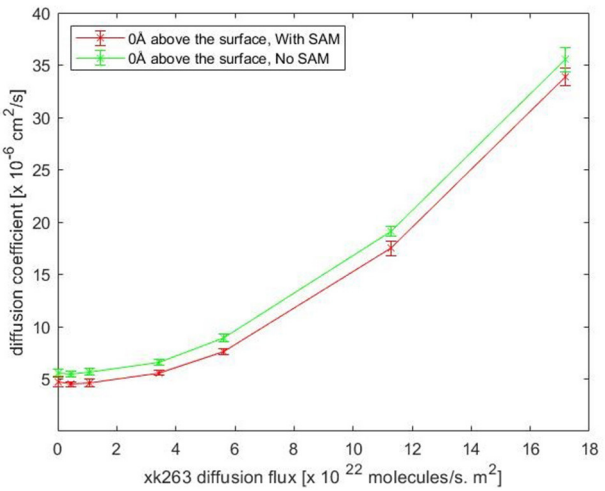


FIGURE 5 | Plot of translational diffusion coefficient of xk263. Diffusion coefficient increases with xk263 diffusion flux. (see **Supplementary Table 1** for numerical data).

Distribution of xk263 in the Simulation Tube

We further investigated whether the environment may affect the concentration gradient of xk263 along the z-direction when the molecule diffuses in the square tube. The system was partitioned into slices of 20 Å each along z-direction, and **Figure 6** shows the distribution of xk263 as a function of their distance above the surface. The distribution of xk263 for a slice, g(z), is the ratio of number of observed xk263 and number of expected xk263 approximated from a random diffusional motion. The systems include HIVp with one orientation shown in **Figure 2A**. When there was no SAM and the xk263 diffusion flux was small (green line in **Figure 6A**), the distribution of xk263 was the same throughout the z-direction (distribution ratio = ~1). In contrast, when CH_3-SAM was present (**Figures 6A–C**), xk263 was double that in the first slide (within 20 Å of the SAM) when diffusion flux of xk263 was $<1.09 \times 10^{22}$ molecules/s m^2. The results suggest that even if the intermolecular attraction between xk263 and CH_3-SAM is weak, the concentration can be increased by 2-fold, which helps to increase the probability of a molecular encounter. Studies for various systems also showed that local molecular concentration can be effected by a surface; for example, surface catalyzed biomolecular reaction using nanoreactors, nucleotidase co-localization, and absorption of biomacromolecules on hydrogels (Roa et al., 2017; Pérez-Mas et al., 2018; Rahmaninejad et al., 2020).

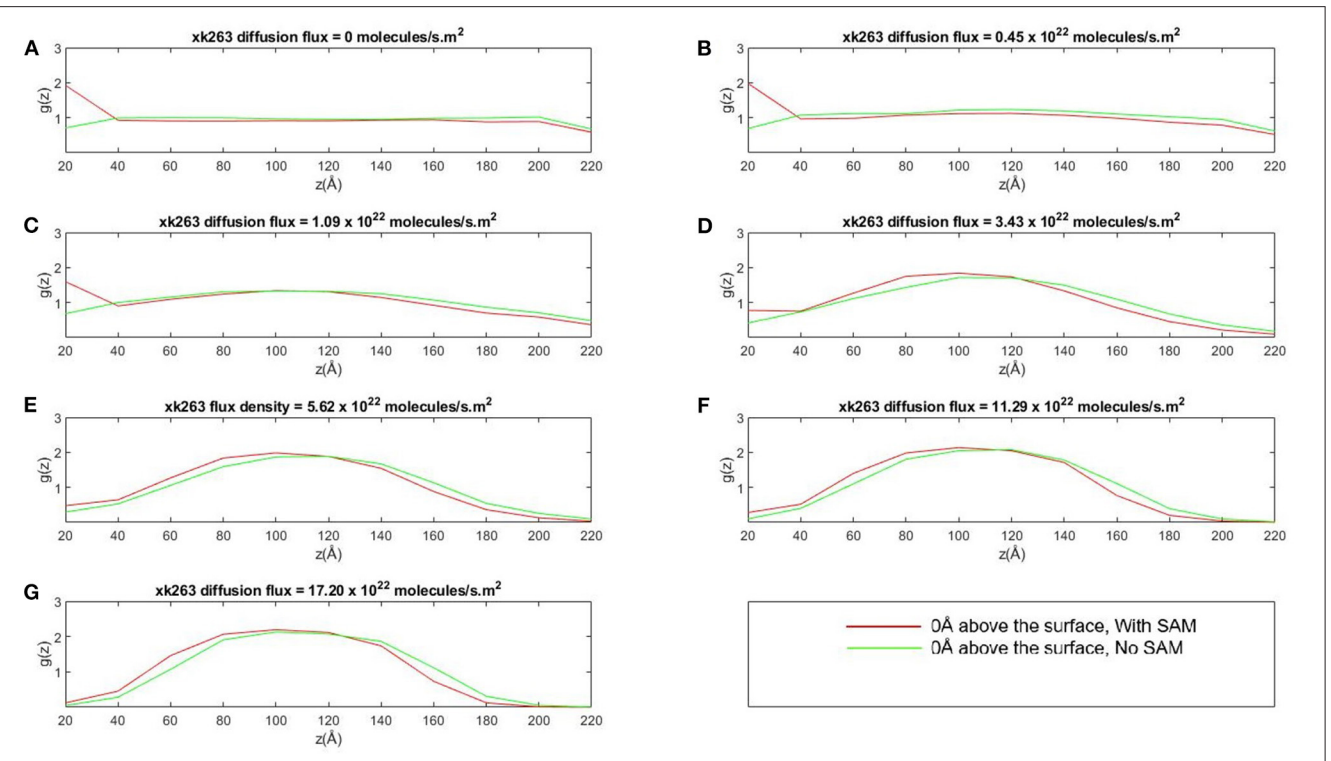


FIGURE 6 | Plot of concentration distribution of xk263, g(z), as a function of the distance above the CH_3-SAM, z, in seven different diffusion fluxes **(A–G)**. The space was partitioned into equally sized slices with a height of 20 Å. g(z) is the ratio of number of observed xk263 and the number of expected xk263 approximated from a random diffusional motion.

TABLE 3 | Average association time (t_{avg}) with different xk263 diffusion fluxes.

xk263 diffusion flux (× 10^{22} /s.m^2)	HIVp perpendicular to the yz–plane, t_{avg} (ns)				HIVp parallel to the yz-plane, t_{avg} (ns)			
	0Å above the surface, With SAM	0Å above the surface, No SAM	20Å above the surface, With SAM	20Å above the surface, No SAM	0Å above the surface, With SAM	0Å above the surface, No SAM	20Å above the surface, With SAM	20Å above the surface, No SAM
0	1491.55	1064.41	1213.45	901.99	1405.77	1084.49	1124.5	905.21
0.45	126.50	122.00	108.70	110.17	145.54	124.90	128.30	111.68
1.09	55.56	51.93	43.61	45.28	53.71	50.05	47.98	44.65
3.43	18.73	18.83	15.042	15.57	17.77	17.83	16.23	15.53
5.62	11.08	10.48	10.31	9.47	10.78	11.01	9.75	9.55
11.29	5.62	5.52	5.52	4.80	5.53	5.38	4.89	4.89
17.20	3.69	3.59	3.33	3.25	3.58	3.69	3.27	3.26

TABLE 4 | Inverse of average association time per molarity and percentage of successful binding with different xk263 diffusion fluxes.

xk263 diffusion flux (× 10^{22} /s.m^2)	0Å above the surface, With SAM		0Å above the surface, No SAM		20Å above the surface, With SAM		20Å above the surface, No SAM	
	$1/t_{avg}$ perM (/nsM)	Association percentage	$1/t_{avg}$ per M (/nsM)	Association percentage	$1/t_{avg}$ per M (/nsM)	Association percentage	$1/t_{avg}$ per M (/nsM)	Association percentage
HIVp perpendicular to the yz-plane								
0	6.09	18.59	8.54	18.14	7.49	18.67	10.08	19.18
0.45	71.86	3.02	74.52	3.82	83.63	3.53	82.52	4.24
1.09	163.63	1.22	175.07	1.73	208.45	1.67	200.75	2.16
3.43	485.44	0.52	482.66	0.64	604.35	1.00	583.76	0.89
5.62	820.41	0.35	867.45	0.33	881.41	0.90	959.31	0.65
11.29	1618.17	0.18	1646.61	0.14	1646.43	0.70	1892.99	0.43
17.20	2461.66	0.11	2532.29	0.07	2732.55	0.46	2793.33	0.34
HIVp parallel to the yz-plane								
0	6.47	17.15	8.38	17.85	8.08	18.99	10.04	20.61
0.45	62.46	3.25	72.78	3.88	70.86	3.43	81.40	4.22
1.09	169.25	1.21	181.64	1.71	189.48	1.86	203.61	1.92
3.43	511.39	0.38	509.78	0.48	559.99	0.71	585.19	0.75
5.62	842.87	0.22	825.62	0.21	931.88	0.40	952.29	0.42
11.29	1645.34	0.05	1690.36	0.05	1858.74	0.14	1858.13	0.15
17.20	2538.08	0.02	2461.19	0.02	2781.88	0.06	2788.36	0.07

The SAM retains xk263 near the surface only when the diffusion flux of the ligand was small. When the flux increases, the local concentration of xk263 near the surface decreased, and xk263 had the highest concentration near the middle of the z-direction (see ∼100 Å in **Figures 6D–G**). The distribution of concentration gradient along the z-direction was the same with or without SAM, which suggests that the weak intermolecular interactions can be altered easily when xk263 has diffusion flux in the x-direction. Our results are consistent with experiments showing the quadratic curve of concentration gradients with flow (Chen et al., 2017). With larger diffusion flux, a significantly smaller amount of xk263 diffused near the surface where HIVp locates, and the reduced distribution also contributed toward probability of small xk263–HIVp association, discussed later.

HIVp–xk263 Association in Different Environments

HIVp Immobilized on the Surface

We modeled the average xk263–HIVp association time and the binding percentage when HIVp was 0 Å above the surface with SAM (**Tables 3**, **4** with SAM) or without SAM. Theoretically, if xk263 employs only 3-dimensional diffusion within a sphere, the average association time to bind to HIVp located in the center of the sphere is 2648.4 ns when the system has no external flux (**Supplementary Methods**) (Adam and Delbrück, 1968). The association time modeled by our simulation is 1491.55 ns is close to the theoretical value. The faster binding time than theory is due to use of a box, and xk263 did not need to search the whole sphere to associate with HIVp. Notably, existing association rate theories describe system without diffusion flux (Berg and Purcell, 1977; Szabo et al., 1980; Shoup and Szabo,

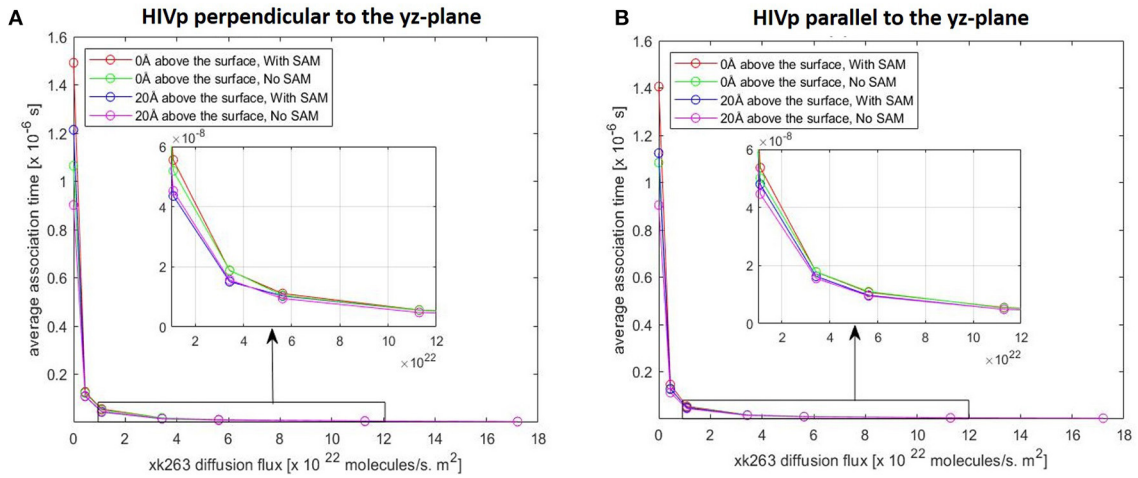


FIGURE 7 | Plot of average association time vs. xk263 diffusion flux. HIVp is placed **(A)** perpendicular to the yz-plane and **(B)** parallel to the yz-plane; Similar data imply no effect of HIVp orientation on the association time. In each plot, coinciding lines (red, green, blue, and magenta) show no effect of hydrophobic CH_3-SAM or height of the receptor from the surface.

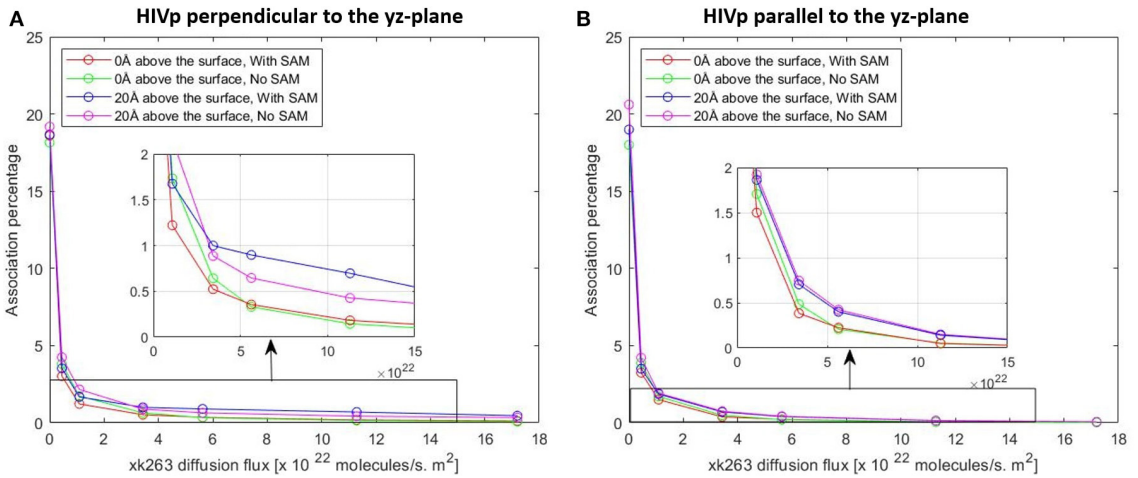


FIGURE 8 | Plot of association percentage vs. xk263 diffusion flux under two different HIVP's orientation **(A,B)**. Association percentage is defined as the fraction of successful binding from the total number of trajectories in the system.

1982) and further theoretical work in molecular association with buffer flowing is needed. Although the interactions between xk263 and CH_3-SAM were small, to examine the environment effects, we turned off the intermolecular vdW and electrostatic interactions, so the system was in a confined space without intermolecular attractions or repulsion (without SAM). The x-direction diffusion flux of xk263 gradually increased from zero to 1.7×10^{23} molecules/s m^2, and the average xk263–HIVp association time decreased non-linearly. The correlation was observed in a previous study, showing that the association time for human immunoglobulin G (IgG) and *Staphylococcus aureus* protein A (SpA) decreased with increasing flow rate (Ogi et al., 2008). The average association time did not differ without or with SAM in the presence of diffusion flux (**Table 3**). This is not surprising because the CH_3-SAM provides only weak attractions to xk263, and the surface neither increased the local concentration of xk263 to assist ligand association nor retained the ligand from binding to HIVp (**Figure 6**) (Cholko et al., 2020). However, if the attraction between a ligand and surface is large, it may affect experimental results. As a result, existing experimental studies considered the choice of the surface for the sensor or SPR to optimize experimental sensitivity and accuracy. For example, phospholipid/alkane bilayers might be a better option for molecular binding in biomolecular systems using SPR (Plant et al., 1995). These phospholipid derivatized surfaces may bring non-specific interactions for xk263 which is absent in our current model, which can result in slightly longer association time.

We report the inverse of association time per molarity and the association percentage (**Table 4**). When xk263 was freely diffused in the tube without diffusion flux, the association rate

approximated using the average association time, 1/1491.55 ns per M, yielded $k_{on} = 7.8 \times 10^9$ 1/Ms. Using the diffusion coefficient from the BD run (**Supplementary Table 2**), our simulated k_{on} was approximately a half of the analytical value, $k_{on} = 4\pi RD = 1.2 \times 10^{10}$ 1/Ms because the analytical formula does not consider molecular geometry in association. Our modeled k_{on} was also slightly slower than that measured using SPR for another highly similar cyclic urea compound, 2.5×10^{10} 1/Ms (Markgren et al., 2002). The first phase in SPR utilizes buffer flowing and detects binding signals, which require successful collision to have the correct ligand and HIVp orientation and enough energy. The flow may influence successful collision, especially on biomolecular systems which usually have complex molecular geometry. Of note, the inverse association time increased linearly with xk263 diffusion flux (**Supplementary Figure 1**), which suggests that the association rate may also increase linearly with the flow. However, our modeled diffusion coefficients (**Supplementary Table 2**) show an exponential increase with the xk263 diffusion flux (**Figure 5**). When particles have flux moving along the +x direction, the ligand did not freely diffuse in all directions, and the equation for the diffusion-controlled association system, $k_{on} = 4\pi RD$, was no longer suitable to approximate the association rate constants. Therefore, we cannot directly estimate k_{on} from the increased diffusion coefficient.

The additional +x direction drift velocity also reduced the initial encounter probability. When the flux was $>3.43 \times 10^{22}$ molecules/s.m^2, xk263 passed the tube without sufficient time to freely diffuse in the y and z directions, which resulted in reduced search space and tremendously decreased association percentage. Large diffusion flux also yielded fewer xk263 molecules staying near both the SAM and HIVp surface. Notably, xk263 may diffuse on the surface of HIVp before binding to the pocket. The xk263 diffusion flux in the +x direction perturbed the interactions between xk263 and HIVp, and the weakened intermolecular attractions also led to decreased association percentage.

In typical experimental settings, an immobilized protein can have multiple orientations. Therefore, we chose two representative orientations—the binding pocket of HIVp perpendicular or parallel to the flux of xk263—to examine whether the orientation against the flow affects ligand binding. **Tables 3**, **4**, **Figures 7**, **8** show no difference in average ligand association time and binding percentage with the two orientations. Because the binding pocket of HIVp is widely accessible for the ligand, the impact of the orientation was insignificant. However, some proteins exhibit a certain direction for ligands to enter the binding pocket, and their orientation against the flux of the ligands can affect the ligand–protein association.

Artificially Localized HIVp 20 Å Above the Surface

Although different HIVp orientations did not affect xk263 association time, existing studies showed that the dimer interface (the bottom part of HIVp) is the most popular region for ligands' initial encounter of HIVp (Roberts and Chang, 2016). After reaching the dimer interface, the ligand can undergo surface diffusion to reach the binding pocket. When we placed HIVp 0 Å above the surface, this bottom region became partially inaccessible to xk263. Therefore, immobilized HIVp was artificially placed 20 Å above the surface so that xk263 can easily reach the bottom region. When the whole HIVp surface was accessible to xk263, the association time was reduced by 10–20% when the xk263 diffusion flux was $<1.09 \times 10^{22}$ molecules/s.m^2, which suggests that the additionally available bottom region accelerates xk263 binding (**Table 3**) and slightly increases the association percentage (**Table 4**). Even when HIVp was placed 20 Å above the surface, the location was not far enough to eliminate the intermolecular xk263–SAM interactions. As a result, xk263 spent longer time near SAM when xk263 diffusion flux was small, thus resulting in longer time to associate with HIVp when SAM was present rather than absent. When the flux exists, the association time was the same when the SAM was present or absent, regardless the position of HIVp (**Table 3**).

CONCLUSION

Ligand–protein encounters can occur in any environment that may provide additional intermolecular interactions or the transporting forces to the ligand. For example, in experimental settings such as using SPR to study ligand–protein binding, the choice of the surface and the flow rate are all optimized for measurements. In this study, we used BD simulations to investigate the effect of the CH_3-SAM surface on xk263–HIVp association. The non-polar surface provided weak intermolecular vdW attractions with xk263 to slightly increase the ligand binding time. The effects quickly vanished when xk263 had an x-direction diffusion flux, and the association time decreased when the diffusion flux increased regardless of the presence of SAM or only a special plane. With no diffusion flux, xk263 was twice more concentrated within 20 Å of the SAM surface because of the xk263–SAM interactions. The concentration gradient did not increase binding time but increased xk263–HIVp binding probability. When the xk263 diffusion flux increased, the middle region of the square tube had the highest xk263 concentration, which is the same as existing experiments showing a quadratic curve of concentration gradients with flow. The results also show that when HIVp was placed on the surface (0 Å above the surface), the bottom part of the protein, a known high-probability site of the xk263–HIVp first encounter, was not accessible to xk263. Because xk263 could bind non-specifically on the HIVp surface and then utilize surface diffusion to reach the binding site, occluding this bottom region increased ligand binding time by 10–20% in the static environment. However, with large xk263 diffusion flux, all the weak intermolecular interactions and searching along the HIVp surface were eliminated, which resulted in a fast association time and small binding probability. This work brings insights into how ligand diffusion flux and the environment may affect ligand–protein association.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

SK ran and analyzed the simulations, produced figures, and wrote the manuscript. C-eC designed experiments, analyzed the simulation, and wrote the manuscript. Both authors contributed to the article and approved the submission version.

FUNDING

This study was supported by the US National Institutes of Health (GM-109045), US National Science Foundation (MCB-1932984), and NSF national supercomputer centers (TG-CHE130009).

ACKNOWLEDGMENTS

We thank Dr. Jason Quan Cheng for helpful discussions regarding SPR experiments. We thank Tim Cholko for help in modifying the GeomBD program and Jianan Sun for providing HIVp open structures.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.659687/full#supplementary-material

REFERENCES

Adam, G., and Delbrück, M. (1968). Reduction of dimensionality in biological diffusion processes. *Struct. Chem. Mol. Biol.* 198, 198–215.

Andrews, S. S. (2017). Smoldyn: particle-based simulation with rule-based modeling, improved molecular interaction and a library interface. *Bioinformatics* 33, 710–717. doi: 10.1093/bioinformatics/btw700

Baron, R., and McCammon, J. A. (2013). Molecular recognition and ligand association. *Annu. Rev. Phys. Chem.* 64, 151–175 doi: 10.1146/annurev-physchem-040412-110047

Berg, H. C., and Purcell, E. M. (1977). Physics of chemoreception. *Biophys. J.* 20, 193–219. doi: 10.1016/S0006-3495(77)85544-6

Bernetti, M., Masetti, M., Rocchia, W., and Cavalli, A. (2019). Kinetics of drug binding and residence time. *Annu. Rev. Phys. Chem.* 70, 143–171. doi: 10.1146/annurev-physchem-042018-052340

Britton, J., Majumdar, S., and Weiss, G. A. (2018). Continuous flow biocatalysis. *Chem. Soc. Rev.* 47, 5891–5918. doi: 10.1039/C7CS00906B

Chang, C.-E., Shen, T., Trylska, J., Tozzini, V., and McCammon, J. A. (2006). Gated binding of ligands to HIV-1 protease: brownian dynamics simulations in a coarse-grained model. *Biophys. J.* 90, 3880–3885. doi: 10.1529/biophysj.105.074575

Chen, X., Hu, Z., Zhang, L., Yao, Z., Chen, X., Zheng, Y., et al. (2017). Numerical and experimental study on a microfluidic concentration gradient generator for arbitrary approximate linear and quadratic concentration curve output. *Int. J. Chem. Reactor Eng.* 16. doi: 10.1515/ijcre-2016-0204

Cholko, T., Barnum, J., and Chang, C. E. (2020). Amyloid-β (Aβ42) peptide aggregation rate and mechanism on surfaces with widely varied properties: insights from brownian dynamics simulations. *J. Phys. Chem. B* 124, 5549–5558. doi: 10.1021/acs.jpcb.0c02926

Cholko, T., Kaushik, S., and Chia-en, A. C. (2019). Dynamics and molecular interactions of single-stranded DNA in nucleic acid biosensors with varied surface properties. *Phys. Chem. Chem. Phys.* 21, 16367–16380. doi: 10.1039/C9CP02441G

De Biase, P. M., Markosyan, S., and Noskov, S. (2015). BROMOC suite: Monte Carlo/Brownian dynamics suite for studies of ion permeation and DNA transport in biological and artificial pores with effective potentials. *J. Comput. Chem.* 36, 264–271. doi: 10.1002/jcc.23799

Di Cera, E. (2017). Mechanisms of ligand binding. *Biophys. Rev.* 1:011303. doi: 10.1063/5.0020997

Dill, K., and Bromberg, S. (2012). *Molecular driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*. New York, NY: Garland Science. doi: 10.4324/9780203809075

Długosz, M., Zieliński, P., and Trylska, J. (2011). Brownian dynamics simulations on CPU and GPU with BD_BOX. *J. Comput. Chem.* 32, 2734–2744. doi: 10.1002/jcc.21847

Ershov, P. V., Mezentsev, Y. V., Kaluzhskiy, L. A., and Ivanov, A. S. (2020). Phenanthridine derivatives as potential HIV-1 protease inhibitors. *Biomed. Rep.* 13:66. doi: 10.3892/br.2020.1373

Ghosh, A. K., Osswald, H. L., and Prato, G. (2016). Recent progress in the development of HIV-1 protease inhibitors for the treatment of HIV/AIDS. *J. Med. Chem.* 59, 5172–5208. doi: 10.1021/acs.jmedchem.5b01697

Hinman, S. S., McKeating, K. S., and Cheng, Q. (2018). Surface plasmon resonance: material and interface design for universal accessibility. *Anal. Chem.* 90:19. doi: 10.1021/acs.analchem.7b04251

Huang, X., Britto, M. D., Kear-Scott, J. L., Boone, C. D., Rocca, J. R., Simmerling, C., et al. (2014). The role of select subtype polymorphisms on HIV-1 protease conformational sampling and dynamics. *J. Biol. Chem.* 289, 17203–17214. doi: 10.1074/jbc.M114.571836

Huang, Y. M., Raymundo, M. A., Chen, W., and Chang, C. A. (2017). Mechanism of the association pathways for a pair of fast and slow binding ligands of HIV-1 protease. *Biochemistry* 56, 1311–1323. doi: 10.1021/acs.biochem.6b01112

Huber, G. A., and McCammon, J. A. (2010). Browndye: a software package for Brownian dynamics. *Comput. Phys. Commun.* 181, 1896–1905. doi: 10.1016/j.cpc.2010.07.022

Huber, G. A., and McCammon, J. A. (2019). Brownian dynamics simulations of biological molecules. *Trends Chem.* 1, 727–738. doi: 10.1016/j.trechm.2019.07.008

Kang, M., Roberts, C., Cheng, Y., and Chang, C. E. (2011). Gating and intermolecular interactions in ligand-protein association: coarse-grained modeling of HIV-1 protease. *J. Chem. Theory Comput.* 7, 3438–3446. doi: 10.1021/ct2004885

Katoh, E., Louis, J. M., Yamazaki, T., Gronenborn, A. M., Torchia, D. A., and Ishima, R. (2003). A solution NMR study of the binding kinetics and the internal dynamics of an HIV-1 protease-substrate complex. *Protein Sci.* 12, 1376–1385. doi: 10.1110/ps.0300703

Kohl, N. E., Emini, E. A., Schleif, W. A., Davis, L. J., Heimbach, J. C., Dixon, R., et al. (1988). Active human immunodeficiency virus protease is required for viral infectivity. *Proce. Natl Acad. Sci. U.S.A.* 85, 4686–4690. doi: 10.1073/pnas.85.13.4686

Lam, P., Jadhav, P. K., Eyermann, C. J., Hodge, C. N., Ru, Y., Bacheler, L. T., et al. (1994). Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science* 263, 380–384. doi: 10.1126/science.8278812

Li, D., Liu, M. S., Ji, B., Hwang, K. C., and Huang, Y. (2012). Identifying the molecular mechanics and binding dynamics characteristics of potent inhibitors to HIV-1 protease. *Chem. Biol. Drug Design* 80, 440–454. doi: 10.1111/j.1747-0285.2012.01417.x

Lin, Y.-C., Kim, W. K., and Dzubiella, J. (2020). Coverage fluctuations and correlations in nanoparticle-catalyzed diffusion-influenced bimolecular reactions. *J. Phys. Chem. C* 124, 24204–24214. doi: 10.1021/acs.jpcc.0c06898

Madura, J. D., Briggs, J. M., Wade, R. C., Davis, M. E., Luty, B. A., Ilin, A., et al. (1995). Electrostatics and diffusion of molecules in solution: simulations

with the University of Houston Brownian Dynamics program. *Comput. Phys. Commun.* 91, 57–95. doi: 10.1016/0010-4655(95)00043-F

Markgren, P.-O., Schaal, W., Hämäläinen, M., Karlén, A., Hallberg, A., Samuelsson, B., et al. (2002). Relationships between structure and interaction kinetics for HIV-1 protease inhibitors. *J. Med. Chem.* 45, 5430–5439. doi: 10.1021/jm0208370

Martinez, M., Bruce, N. J., Romanowska, J., Kokh, D. B., Ozboyaci, M., Yu, X., et al. (2015). SDA 7: A modular and parallel implementation of the simulation of diffusional association software. *J. Comput. Chem.* 36, 1631–1645. doi: 10.1002/jcc.23971

McCammon, J. A. (2011). Gated diffusion-controlled reactions. *BMC Biophys.* 4, 1–5. doi: 10.1186/2046-1682-4-4

Northrup, S., Laughner, T., and Stevenson, G. (1999). *MacroDox Macromolecular Simulation Program*. Cookeville, TN: Tennessee Technological University, Department of Chemistry.

Northrup, S. H., Allison, S. A., and McCammon, J. A. (1984). Brownian dynamics simulation of diffusion-influenced bimolecular reactions. *J. Chem. Phys.* 80, 1517–1524. doi: 10.1063/1.446900

Ogi, H., Fukunishi, Y., Omori, T., Hatanaka, K., Hirao, M., and Nishiyama, M. (2008). Effects of flow rate on sensitivity and affinity in flow injection biosensor systems studied by 55-MHz wireless quartz crystal microbalance. *Anal. Chem.* 80, 5494–5500. doi: 10.1021/ac800459g

Ozbabacan, S. A., Gursoy, A., Keskin, O., and Nussinov, R. (2010). Conformational ensembles, signal transduction and residue hot spots: application to drug discovery. *Curr. Opin. Drug Discov. Devel.* 13, 527–537.

Pang, X., and Zhou, H.-X. (2017). Rate constants and mechanisms of protein–ligand binding. *Ann. Rev. Biophys.* 46, 105–130. doi: 10.1146/annurev-biophys-070816-033639

Pérez-Mas, L., Martín-Molina, A., Quesada-Pérez, M., and Moncho-Jordá A. (2018). Maximizing the absorption of small cosolutes inside neutral hydrogels: steric exclusion versus hydrophobic adhesion. *Phys. Chem. Chem. Phys.* 20, 2814–2825. doi: 10.1039/C7CP07679G

Pfitzner, J. (1976). Poiseuille and his law. *Anaesthesia* 31, 273–275. doi: 10.1111/j.1365-2044.1976.tb11804.x

Planchestainer, M., Contente, M. L., Cassidy, J., Molinari, F., Tamborini, L., and Paradisi, F. (2017). Continuous flow biocatalysis: production and in-line purification of amines by immobilised transaminase from Halomonas elongata. *Green Chem.* 19, 372–375. doi: 10.1039/C6GC01780K

Plant, A. L., Brighamburke, M., Petrella, E. C., and Oshannessy, D. J. (1995). Phospholipid/alkanethiol bilayers for cell-surface receptor studies by surface plasmon resonance. *Anal. Biochem.* 226, 342–348. doi: 10.1006/abio.1995.1234

Prabowo, B. A., Purwidyantri, A., and Liu, K.-C. (2018). Surface plasmon resonance optical sensor: a review on light source technology. *Biosensors* 8:80. doi: 10.3390/bios8030080

Rahmaninejad, H., Pace, T., Bhatt, S., Sun, B., and Kekenes-Huskey, P. (2020). Co-localization and confinement of ecto-nucleotidases modulate extracellular adenosine nucleotide distributions. *PLoS Comput. Biol.* 16:e1007903. doi: 10.1371/journal.pcbi.1007903

Roa, R., Kim, W. K., Kanduc M., Dzubiella, J., and Angioletti-Uberti, S. (2017). Catalyzed bimolecular reactions in responsive nanoreactors. *ACS Catal.* 7, 5604–5611. doi: 10.1021/acscatal.7b01701

Roberts, C. C., and Chang, C.-E. (2016). Analysis of ligand–receptor association and intermediate transfer rates in multienzyme nanostructures with all-atom brownian dynamics simulations. *J. Phys. Chem. B* 120, 8518–8531 doi: 10.1021/acs.jpcb.6b02236

Roberts, C. C., and Chang, C. E. (2015). Modeling of enhanced catalysis in multienzyme nanostructures: effect of molecular scaffolds, spatial organization, and concentration. *J. Chem. Theory Comput.* 11, 286–292 doi: 10.1021/ct5007482

Saadat, A., and Khomami, B. (2015). Matrix-free Brownian dynamics simulation technique for semidilute polymeric solutions. *Phys. Rev. E* 92:033307 doi: 10.1103/PhysRevE.92.033307

Schöneberg, J., and Noé, F. (2013). ReaDDy-a software for particle-based reaction-diffusion dynamics in crowded cellular environments. *PLoS ONE* 8:e74261(e2013). doi: 10.1371/journal.pone.0074261

Shoup, D., and Szabo, A. (1982). Role of diffusion in ligand binding to macromolecules and cell-bound receptors. *Biophys. J.* 40, 33–39. doi: 10.1016/S0006-3495(82)84455-X

Spinelli, S., Liu, Q., Alzari, P., Hirel, P., and Poljak, R. (1991). The three-dimensional structure of the aspartyl protease from the HIV-1 isolate BRU. *Biochimie* 73, 1391–1396. doi: 10.1016/0300-9084(91)90169-2

Szabo, A., Schulten, K., and Schulten, Z. (1980). First passage time approach to diffusion controlled reactions. *J. Chem. Phys.* 72, 4350–4357. doi: 10.1063/1.439715

Szabo, A., Shoup, D., Northrup, S. H., and McCammon, J. A. (1982). Stochastically gated diffusion-influenced reactions. *J. Chem. Phys.* 77, 4484–4493. doi: 10.1063/1.444397

Tozzini, V., and McCammon, J. A. (2005). A coarse grained model for the dynamics of flap opening in HIV-1 protease. *Chem. Phys. Lett.* 413, 123–128. doi: 10.1016/j.cplett.2005.07.075

Votapka, L. W., Jagger, B. R., Heyneman, A. L., and Amaro, E. R. (2017). SEEKR: simulation enabled estimation of kinetic rates, a computational tool to estimate molecular kinetics and its application to trypsin–benzamidine binding. *J. Phys. Chem. B* 121, 3597–3606. doi: 10.1021/acs.jpcb.6b09388

Zotter, A., Bäuerle, F., Dey, D., Kiss, V., and Schreiber, G. (2017). Quantifying enzyme activity in living cells. *J. Biol. Chem.* 292, 15838–15848. doi: 10.1074/jbc.M117.792119

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 © 2021 Kaushik and Chang. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 19 May 2021
doi: 10.3389/fmolb.2021.663121

HEMNMA-3D: Cryo Electron Tomography Method Based on Normal Mode Analysis to Study Continuous Conformational Variability of Macromolecular Complexes

Mohamad Harastani[1], *Mikhail Eltsov*[2], *Amélie Leforestier*[3] *and Slavica Jonic*[1]*

[1] IMPMC-UMR 7590 CNRS, Sorbonne Université, Muséum National d'Histoire Naturelle, Paris, France, [2] Department of Integrated Structural Biology, Institute of Genetics and Molecular and Cellular Biology, Illkirch, France, [3] Laboratoire de Physique des Solides, UMR 8502 CNRS, Université Paris-Saclay, Paris, France

OPEN ACCESS

Edited by:
Fabio Trovato,
Freie Universität Berlin, Germany

Reviewed by:
Dustin Morado,
Science for Life Laboratory (SciLifeLab), Sweden
Michael Habeck,
University Hospital Jena, Germany
Jose-Maria Carazo,
Consejo Superior de Investigaciones Científicas (CSIC), Spain

***Correspondence:**
Slavica Jonic
slavica.jonic@upmc.fr

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: 02 February 2021
Accepted: 09 April 2021
Published: 19 May 2021

Citation:
Harastani M, Eltsov M, Leforestier A and Jonic S (2021) HEMNMA-3D: Cryo Electron Tomography Method Based on Normal Mode Analysis to Study Continuous Conformational Variability of Macromolecular Complexes. Front. Mol. Biosci. 8:663121. doi: 10.3389/fmolb.2021.663121

Cryogenic electron tomography (cryo-ET) allows structural determination of biomolecules in their native environment (*in situ*). Its potential of providing information on the dynamics of macromolecular complexes in cells is still largely unexploited, due to the challenges of the data analysis. The crowded cell environment and continuous conformational changes of complexes make difficult disentangling the data heterogeneity. We present HEMNMA-3D, which is, to the best of our knowledge, the first method for analyzing cryo electron subtomograms in terms of continuous conformational changes of complexes. HEMNMA-3D uses a combination of elastic and rigid-body 3D-to-3D iterative alignments of a flexible 3D reference (atomic structure or electron microscopy density map) to match the conformation, orientation, and position of the complex in each subtomogram. The elastic matching combines molecular mechanics simulation (Normal Mode Analysis of the 3D reference) and experimental, subtomogram data analysis. The rigid-body alignment includes compensation for the missing wedge, due to the limited tilt angle of cryo-ET. The conformational parameters (amplitudes of normal modes) of the complexes in subtomograms obtained through the alignment are processed to visualize the distribution of conformations in a space of lower dimension (typically, 2D or 3D) referred to as space of conformations. This allows a visually interpretable insight into the dynamics of the complexes, by calculating 3D averages of subtomograms with similar conformations from selected (densest) regions and by recording movies of the 3D reference's displacement along selected trajectories through the densest regions. We describe HEMNMA-3D and show its validation using synthetic datasets. We apply HEMNMA-3D to an experimental dataset describing *in situ* nucleosome conformational variability. HEMNMA-3D software is available freely (open-source) as part of ContinuousFlex plugin of Scipion V3.0 (http://scipion.i2pc.es).

Keywords: cryo electron tomography, continuous conformational changes, flexible-reference alignment, normal mode analysis, nucleosomes *in situ*

1. INTRODUCTION

Cryogenic electron microscopy (cryo-EM) image collection and analysis technique referred to as single-particle analysis (SPA) allows near-atomic structural resolution of purified biomolecular complexes (*in vitro*). It is based on the principle of reconstructing a three dimensional (3D) structure from two dimensional (2D) parallel-beam projection images of vitrified specimens containing many copies of the same macromolecular complex at unknown orientations and positions. The 3D reconstruction requires extracting macromolecular complexes (particles) from the collected images into individual (single-particle) images and determining the particle orientation and position in every single-particle image. On the other hand, cryogenic electron tomography (cryo-ET) is gaining popularity for studying biomolecular complexes in their native environments (*in situ*). Cryo-ET requires the acquisition of multiple 2D projection images of the specimen in a range of orientations. In most practices, the specimen is physically rotated around a single axis (perpendicular to the electron beam) inside the cryo electron microscope. An image is collected at each tilting angle in a specific range (e.g., -60 to 60° with a step of 1°), yielding a tilt series representing 2D projections of the specimen. The tilt series is then used to computationally reconstruct a 3D volume called tomogram. A tomographic reconstruction typically contains hundreds of copies of a target biomolecular complex at unknown orientations and positions. These copies are then identified and extracted into individual (single-particle) volumes called subtomograms, either manually or semi-automatically (via template matching methods). Subtomograms suffer from a low signal-to-noise ratio (SNR), which is due to exposing the sample to a low electron dose during data acquisition in order to preserve the fragile biological structure. Additionally, subtomograms suffer from the so-called missing wedge artifacts, which are due to inability to include in the 3D reconstruction the images from all orientations (the maximum tilt angle in the microscope is usually limited to $\pm$ 60°). The missing wedge artifacts are often observed as elongation along the beam axis, blurring, and distracting caustics in the subtomograms. Due to the low SNR and the missing wedge artifacts, cryo-ET data processing is mainly based on rigid-body aligning and averaging many subtomograms to enhance the data quality and reveal the targeted biomolecular structure (Leigh et al., 2019).

The primary technique for macromolecular structural determination is so-called subtomogram averaging (StA), in which subtomograms are classified, rigid-body aligned and averaged into 3D density maps iteratively (Mahamid et al., 2016; Schur et al., 2016; Wan and Briggs, 2016; Albert et al., 2017; Böck et al., 2017; Bykov et al., 2017; Pfeffer et al., 2017; Riedel et al., 2017; Wan et al., 2017; Davies et al., 2018; Guo et al., 2018; Hutchings et al., 2018; Kovtun et al., 2018; Mosalaganti et al., 2018; Park et al., 2018; Kaplan et al., 2019; Rapisarda et al., 2019). However, with recent instrumentation and software development, more research moves in the direction of studying single-particle subtomograms individually (with no or a minimum of averaging) by developing new methods for denoising, missing wedge correction, and 3D reconstruction (Zhang and Ren, 2012; Moebel and Kervrann, 2020; Zhai et al., 2020).

Biomolecular complexes are not rigid but flexible entities with gradual (continuous) conformational transitions, and this flexibility is usually referred to as continuous conformational variability. If not properly taken into account, conformational heterogeneity limits the resolution of the resulting 3D structure. However, SPA research in the last decade has shown that disentangling the different conformations and identifying the conformational transitions from heterogeneous samples is valuable to study molecular mechanisms of action of complexes (Dashti et al., 2014; Jin et al., 2014; Zhou et al., 2015; Abeyrathne et al., 2016; Banerjee et al., 2016; Haselbach et al., 2018). The majority of available SPA computational methods rely on optimized biochemical specimen preparation protocols and data classification, and simplify the problem of conformational heterogeneity by assuming that the data can be classified into a small number of different conformations (Penczek et al., 2006, 2011; Fu et al., 2007; Elad et al., 2008; Scheres, 2012; Lyumkis et al., 2013). However, some SPA methods explicitly take into account continuous conformational variability and aim at determining the full conformational distribution (Dashti et al., 2014; Jin et al., 2014; Sorzano et al., 2014; Katsevich et al., 2015; Tagare et al., 2015; Frank and Ourmazd, 2016; Andén and Singer, 2018; Harastani et al., 2020). They represent images in a low-dimensional space, referred to as space of conformations or energy landscape, and allow a 3D visualization of conformational changes along trajectories in this space. For more information on SPA methods for continuous conformational variability analysis, the reader is referred to the recent reviews by Jonić (2017) and Sorzano et al. (2019).

Methods reported to deal with cryo-ET data heterogeneity are based on rigid-body alignment and can be classified into (i) post-alignment classification approaches, and (ii) simultaneous alignment and classification approaches (Förster et al., 2008; Scheres et al., 2009; Stölken et al., 2011; Xu et al., 2012; Chen et al., 2014; Bharat and Scheres, 2016; Himes and Zhang, 2018). In the first family, the starting point is usually the covariance matrix representing the similarities of each pair of aligned subtomograms. The covariance matrix serves as a basis for a classification technique with some variants including dimensionality reduction. The second family of methods is based on competitive alignment. An example of the competitive alignment is a multireference alignment in which a subtomogram is compared with a set of different references provided by an expert user based on a prior knowledge and, then, attributed to the reference that yields the highest similarity score. Another example is maximum-likelihood-based alignment, where each subtomogram contributes to all references with a probability. The main drawback of the post-alignment classification approaches is that the classification is heavily dependent on the alignment quality that degenerates with broadly heterogeneous specimens. The main drawback of the simultaneous alignment and classification methods is that the number of classes must be decided and set prior to the use of the methods. Besides, the methods that require prior knowledge of the specimen's anticipated conformations

are prone to overfitting and data misinterpretation. Finally, as macromolecular complexes are not rigid but flexible entities with continuous conformational transitions, particles assigned to the same class will rarely, if ever, have perfectly identical conformations. For more information regarding the classification based techniques for cryo-ET conformational heterogeneity, the reader is referred to a recent review on the available techniques by Castaño-Díez and Zanetti (2019).

Existing multivariate statistical analysis techniques adapted to cryo-ET analysis of continuous flexibility of particular systems have been used previously (e.g., Mattei et al., 2016). However, to the best of our knowledge, no method is currently available that has been specifically designed for cryo-ET analysis of continuous conformational variability of a general-case macromolecular complex. In this article, we present one such method, named HEMNMA-3D, which allows analyzing continuous conformational variability of macromolecular complexes by cryo-ET. It is inspired by HEMNMA, a method for continuous conformational variability analysis in SPA (Jin et al., 2014; Sorzano et al., 2014; Harastani et al., 2020). HEMNMA interprets the conformation in each cryo-EM single-particle image by comparing this image with 2D projections of a 3D reference (an atomic structure or a density map) deformed elastically using normal modes. Normal Mode Analysis (NMA) is a method for molecular mechanics simulation. One of its main applications is elastic deformation of an existing atomic structure of one conformation to fit an electron microscopy density map (EM map) of a different conformation of the same macromolecule, which is usually known as normal mode flexible fitting and allows obtaining atomic resolution models for the EM map (Tama et al., 2004a,b). However, as HEMNMA, HEMNMA-3D can use normal modes of an atomic structure or a density map. As in the case of the reference density-map structure in HEMNMA, HEMNMA-3D converts the density map into a collection of 3D Gaussian functions, referred to as pseudoatoms, and computes normal modes of this pseudoatomic structure, following the procedures described in Nogales-Cadenas et al. (2013), Jin et al. (2014), and Jonić and Sorzano (2016). HEMNMA-3D uses the atomic or pseudoatomic normal modes to elastically deform the 3D reference to match the conformation of the complex in the given series of subtomograms (3D data) (see **Figure 1**). More precisely, HEMNMA-3D uses a combination of elastic (based on normal modes) and rigid-body 3D-to-3D iterative alignments of the 3D reference to match the conformation, orientation, and position of the complex in each subtomogram, and includes compensation for the missing wedge. The conformational parameters (amplitudes of normal modes) of the complexes in subtomograms obtained through the alignment are then processed to visualize the distribution of conformations in a space of lower dimension (typically, 2D or 3D) referred to as space of conformations. This space allows a visually interpretable insight into the dynamics of complexes, by calculating 3D averages of subtomograms with similar conformations from selected (densest) regions and by recording movies of the 3D reference's displacement along selected trajectories in the densest regions.

In this article, we describe HEMNMA-3D and show its validation using synthetic datasets. Additionally, we show an application of HEMNMA-3D with an experimentally obtained dataset for *in situ* nucleosome conformational variability. HEMNMA-3D software is available freely (open-source) as part of ContinuousFlex plugin of Scipion V3.0 (http://scipion.i2pc.es). The article is organized as follows: section 2 describes building blocks of HEMNMA-3D workflow, notably, in 2.5 we describe the software developed for this method. In section 3, we present (i) the process of synthesis of test datasets and HEMNMA-3D validation using these synthetic test data, and (ii) use of HEMNMA-3D with experimental, *in situ* nucleosome data, and we discuss these results. The conclusions are provided in section 4.

2. MATERIALS AND METHODS

The flowchart in **Figure 2** describes the workflow of the proposed method, which was inspired by the workflow of HEMNMA (Jin et al., 2014; Harastani et al., 2020). A graphical summary of the method is presented in **Figure 3**. The workflow comprises the following steps: (1) Input: the input to the method are a reference structure and a set of subtomograms. In the case where the reference structure is a density map (a 3D volume, such as an EM map or a subtomogram average), a conversion to 3D Gaussian functions (pseudoatoms) takes place. (2) Normal mode analysis of the reference atomic structure or the reference pseudoatomic structure (obtained by converting the reference density map into 3D Gaussian functions in the previous step). (3) Combined iterative elastic and rigid-body 3D-to-3D alignment of the reference structure with each input subtomogram independently from other subtomograms, with missing wedge compensation. (4) Visualization of the computed conformations, after projecting the conformational parameters obtained for all subtomograms onto a low-dimensional space. In the remaining part of this section, we describe these steps in more detail. Please note that the first two steps of the workflow are exactly as those of HEMNMA and were thoroughly presented, tested and discussed in our previously published works on HEMNMA, its tools and applications (Nogales-Cadenas et al., 2013; Jin et al., 2014; Sorzano et al., 2014; Jonić and Sorzano, 2016; Harastani et al., 2020). However, for completeness of the present article, we here recall their basic principles. We close this section by a brief description of the software implemented for HEMNMA-3D.

2.1. Input Reference and Conversion of Reference Density Maps Into Pseudoatoms

A reference structure of the molecule targeted in the subtomograms can be used in the form of an atomic model (PDB formatted files) or a density map, such as an EM map (SPA reconstruction) or a subtomogram average (obtained using classical StA without taking into account conformational heterogeneity). Although our method can be used with both atomic and density-map reference structures, one should prefer the use of a reference density map from the data at hand, if it can be obtained. If a reference density map is used, it must be

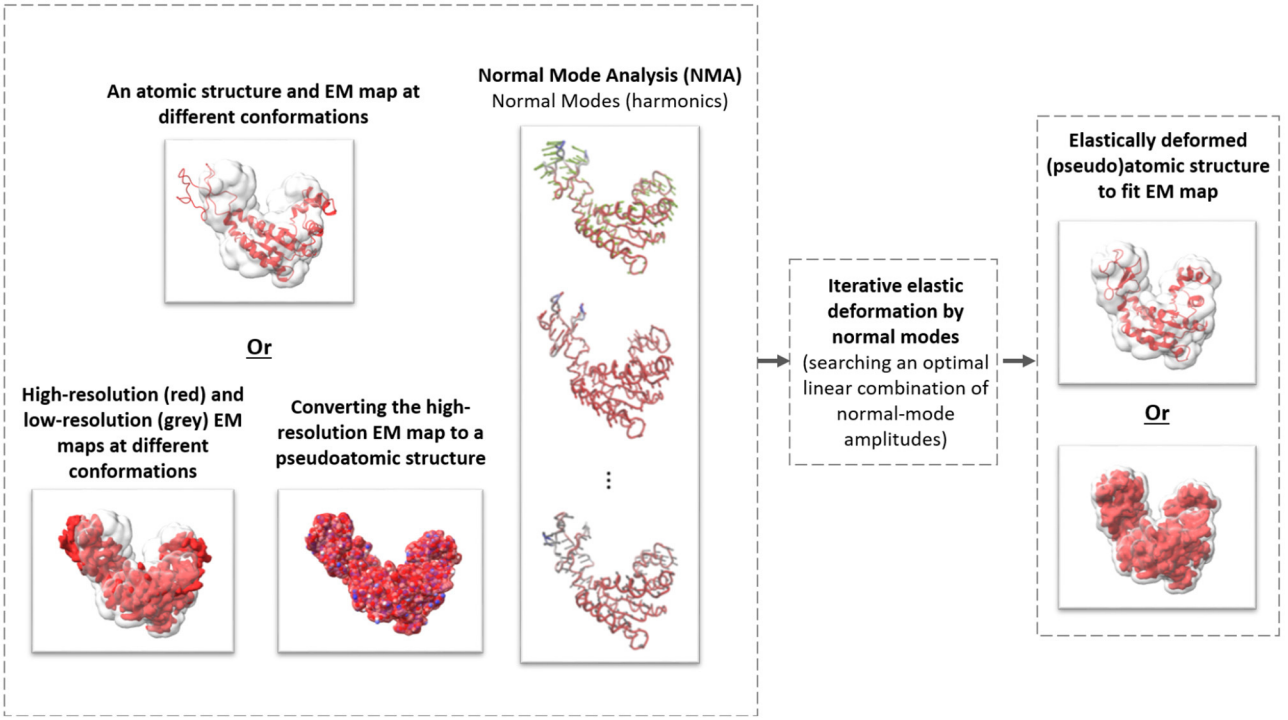


FIGURE 1 | The general scheme of elastic deforming of a reference structure (atomic or pseudoatomic) using normal modes to fit a density map (e.g., an EM map or a subtomogram average).

converted into a collection of Gaussian functions (pseudoatoms) with a carefully selected standard deviation (pseudoatom size, whose default value is 1 voxel Nogales-Cadenas et al., 2013; Jonić and Sorzano, 2016). The pseudoatom size should lead to a structure (called pseudoatomic structure) that, converted back to a density map, approximates the input density map with a small error (given a target approximation error, whose default value is 5% Nogales-Cadenas et al., 2013; Jonić and Sorzano, 2016). Optionally, a mask on the density map can be used prior to the conversion into pseudoatoms (e.g., a spherical binary mask of a given radius) to reduce background noise. Such masks may also be useful if applied on input cryo-ET subtomograms to maximize the chance of having a single molecular complex in each subtomogram (Preprocessing block in the workflow in **Figure 2A**).

2.2. Normal Mode Analysis

This step involves computing normal modes of a reference atomic or pseudoatomic structure, for the 3D-to-3D elastic alignment in the next step. The computation of normal modes is based on the elastic network model (Tirion, 1996; Tama et al., 2002) by representing the interaction between the (pseudo-)atoms as if they are locally connected by elastic springs (within a cutoff distance). Normal Mode Analysis requires the diagonalization of a $3N \times 3N$ matrix of second derivatives of the potential energy (Hessian matrix), where N is the number of nodes in the elastic network model determined by the total number of atoms (or pseudoatoms) in the input reference. In the case of atomic structures, we use the rotation-translation block (RTB) method, which divides the structure into blocks (one or a few consecutive residues per block) whose rotations and translations are considered rather than all degrees of freedom for all atoms (Durand et al., 1994; Tama et al., 2000). Since the RTB method reduces the basis for Hessian diagonalization, it allows fast computing of normal modes. Since pseudoatomic structures usually contain fewer nodes (pseudoatoms) than atomic structures, normal modes can be obtained by a direct diagonalization of the $3N \times 3N$ Hessian, which is referred to as the Cartesian method. As in the case of HEMNMA, we here use the RTB and Cartesian method implementations of Tama et al. (2002) and Suhre and Sanejouand (2004), respectively. Larger values of the interaction cutoff distance (the distance below which atoms or pseudoatoms do not interact) lead to more rigid motions. The atomic interaction cutoff distance may be set manually (by default 8 Å) and the pseudoatomic cutoff distance is recommended to be computed automatically based on the distribution of the pseudoatomic pairwise distances (e.g., as the value below which is a given percentage of all distances as in Nogales-Cadenas et al., 2013; Jin et al., 2014; Jonić and Sorzano, 2016; Harastani et al., 2020). The modes are computed along with their respective collectivity degrees, which count the number of atoms or pseudoatoms affected by the mode as in Brüschweiler (1995). To allow faster data analysis and avoid noise overfitting in the 3D-to-3D elastic alignment in the next step, we select a subset of normal modes (usually, less than 10) with lowest frequencies and highest collectivities, as previously described (Jin

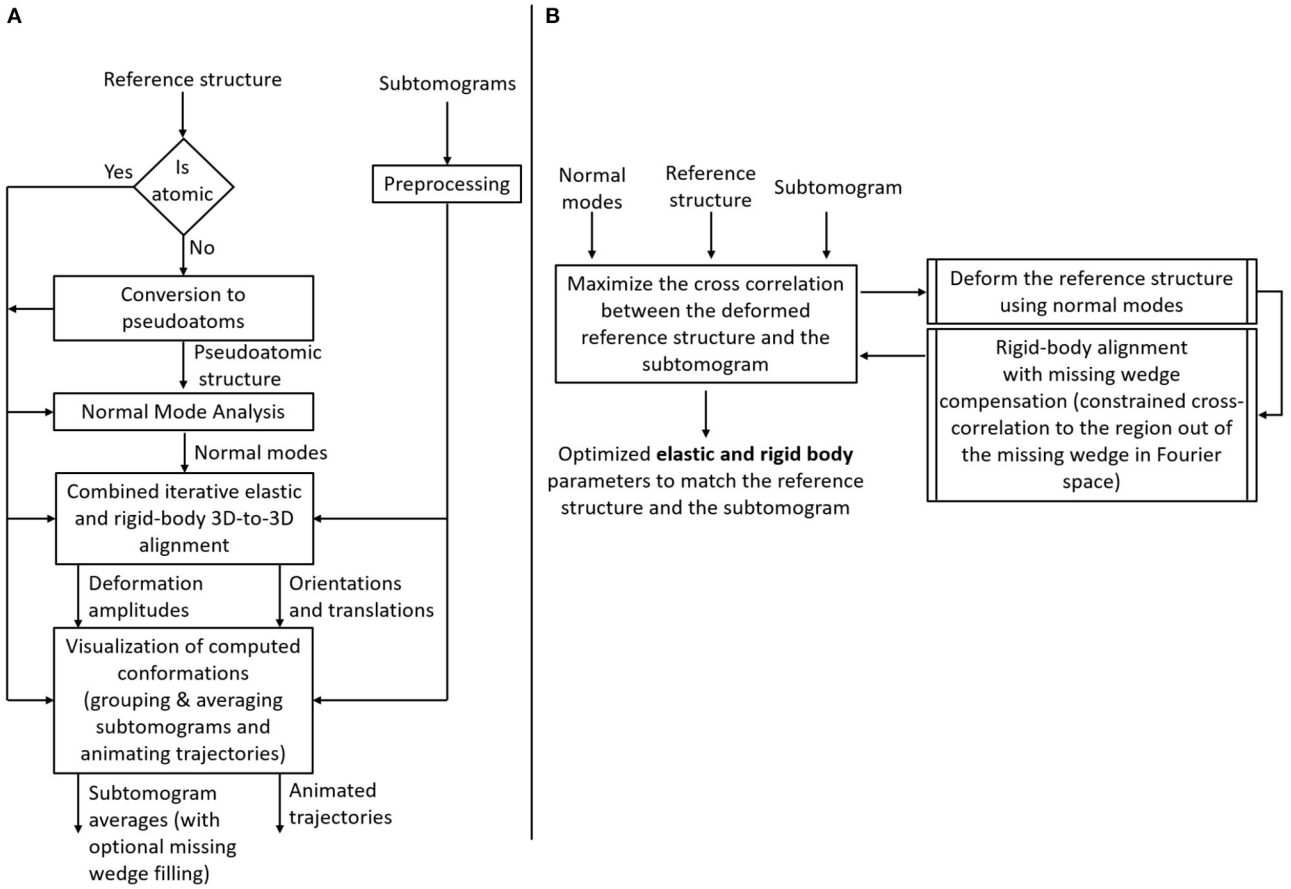


FIGURE 2 | Flowchart of HEMNMA-3D. **(A)** Workflow. **(B)** Combined iterative elastic and rigid-body 3D-to-3D alignment step (the core module of HEMNMA-3D).

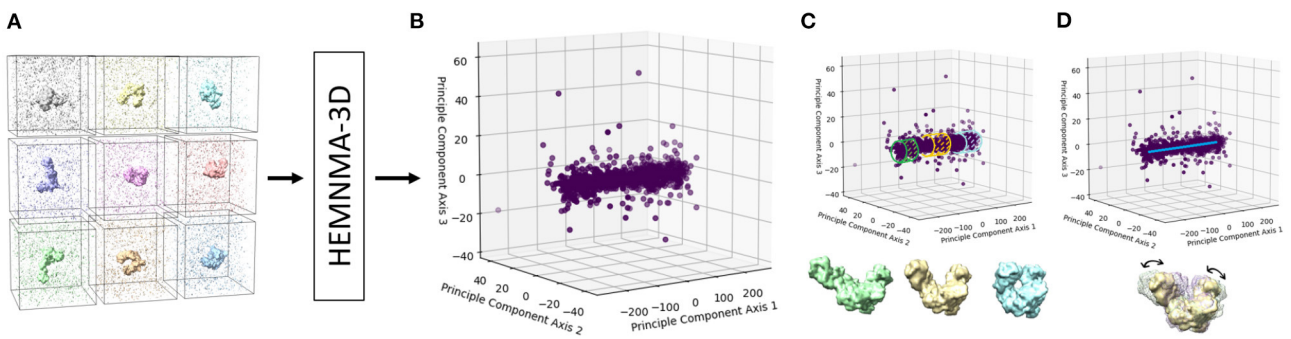


FIGURE 3 | A graphical summary of the dataflow of HEMNMA-3D. **(A)** Input subtomograms containing the same biomolecule but at different orientations, positions and conformations (here represented with a low level of noise for illustration). **(B)** Input subtomograms projected onto a low-dimensional "space of conformations," describing and visualizing the biomolecular conformational variability contained in the subtomograms. **(C)** Grouping of close points (subtomograms with similar biomolecular conformations) and averaging of subtomograms in these groups. **(D)** Animating biomolecular motion along trajectories identified in the densest regions.

et al., 2014; Sorzano et al., 2014; Harastani et al., 2020). Low-frequency high-collectivity normal modes have been shown to be relevant to functional conformational changes (Tama and Sanejouand, 2001; Delarue and Dumas, 2004; Wang et al., 2004; Ma, 2005; Suhre et al., 2006; Tama and Brooks, 2006). The first six (lowest-frequency) normal modes are related to rigid-body transformations and are thus not used for the 3D-to-3D elastic alignment in the next step. The rigid-body 3D-to-3D alignment is done without using these rigid-body normal modes, as explained in the next paragraph. Additionally, a prior knowledge about the

conformational transitions of the complex under study can be used to select the normal modes for the use in the next step. The HEMNMA-3D graphical interface helps the user decide which normal modes to select. The reader is referred to Ma (2005) and Tama and Brooks (2006) for reviews on the usefulness and limitations of NMA.

2.3. Combined Iterative Elastic and Rigid-Body 3D-to-3D Alignment

This step, represented in **Figure 2B**, is the backbone of the proposed method. It has been inspired by the combined iterative elastic and rigid-body 3D-to-3D alignment step of StructMap method (Sorzano et al., 2016), which was proposed for pairwise similarity analysis of SPA high-resolution EM maps (no missing wedge). In HEMNMA-3D proposed here, this step comprises simultaneous NMA-based elastic alignment (search for amplitudes of a linear combination of normal modes) and rigid-body alignment (search for orientation and position, meaning three Euler angles and x, y, and z shifts) of the reference structure with each given subtomogram. It refines the amplitudes of displacement along each used normal mode (elastic parameters) as well as the angles and shifts (rigid-body parameters) of the reference structure until the best match is obtained between this reference structure and the given subtomogram. The latter is achieved by maximizing the similarity between the subtomogram and the density volume from the elastically deformed, oriented and shifted reference, and includes missing wedge compensation. The missing wedge compensation is done by calculating the cross-correlation between the reference and subtomogram density maps only in the region of the Fourier space where the data can be trusted, i.e., by constraining the cross-correlation evaluation to the Fourier space region that excludes the missing wedge region (the region outside of the one specified by the tilt angle range, e.g., −60 to +60°). To maximize this constrained cross-correlation (CCC), we use a variant of Powell's UOBYQA method, which subjects the objective function to a trust-region radius (Berghen and Bersini, 2005). To control the elastic deformation with highly noisy data, the radius of the trust region is adjusted iteratively. The scaling factor of the initial trust-region radius is a parameter that controls the normal-mode amplitude search range and can be modified by the user. It should have a positive value and its default value of 1 produces good results in general. It may be increased (typically to a value between 1 and 2) or decreased (e.g., between 0.5 and 0.9), if expecting larger or smaller conformational changes, respectively. For each subtomogram, the normal mode amplitudes are initiated with zeros, meaning that the non-deformed reference is used in the first iteration. As the iterations evolve, the reference model is displaced with the new guesses of the normal mode displacement amplitudes, converted into a volume and rigid-body aligned with the subtomogram using the method of fast rotational matching. Fast rotational matching has been largely used for rigid-body fitting of atomic models to high-SNR consensus EM maps (Kovacs and Wriggers, 2002; Kovacs et al., 2003). It has been extended to alignment of noisy subtomograms in Chen et al. (2013) and this implementation is used in our work. At the end of each iteration, the CCC is found and fed to the numerical optimizer (Berghen and Bersini, 2005). The iterations repeat until the final value of the trust-region radius or the maximum number of iterations is reached.

2.4. Visualizing and Utilizing the Space of Conformations

The number of elastic alignment parameters (normal mode amplitudes) is determined by the number of selected normal modes for the 3D-to-3D elastic alignment. The ensemble of normal mode amplitudes (for all subtomograms) can be projected onto a lower-dimensional space, so-called conformational space, using a dimensionality reduction technique. Here, we use linear Principal Component Analysis (PCA) as it is the most widely known and intuitively clear dimensionality reduction method, but other dimension reduction methods could also be used (linear or nonlinear). The dimensionality reduction is usually performed to two or three dimensions, which allows a global data display and easier modeling of conformational changes. Each point in the conformational space represents a subtomogram and close points correspond to similar conformations in the subtomograms. The points that differ significantly from the remaining observations (too isolated, outlier points) may be excluded from the further analysis, by excluding the points below a certain p-value based on the Mahalanobis distance (the distance between each point and the whole distribution) (Mahalanobis, 1936). The excluded points can be explained by the fact that some orientations of the molecule combined with the missing wedge artifacts and the high noise make some volumes more difficult to align with the elastically deformed reference. After excluding such outlier points, the space of conformations can be analyzed to reveal molecular dynamics. This can be done by averaging subtomograms of similar conformations in the densest regions of the conformational space or by exploring the densest regions by fitting curves (approximation by line segments) through the data and displacing the reference structure along these curves (referred to as trajectories) to animate the motion along them. The 1-CCC color bar of the conformational space shows coloring the points according to the value of the CCC between the subtomograms and the density maps from the elastically deformed reference model. Subsequently, the colors provide the level of confidence in the obtained conformations. Those subtomograms in which we have less confidence (subtomograms with lower CCC values) than in the "consensus" observations (subtomograms with higher CCC values) can be eliminated from the group averages.

2.4.1. Averaging Subtomograms of Similar Conformations

Close points in the conformational space can be grouped, which results in grouping subtomograms of similar conformations and averaging them. Before computing group averages, the rigid-body alignment parameters found during the combined iterative elastic and rigid-body alignment are applied on the subtomograms. Optionally, before computing group averages, the missing-wedge Fourier space region of individual

subtomograms may be filled in with the corresponding region of the global average computed from all subtomograms. A similar procedure of missing wedge filling of individual subtomograms is used in EMAN2 software package (Galaz-Montoya et al., 2015). The subtomogram averages obtained from the selected groups of subtomograms can be overlapped and compared to understand the conformational changes of the complex in the given set of subtomograms.

2.4.2. Animating Motions (Trajectories)

Distinct trajectories can be determined through the data in the conformational space, and animated to see the motion of the biomolecule while it is displaced along the trajectory. To animate a trajectory, several points (e.g., 10) along the trajectory should be mapped back to the original displacement space (e.g., using inverse PCA), resulting in elastic alignment parameters that can be used to deform the reference atomic or pseudoatomic structure. Concatenating and displaying the resulting structures can show a movie-like animation of the reference biomolecule traveling across the specified trajectory.

2.5. Software Implementation and Technical Details

The software of HEMNMA-3D method proposed here is freely available (open-source). It is a part of ContinuousFlex plugin for the open-source software Scipion3 (De la Rosa-Trevín et al., 2016). ContinuousFlex was introduced in Harastani et al. (2020) and also contains HEMNMA software. The software provides a graphical user interface (GUI) and is empowered with a C++ backend with a message passing interface (MPI) parallelization scheme to efficiently analyze large datasets (simultaneous analysis of N subtomograms using N computing threads). We tested the software on our local workstations and on supercomputer centers. On our local workstations (2.2 GHz Intel Xeon Silver 4214 CPU processors), the current implementation takes around 10 and 30 min to analyze a subtomogram of size 64^3 voxels with three and six normal modes, respectively. These times are reported for the two types of complexes used in this article, together with the number of normal modes used in these two cases. They are the average times required for all iterations of analyzing one subtomogram using a single computing thread while the different subtomograms are analyzed in parallel using different computing threads (if the time is measured for the entire dataset, it will vary with the size of the dataset). It should be noted that the software allows the use of any number of MPI threads and any number of normal modes. However, the more modes are used, the slower the processing. Finally, it should be noted that there is no constraint regarding the size of the dataset (the number of subtomograms) or the size of the individual subtomograms (the number of voxels) that can be analyzed with our software.

3. RESULTS AND DISCUSSION

In this section, we present and discuss the results of HEMNMA-3D with synthetic and experimental subtomograms.

3.1. Synthesizing Datasets for Testing the Method Performance

For testing HEMNMA-3D in general, and the combined elastic and rigid-body 3D-to-3D alignment module in particular (which is the core module of the proposed method), we synthesized two datasets of conformationally heterogeneous subtomograms that mimic discrete and continuous conformational variability, called "Discrete" and "Continuous" datasets, respectively. The flowchart for the data generation procedure is shown in **Figure 4** and is detailed in the following.

The "Discrete" dataset comprises 900 synthetic subtomograms representing three different (synthetic) conformations of the atomic PDB:4AKE structure (Müller et al., 1996) of adenylate kinase chain A (1656 atoms), i.e., 300 subtomograms per conformation. We generated this dataset using the atomic PDB:4AKE structure and its first two non-rigid-body normal modes, i.e., modes 7 and 8. Precisely, the three conformations are represented by the following amplitudes of modes 7 and 8: (mode 7, mode 8) $\in \{(-150, 0), (+150, 0), (0, +150)\}$.

The "Continuous" dataset comprises 1,000 synthetic subtomograms representing a continuum of conformations of the same PDB:4AKE structure. We generated this dataset using this atomic structure and its modes 7 and 8 using a linear relationship between the amplitudes of the two modes. More precisely, the synthesized amplitudes of modes 7 and 8 were identical and randomly distributed in the range $[-200, +200]$ (uniform distribution).

Normal-mode amplitudes do not have a physical unit. Nonetheless, the Root Mean Square Deviation (RMSD) (Kufareva and Abagyan, 2011) between the reference atomic coordinates and these coordinates displaced using normal-mode amplitudes transforms the normal mode amplitudes in physical units. To provide a basis for further evaluation of the method performance, we found a RMSD of 6.95 Å corresponding to the displacement using the amplitude of 200 for each of the two combined modes 7 and 8 (this represents one half of the full range of the synthesized motion).

To generate a subtomogram, first, we deform the atomic structure using appropriate amplitudes for the selected normal modes depending on the dataset in hand, i.e., we use (mode 7, mode 8) $=$ (+150, 0) or (−150, 0) or (0, +150) to create a subtomogram in the "Discrete" dataset, while we assign a random value in the range $[-200, 200]$ for both mode 7 and mode 8 to generate a subtomogram in the "Continuous" dataset. Then, we convert the deformed structure to a volume of size 64^3 voxels and the voxel size of 2.2 $Å^3$ (Peng et al., 1996). Afterwards, we rotate and shift this volume in 3D space using random Euler angles (each of the three Euler angles was randomized in the range $[0, 360^\circ]$) and random shifts (the shift along each of the x, y, and z axes was randomized in the range $[-5, +5]$ voxels), and we project the rotated and shifted volume using tilt values -60 to $+60^\circ$ to obtain a tilt series. We simulate microscope conditions by adding heavy noise (signal to noise ratio SNR $= 0.01$) and modulating the images with a contrast transfer function (CTF) of defocus -1 μm, so that one part of the noise is affected by the CTF and the other is not (Sorzano et al., 2007; Chen

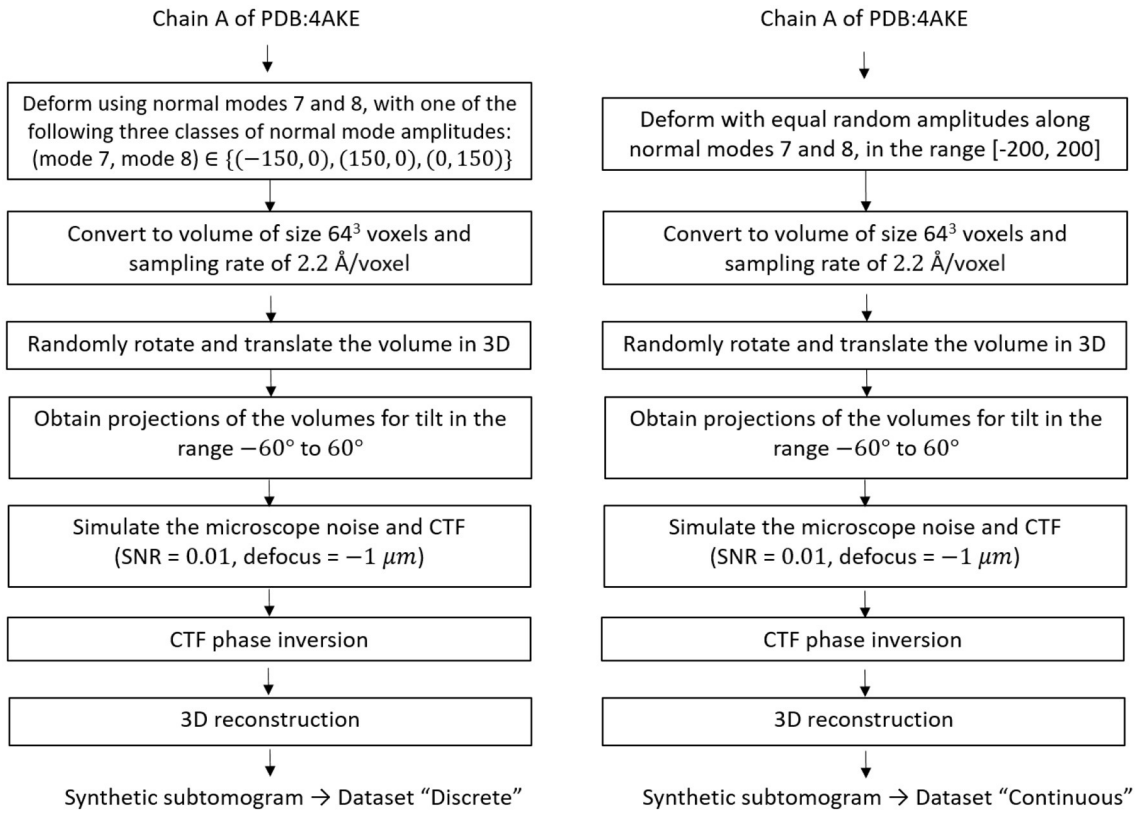


FIGURE 4 | Flowcharts of synthesis of the datasets used for testing and validating HEMNMA-3D, namely "Discrete" dataset (left) and "Continuous" dataset (right).

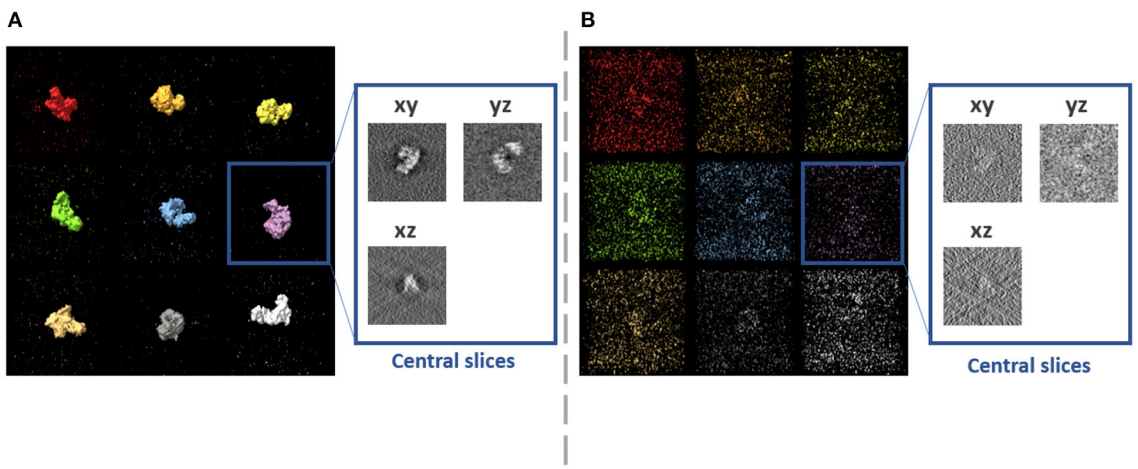


FIGURE 5 | Examples of synthetic subtomograms containing the same molecule but at different orientations, positions and conformations, for two different noise levels. **(A)** Low level of noise (SNR = 0.5). **(B)** High level of noise (SNR = 0.01).

et al., 2013). Finally, we reconstruct a volume (our synthetic subtomogram) from the tilt series using a Fourier reconstruction method (Sorzano et al., 2013). A few examples of the synthesized subtomograms (SNR = 0.01) and their less noisy version (SNR = 0.5, for illustration) is presented in **Figure 5**.

3.2. Synthetic Discrete-Type Conformational Variability

In this experiment, our goal is to retrieve the ground-truth amplitudes of normal modes 7 and 8 by the combined elastic and rigid-body alignment (the core module of HEMNMA-3D) of a

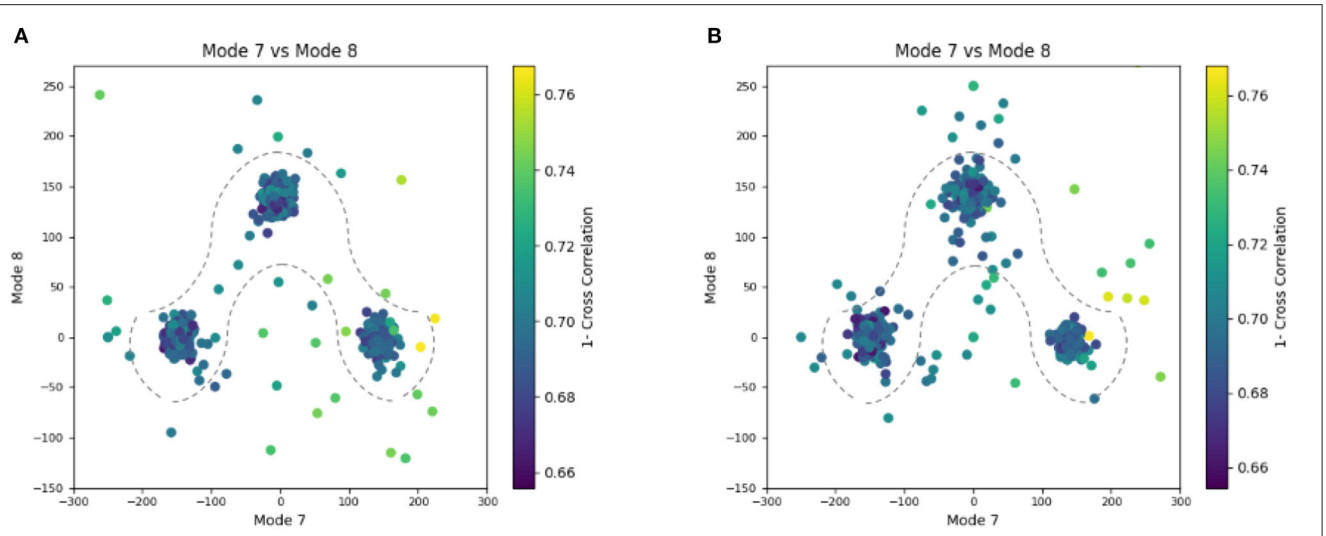


FIGURE 6 | Plots showing the output of the 3D-to-3D elastic and rigid-body alignment module of HEMNMA-3D with "Discrete" dataset (synthetic subtomograms are simulating discrete conformational heterogeneity). **(A)** Use of the atomic structure (chain A of PDB:4AKE) and its normal modes to estimate the conformational parameters (normal-mode amplitudes) and rigid-body parameters (orientation and shift) of the molecules in the input synthetic subtomograms. **(B)** Use of a pseudoatomic structure (from a simulated density map) and its normal modes to estimate the conformational and rigid-body parameters of the molecules in the input synthetic subtomograms. The goal was the retrieval of the ground-truth relationship between the amplitudes along normal modes 7 and 8; ideally, all data should lay in one of the following three clusters of normal-mode amplitudes: (mode 7, mode 8) ∈ {(-150, 0), (150, 0), (0, 150)}; each point in the plot represents a subtomogram and close points represent similar conformations. Note that the dashed curves enclose the data points where p-value > 0.01 in **Table 1**. See the text for more details on this experiment.

TABLE 1 | Mean absolute error and standard deviation between the estimated and ground-truth normal-mode amplitudes along with the angular and shift distances obtained with HEMNMA-3D and "Discrete" synthetic dataset, using an atomic structure (Atomic) and simulated EM map (Volume) as input references.

Experiment		Mode 7		Mode 8		Mode 9		Angular (deg)		Shifting (vox)		*p*-Value	Samples
Ref	Dataset	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std		
Atomic	"Discrete"	16.51	11.87	10.91	7.64	10.70	6.66	1.33	0.77	0.19	0.09	$p > 0.01$	871/900
Volume	"Discrete"	17.70	13.26	11.90	10.13	12.29	8.03	1.33	0.81	0.21	0.10	$p > 0.01$	870/900

The data points below the p*-value of 0.01 were excluded from the error evaluation based on the Mahalanobis distance measure (few data points differing significantly from the remaining observations, which would not be selected in real-case experiments as being too isolated and far from other points). The number of points used for the error computation is shown in the last column of the table (column Samples) and the region with the kept points (*p*-value > 0.01) is shown in* ***Figure 6****.*

reference model with the subtomograms in the "Discrete" dataset. In other words, the goal is to find a solution for the challenging inverse problem of finding the conformation of the structure in each subtomogram. Since the proposed method can use two choices for the reference model, namely, an atomic structure and a density map (e.g., an EM map or a subtomogram average), we performed two types of tests. In the first test type, the atomic structure used to generate the synthetic subtomograms (chain A of the PDB:4AKE) was used as a reference for retrieving normal mode amplitudes of the synthetic subtomograms. In the second test type, we converted (Peng et al., 1996) the atomic structure into a density map (volume) of size 128^3 voxels and voxel size of 1 $Å^3$, and we used this density map as a reference for retrieving normal mode amplitudes of the synthetic subtomograms. In the case of the reference density map, normal modes were computed from the corresponding structure obtained by converting the density map into pseudoatoms (1675 pseudoatoms for the given pseudoatom radius of 1.25 voxels and the target approximation error of 5%). In both cases (reference atomic structure and reference pseudoatomic structure, with their corresponding normal modes), we used three modes (modes 7, 8, and 9) instead of only two modes (modes 7 and 8 that were used to generate synthetic subtomograms), to make the 3D-to-3D elastic and rigid-body alignment task even more challenging. **Figure 6** presents the estimated amplitudes of normal modes 7 and 8 (the estimated amplitude of normal mode 9 is close to 0 and is therefore not shown graphically). **Table 1** presents the mean absolute error and the standard deviation between the estimated and ground-truth normal-mode amplitudes along with the angular and shift distances. In both test cases, the three distinct synthetic groups of subtomograms are correctly separated, taking into account the extreme noise level. The results show a less accurate alignment in the second case, which is expected since, in that case, the atomic structure was used to

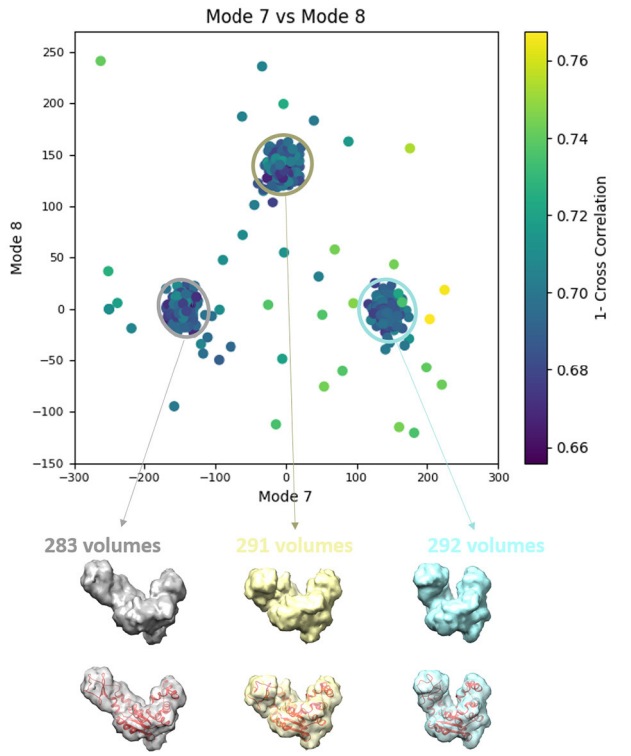


FIGURE 7 | Averages of the three groups (enclosed by ellipses) of subtomograms identified from the output of the 3D-to-3D elastic and rigid-body alignment module of HEMNMA-3D with "Discrete" dataset (shown in **Figure 6A**), using the atomic structure (chain A of PDB:4AKE) and its normal modes to estimate the conformational parameters (normal-mode amplitudes) and rigid-body parameters (orientation and shift) of the molecules in the input synthetic subtomograms. Subtomograms are represented by points and close points represent similar conformations. The numbers of volumes written above the shown subtomogram averages are the numbers of synthetic subtomograms used for computing these subtomogram averages (the numbers of points enclosed by the corresponding ellipses). On the bottom, the subtomogram averages are shown at 50% transparency along with the corresponding ground-truth deformed atomic structure (in red).

generate the dataset and the pseudoatomic structure was used as the reference model for the method to estimate the normal-mode amplitudes from this generated dataset. This is in contrast to the first test case where the same atomic structure was used to create the dataset and as the reference for the method to estimate the normal-mode amplitudes from this dataset. We found a RMSD of 0.55 and 0.60 Å corresponding to a combined displacement along modes 7, 8, and 9 with the mean absolute errors in **Table 1** for the tests with atomic and pseudoatomic structures, respectively. Similarly, we found a RMSD of 0.94 and 1.06 Å corresponding a combined displacement along modes 7, 8, and 9 with the sum of the mean and standard deviation of the absolute errors in **Table 1** for the tests with atomic and pseudoatomic structures, respectively. Hence, the error range is significantly inferior to the half range of the synthesized motion (6.95 Å) and the pixel size used to create the data (2.2 Å). **Figure 7** shows grouping and averaging the subtomograms in the first test type (atomic reference). We compared the obtained subtomogram averages with the corresponding ground-truth volumes (density maps from ground-truth deformed models, without noise and missing wedge, used for synthesizing noisy and CTF-affected tilt-series from which subtomograms were obtained by 3D reconstruction). The visual comparison shows no significant difference between them. The resolutions calculated using Fourier shell correlation—FSC (threshold value of 0.143), after applying onto the subtomogram average a large spherical mask (radius of 28 voxels) with smooth edges (Gaussian smoothing with Gaussian standard deviation of five voxels), are 5.30, 5.35, and 5.74 Å for the three subtomogram averages shown from left to right in **Figure 7**, respectively. Without masking subtomogram averages, these resolutions are 6.31, 6.10, and 6.43 Å, respectively. As a basis for comparison, we provide the resolutions of three individual subtomograms arbitrarily chosen from the three corresponding subtomogram averaging groups. The resolutions of individual subtomograms with masking (the mask already described) are 10.58, 10.57, and 12.75 Å. The resolutions of the same subtomograms without masking are 12.67, 13.55, and 14.72 Å. These results show that a twice better resolution is obtained after averaging only about 300 individual subtomograms per group (**Figure 7**).

3.3. Synthetic Continuous-Type Conformational Variability

Similarly to the previous experiment, our goal in this experiment is to find a solution for the inverse problem of finding the conformation of the structure in each subtomogram using the combined elastic and rigid-body alignment of a reference model with the subtomograms in the "Continuous" dataset. We used the same two reference models as in the previous experiment to estimate the normal-mode amplitudes: an atomic structure (chain A of PDB:4AKE) and a density map from this atomic structure. Also, as in the previous experiment, we used three modes for both tests (atomic or pseudoatomic modes 7, 8, and 9). **Figure 8** presents the estimated amplitudes of modes 7 and 8 (the estimated amplitude of mode 9 is close to 0 and is not shown in the plots). **Table 2** shows the mean absolute error and the standard deviation between the estimated and ground-truth normal-mode amplitudes along with the angular and shift distances. In both test cases, a linear relationship between the estimated amplitudes of normal modes 7 and 8 is clearly distinguishable, which is close to the identity relationship between the ground-truth amplitudes taking into account strong noise present in the data. As in the previous experiment, the results show a slightly less accurate alignment in the second test type (pseudoatomic reference) for the same aforementioned reason. We found a RMSD of 0.66 and 0.75 Å corresponding to a combined displacement along modes 7, 8, and 9 with the mean absolute errors in **Table 2** for the tests with atomic and pseudoatomic structures, respectively. Similarly, we found a RMSD of 1.09 and 1.21 Å corresponding a combined displacement along modes 7, 8, and 9 with the sum of the mean and standard deviation of the absolute errors in **Table 2** for the tests with atomic and pseudoatomic structures, respectively.

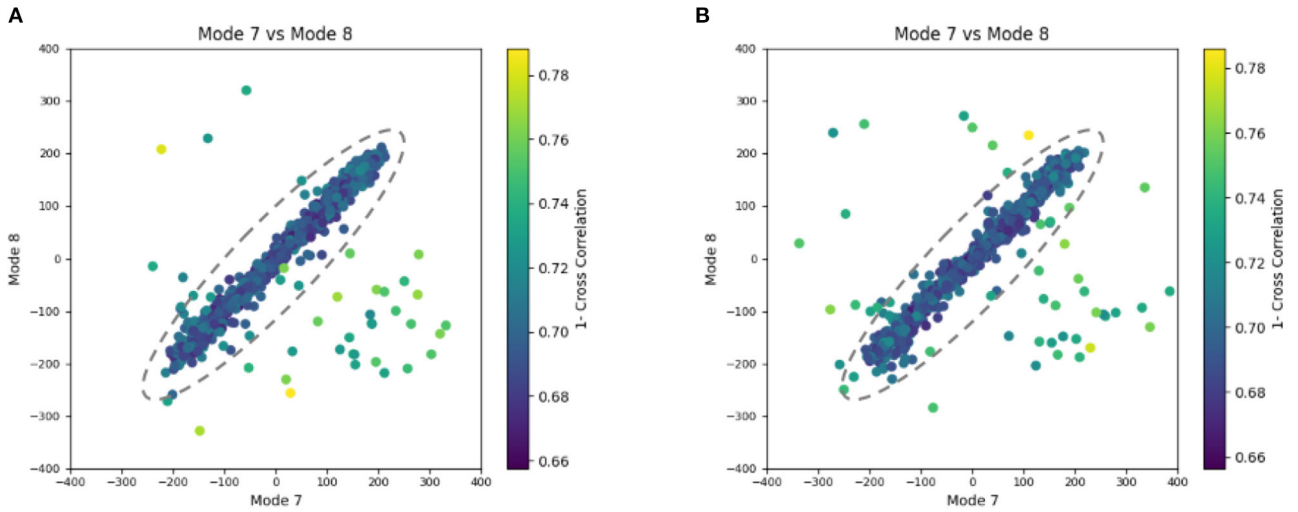


FIGURE 8 | Plots showing the output of the 3D-to-3D elastic and rigid-body alignment module of HEMNMA-3D with "Continuous" dataset (synthetic subtomograms are simulating continuous conformational heterogeneity). **(A)** Use of the atomic structure (chain A of PDB:4AKE) and its normal modes to estimate the conformational parameters (normal-mode amplitudes) and rigid-body parameters (orientation and shift) of the molecules in the input synthetic subtomograms. **(B)** Use of a pseudoatomic structure (from a simulated density map) and its normal modes to estimate the conformational and rigid-body parameters of the molecules in the input synthetic subtomograms. The goal was the retrieval of the ground-truth relationship between the amplitudes along normal modes 7 and 8 (ideally linear relationship, with equal amplitudes of normal modes 7 and 8); each point in the plot represents a subtomogram and close points represent similar conformations. Note that the dashed ellipses enclose the data points where *p*-value > 0.001 in **Table 2**. See the text for more details on this experiment.

TABLE 2 | Mean absolute error and standard deviation between the estimated and ground-truth normal-mode amplitudes along with the angular and shift distances obtained with HEMNMA-3D and "Continuous" synthetic dataset, using an atomic structure (Atomic) and simulated EM map (Volume) as input references.

Experiment		Mode 7		Mode 8		Mode 9		Angular (deg)		Shifting (vox)		*p*-Value	Samples
Ref	Dataset	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std		
Atomic	"Continuous"	20.12	11.30	12.78	11.24	12.74	7.71	1.31	0.79	0.19	0.10	$P > 0.001$	960/1,000
Volume	"Continuous"	21.94	12.59	14.03	9.92	15.68	10.04	1.34	0.80	0.21	0.10	$P > 0.001$	957/1,000

The data points below the p*-value of 0.001 were excluded from the error evaluation based on the Mahalanobis distance measure (few data points differing significantly from the remaining observations, which would not be selected in real-case experiments as being too isolated and far from other points). The number of points used for the error computation is shown in the last column of the table (column Samples) and the region where* p*-value > 0.001 is shown in* ***Figure 8****.*

Hence, the error range is significantly inferior to the half range of the synthesized motion (6.95 Å) and the pixel size used to create the data (2.2 Å). **Figure 9** shows grouping and averaging of subtomograms in this experiment, with eight subtomogram averages calculated along the distribution of the points for the first test type (atomic reference). The subtomogram averages show different conformations of adenylate kinase chain A. Note that the noise contained in the individual subtomograms (SNR = 0.01, **Figure 5B**) was reduced through subtomogram averaging (**Figure 9**). Additional experiments, for other noise levels in input subtomograms, can be found in **Supplementary Figure 1** and **Supplementary Table 1**.

3.4. Experimental Cryo-ET Data: Nucleosomes *in situ*

We applied our method on a dataset comprising 650 *in situ* subtomograms of nucleosomes collected from a cell of a Drosophila embryonic brain, whose conformational variability was detected but not fully explored in a previous work (Eltsov et al., 2018). The subtomograms had the size of 64^3 voxels and the voxel size of 4.4 $Å^3$. A density map obtained with classical subtomogram averaging (without taking into account conformational heterogeneity) was used as the reference density map for HEMNMA-3D (**Figure 10C**). The resolution of this reference density map is around 2 nm (as determined by Fourier Shell Correlation between the reference density map and the density map from the atomic nucleosome structure PDB:3w98 Iwasaki et al., 2013 shown in **Figure 10B**). For more information on how this reference density map (global initial subtomogram average) was obtained, please see section 1 of the **Supplementary Material** (Nucleosome data preparation and acquisition). This reference density map was converted into pseudoatoms (1368 pseudoatoms for the pseudoatom radius of 0.5 voxels and the target approximation error of 5%) and normal mode analysis of the obtained reference pseudoatomic structure was performed. The combined elastic and rigid-body alignment was performed using the pseudoatomic structure and a set of its six low-frequency high-collectivity normal modes.

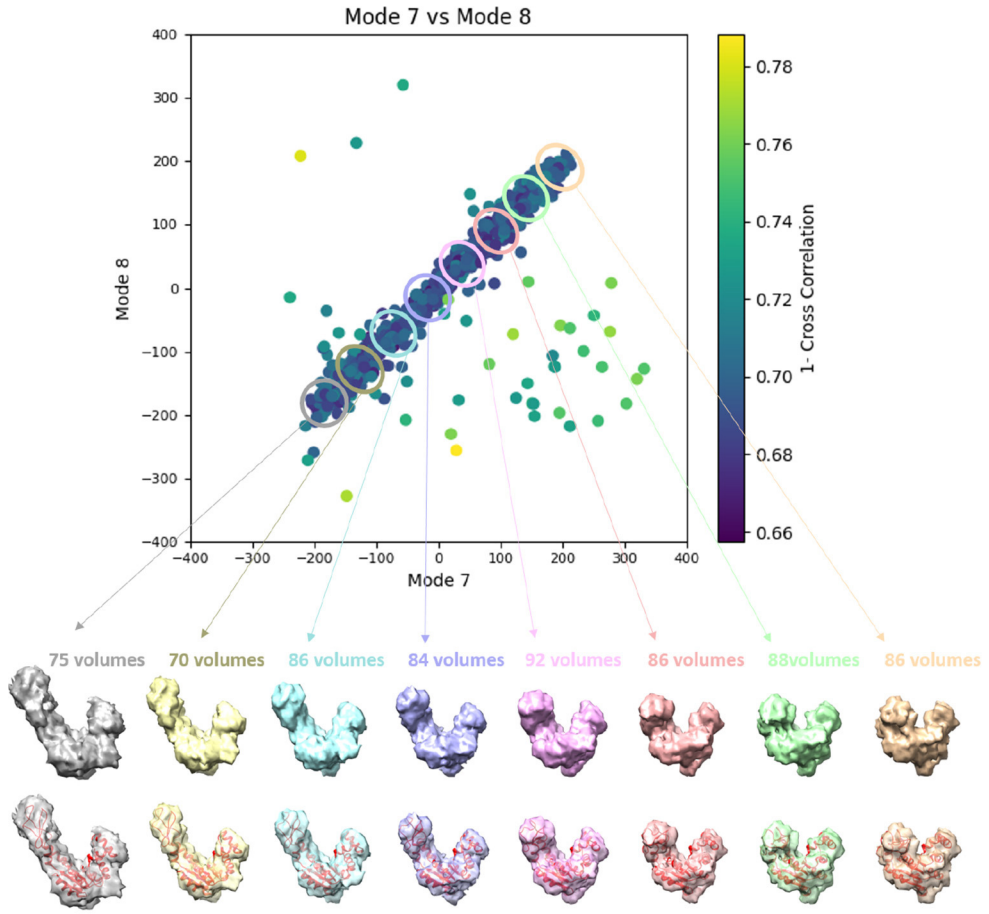


FIGURE 9 | Averages of eight groups (enclosed by ellipses) of subtomograms identified from the output of the 3D-to-3D elastic and rigid-body alignment module of HEMNMA-3D with "Continuous" dataset (shown in **Figure 8A**), using the atomic structure (chain A of PDB:4AKE) and its normal modes to estimate the conformational parameters (normal-mode amplitudes) and rigid-body parameters (orientation and shift) of the molecules in the input synthetic subtomograms. Subtomograms are represented by points and close points represent similar conformations. The numbers of volumes written above the shown subtomogram averages are the numbers of synthetic subtomograms used for computing these subtomogram averages (the numbers of points enclosed by the corresponding ellipses). On the bottom, the subtomogram averages are shown at 50% transparency along with the corresponding theoretical centroid deformed atomic structure (in red).

We selected modes 7–11 and mode 16, as described above and in our previous works (Jin et al., 2014; Sorzano et al., 2014; Harastani et al., 2020). Modes 7–11 were selected as being the five lowest-frequency non-rigid-body modes with collectivities above 0.5. They include the mode related to gaping motion (mode 7) and the mode related to breathing motion (mode 9), which have been described in previous nucleosome studies (Zlatanova et al., 2009; Eltsov et al., 2018). Mode 16 was selected as being related to a motion that could be potentially interesting but it is more complex (a slightly higher frequency motion), potentially including gaping- and breathing-like motions. The normal-mode amplitudes estimated through the alignment (six normal mode amplitudes per subtomogram) were then projected onto a 2D space of conformations using PCA. The space of conformations is presented in **Figure 10**. Recall that each of the points represents a subtomogram, and close points represent similar conformations. By inspecting this conformational space, we identified four densest regions with 70, 183, 74, and 64 points from left to right in **Figure 10D**. Following this analysis, we grouped the subtomograms in each of these four regions and averaged them. Before averaging, we filled in the missing-wedge Fourier space region of the individual subtomograms with the corresponding region of the global average computed from all subtomograms (please note that this global average was computed after aligning subtomograms using the rigid-body alignment parameters found along with the 3D-to-3D elastic alignment by HEMNMA-3D, which is a similar density map to the initial global average map shown in **Figure 10C** as both density maps result from averaging conformational heterogeneous subtomograms). The displacement of the reference pseudoatomic structure (converted into a density map) along two directions D1 and D2 in the space of conformations is shown in **Figure 11** and in **Supplementary Videos 1, 2**. The significant difference between the four group averages (**Figure 10D**) and the reference density map (**Figure 10C**) as well as the motion observed along the two

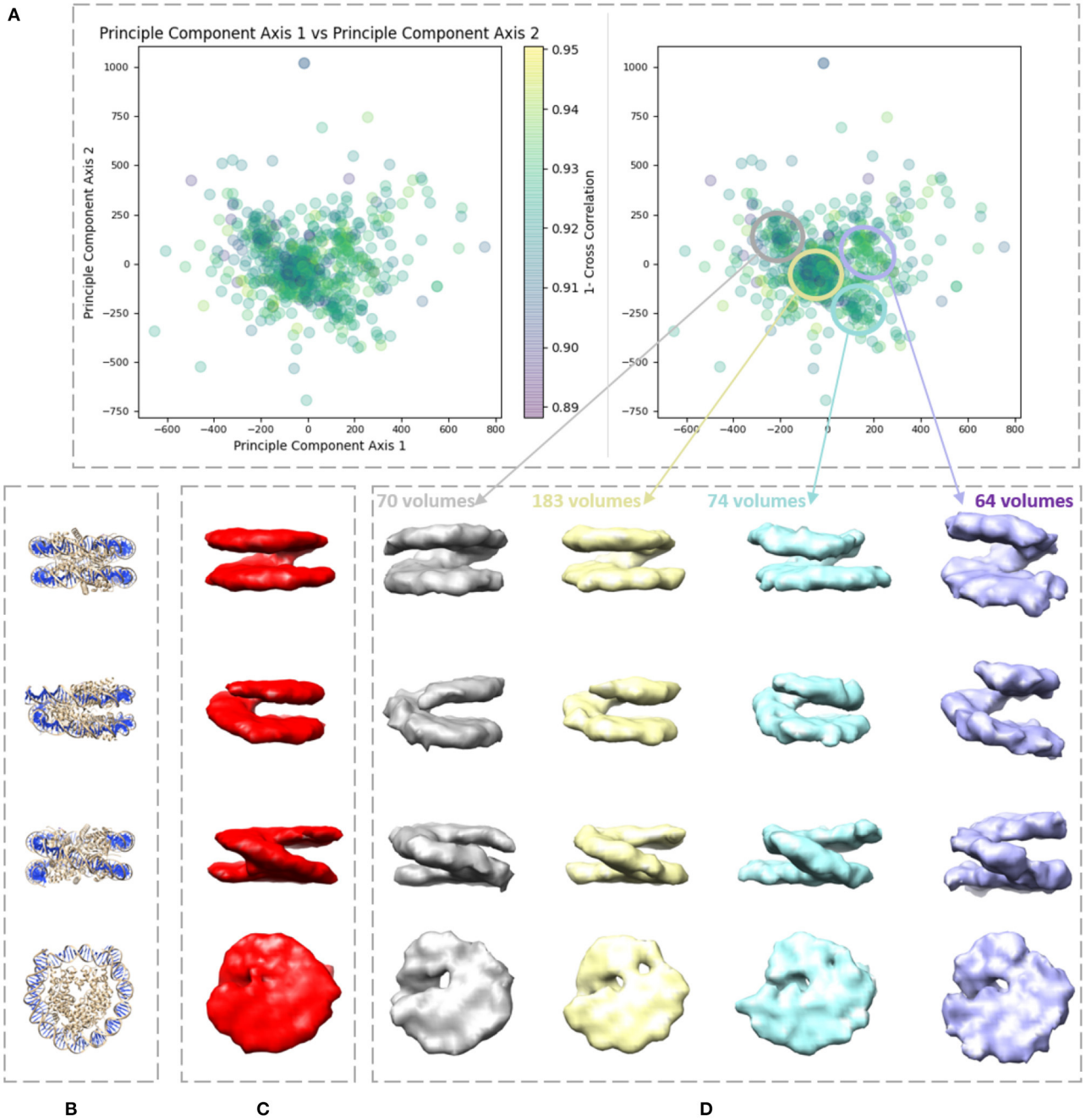


FIGURE 10 | Illustration of HEMNMA-3D use with *in situ* cryo-ET nucleosome dataset. **(A)** Space of conformations resulting from projecting the estimated amplitudes of six normal modes onto a two-dimensional space using PCA. **(B)** Nucleosome atomic structure PDB:3w98, for comparison purposes. **(C)** Nucleosome subtomogram average (around 2 nm resolution) used as the input reference density map for HEMNMA-3D, obtained by classical subtomogram averaging, without taking into account conformational heterogeneity [for more information on how this global initial subtomogram average was obtained, see section 1 of the **Supplementary Material** (Nucleosome data preparation and acquisition)]. **(D)** Four subtomogram averages from four densest regions in the space of conformations (regions encircled with ellipses) showing different nucleosome conformations, mainly, different gap distances between the nucleosome gyres. The numbers of volumes written above the subtomogram averages shown in **(D)** are the numbers of *in situ* cryo-ET subtomograms used for computing these subtomogram averages (the numbers of points enclosed by the corresponding ellipses).

directions D1 and D2 (**Figure 11**, **Supplementary Videos 1, 2**) can be described, mainly in terms of opening the nucleosome by increasing the distance between the two gyres of the DNA superhelix. This result consents the previous findings, observed but not fully explored in a previous study (manual analysis) of the nucleosome conformational variability (Eltsov et al., 2018). The

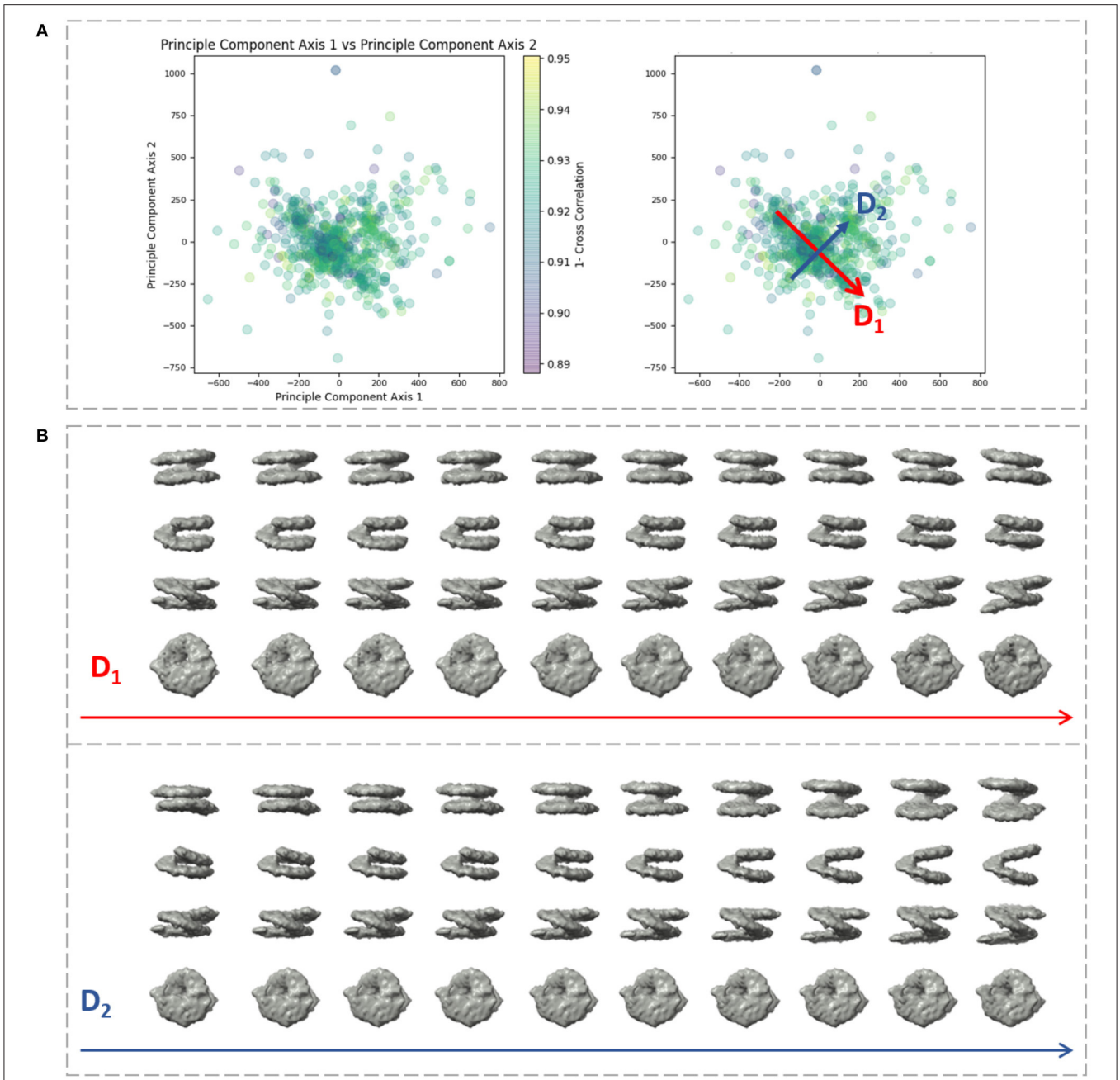


FIGURE 11 | Displacement of the reference density map along two directions D1 and D2 in the space of conformations obtained (**Figure 10**) with HEMNMA-3D with *in situ* cryo-ET nucleosome dataset. **(A)** Space of conformations (left) as shown in **Figure 10** and two directions D1 and D2 used to displace the reference density map (**Figure 10C**) in this space (right). **(B)** Displacement of the reference density map along the D1 and D2 directions (10 frames of the corresponding trajectory are shown row-wise).

group averages are also compared with the atomic nucleosome structure PDB:3w98 in **Supplementary Figure 2**.

4. DISCUSSION AND CONCLUSIONS

This article presents HEMNMA-3D, the first cryo-ET subtomogram data analysis approach to study continuous conformational variability of biomolecular complexes, which maps a set of subtomograms into a space of conformations using a reference model and its normal modes. The conformational space permits (i) grouping (and averaging) subtomograms with similar conformations and revealing hidden conformations and (ii) recording animated displacements of the reference model along the densest regions of the space, along trajectories identified by curve fitting of the

data in these regions. These HEMNMA-3D outputs could be valuable to cryo-ET studies of molecular mechanisms involved in conformational changes of complexes *in vitro* and *in situ*. HEMNMA-3D is thoroughly tested using synthetic subtomograms and applied to a cryo-ET experimental dataset (nucleosome subtomograms recorded *in situ* in Drosophila interphase nucleus). It provides promising results coherent with previous findings. An open-source software with a graphical user interface is provided for this method with a C++ backend and a Message Passing Interface parallelization scheme.

Both HEMNMA-3D and NMA-based flexible fitting are NMA applications concerned with estimating the molecular conformation in density maps based on an atomic or pseudoatomic reference. However, the purpose of HEMNMA-3D is different from that of the classical NMA-based fitting methods. Classical NMA-based fitting methods aim at determining an atomic representation of an EM density map, which is done by flexible fitting of a given atomic structure into that EM map. The purpose of HEMNMA-3D is to get a low-dimensional representation of the heterogeneity of a given set of EM maps, such as subtomograms. Such low-dimensional representations do not require pushing the limits of the fitting accuracy as in the case of classical flexible fitting of atomic structures into EM maps, which also prevents overfitting. Besides, HEMNMA-3D performs a rigid-body alignment simultaneously with the flexible alignment, which accounts for the missing wedge of the low-SNR subtomograms, whereas classical NMA-based fitting methods typically use high-SNR average consensus EM maps reconstructed from single particle cryo-EM images without missing wedge.

The uniform random distribution of conformations was used in our experiments with synthetic data to show that HEMNMA-3D finds the correct values (within an acceptable error) for any conformation, rotation and translation, and that it does not yield wrong biased solutions (e.g., systematic alignment errors, such as a wrong biased alignment to one or the other conformation and systematic rotational or translational errors). Taking into account the independent analysis of each individual subtomogram, HEMNMA-3D should be able to recover any other conformational distribution, with similar errors to those obtained with the conformational distributions used in this article.

The tests with synthetic data using a pseudoatomic structure from a simulated density map as a reference were used to demonstrate the ability of the method to retrieve the ground truth conformations with a comparable accuracy to the accuracy achieved using an atomic reference despite that (i) the pseudoatomic reference was not used to synthesize the data (the datasets were synthesized using the atomic reference), (ii) pseudoatomic coordinates unlikely coincide with atomic coordinates (pseudoatomic coordinates are obtained through volume-to-pseudoatoms conversion, which does not use any prior information about atoms), and (iii) the method for calculating normal modes is different for a pseudoatomic reference (Cartesian method) and an atomic reference (RTB method). In experimental cases, one could obtain pseudoatoms from a density map of higher resolution than the data itself, if such density map is already available (e.g., an EMDB map obtained by high-resolution single particle cryo-EM reconstruction) or if it can be simulated from an available atomic structure (as in our synthetic data experiments). However, a preferred choice for the reference density map should be a density map from the data itself, which can be obtained by classical subtomogram averaging (without taking into account conformational heterogeneity), as was the case in the nucleosome experiments shown in this article.

The synthetic subtomogram datasets used in this work do not account for crowded molecular environments, radiation dose accumulation during tilting, and differences in CTF defocus over the tilted planes. Nevertheless, the synthesized subtomograms used here were challenging, as containing a small number of voxels and as being obtained by 3D reconstruction from synthetic tilt images affected by strong noise and CTF, which altogether lowered the resolution of the reconstructed subtomograms. The subtomograms were additionally affected by the missing wedge artifacts. Despite such difficult conditions, HEMNMA-3D finds the correct conformational, orientational, and translational parameters, which suggests that it can be useful in practice, and this usefulness was here demonstrated with experimental nucleosome subtomograms.

In experimental cases, such as the nucleosome study shown in this article, the number of used normal modes will always be smaller than the actual number of normal modes (the entire set of modes is too large to be included in our calculations as this would require too long computing times). Therefore, the conformational landscape will always be an approximation of the actual conformational landscape. Small sets of selected potentially relevant modes have been shown to produce good approximation of the actual conformational landscape of the nucleosome studied here by HEMNMA-3D as well as of other complexes studied by HEMNMA. In some cases, a single normal mode could be enough, such as in the case of 70S ribosomes, where the normal mode describing the rotation between the two subunits 30S and 50S was used to analyze conformational and compositional variability of EF-G bound and unbound 70S ribosome cryo-EM dataset (Jin et al., 2014). HEMNMA-3D uses the same software for NMA and the same numerical optimizer (to estimate the amplitudes of normal modes iteratively) as HEMNMA, and it should thus have similar performance as HEMNMA regarding the determination of the conformational landscape using a smaller number of normal modes. HEMNMA-3D software (ConinuousFlex plugin of Scipion V3.0) helps the user decide which normal modes to select, based on the lowest-frequency highest-collectivity criterion and including or not a prior knowledge about the conformational transitions, as is the case with HEMNMA software (Harastani et al., 2020).

HEMNMA-3D can deal with larger sets of subtomograms than those shown in this article. Each subtomogram is analyzed independently of other subtomograms, meaning on a separate computing thread. The same computing time per subtomogram

is required for larger and smaller datasets of the same molecular complex and the time required to process the entire dataset varies with the size of the dataset.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Files**. HEMNMA-3D software code is publicly available on Github (https://github.com/scipion-em/scipion-em-continuousflex.git) and is also part of the open-source ContinuousFlex plugin of Scipion V3.0. The nucleosome data used in this article have been deposited in EMPIAR and EMDB databases under the accession codes EMPIAR-10679 and EMD-12699, respectively. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

MH and SJ designed the method and the experiments with synthetic data. ME obtained the nucleosome subtomograms and their initial average used as the reference for the method in nucleosome experiments. MH implemented the method and performed all the experiments, under the supervision of SJ, and wrote the first draft of the article manuscript, which was finalized by SJ with input from all authors. ME and AL participated in the results interpretation. All authors designed the experiments with nucleosome data, contributed to the manuscript preparation, read, and approved the final version of the manuscript.

FUNDING

We acknowledge the support of the French National Research Agency—ANR (ANR-20-CE11-0020-01 to AL, ANR-20-CE11-0020-02 to ME, ANR-20-CE11-0020-03 and ANR-19-CE11-0008-01 to SJ); German Research Foundation (DFG EL 861/1 to ME); the French Microscopy Society (2019 Master internship Grant to MH); the Sorbonne University (2019 Interface pour le Vivant Ph.D. scholarship Grant to MH); and the access to the HPC resources of CINES and IDRIS granted by GENCI (2019-A0070710998, AP010712190, AD011012188 to SJ).

ACKNOWLEDGMENTS

MH and SJ thank the Scipion team (CNB-CSIC, Madrid, Spain) for the help with porting the HEMNMA-3D code from Scipion V2.0 to Scipion V3.0, and Ilyes Hamitouche and Rémi Vuillemot (IMPMC-UMR 7590, Paris, France) for constructive discussions on HEMNMA-3D development. ME thanks Diana Grewe for cryo-sectioning, and Achilleas Frangakis (BMLS, Frankfurt, Germany) for providing scripts for subtomogram averaging, and the access to computational resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.663121/full#supplementary-material

REFERENCES

Abeyrathne, P. D., San Koh, C., Grant, T., Grigorieff, N., and Korostelev, A. A. (2016). Ensemble cryo-em uncovers inchworm-like translocation of a viral ires through the ribosome. *Elife* 5:e14874. doi: 10.7554/eLife.14874

Albert, S., Schaffer, M., Beck, F., Mosalaganti, S., Asano, S., Thomas, H. F., et al. (2017). Proteasomes tether to two distinct sites at the nuclear pore complex. *Proc. Natl. Acad. Sci. U.S.A.* 114, 13726–13731. doi: 10.1073/pnas.1716305114

Andén, J., and Singer, A. (2018). Structural variability from noisy tomographic projections. *SIAM J. Imaging Sci.* 11, 1441–1492. doi: 10.1137/17M1153509

Banerjee, S., Bartesaghi, A., Merk, A., Rao, P., Bulfer, S. L., Yan, Y., et al. (2016). 2.3 å resolution cryo-em structure of human p97 and mechanism of allosteric inhibition. *Science* 351, 871–875. doi: 10.1126/science.aad7974

Berghen, F. V., and Bersini, H. (2005). Condor, a new parallel, constrained extension of Powell's Uobyqa algorithm: experimental results and comparison with the dfo algorithm. *J. Comput. Appl. Math.* 181, 157–175. doi: 10.1016/j.cam.2004.11.029

Bharat, T. A., and Scheres, S. H. (2016). Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in relion. *Nat. Protoc.* 11, 2054–2065. doi: 10.1038/nprot.2016.124

Böck, D., Medeiros, J. M., Tsao, H. F., Penz, T., Weiss, G. L., Aistleitner, K., et al. (2017). *In situ* architecture, function, and evolution of a contractile injection system. *Science* 357, 713–717. doi: 10.1126/science.aan7904

Brüschweiler, R. (1995). Collective protein dynamics and nuclear spin relaxation. *J. Chem. Phys.* 102, 3396–3403. doi: 10.1063/1.469213

Bykov, Y. S., Schaffer, M., Dodonova, S. O., Albert, S., Plitzko, J. M., Baumeister, W., et al. (2017). The structure of the copi coat determined within the cell. *Elife* 6:e32493. doi: 10.7554/eLife.32493

Castaño-Díez, D., and Zanetti, G. (2019). In situ structure determination by subtomogram averaging. *Curr. Opin. Struct. Biol.* 58, 68–75. doi: 10.1016/j.sbi.2019.05.011

Chen, Y., Pfeffer, S., Fernández, J. J., Sorzano, C. O. S., and Förster, F. (2014). Autofocused 3D classification of cryoelectron subtomograms. *Structure* 22, 1528–1537. doi: 10.1016/j.str.2014.08.007

Chen, Y., Pfeffer, S., Hrabe, T., Schuller, J. M., and Förster, F. (2013). Fast and accurate reference-free alignment of subtomograms. *J. Struct. Biol.* 182, 235–245. doi: 10.1016/j.jsb.2013.03.002

Dashti, A., Schwander, P., Langlois, R., Fung, R., Li, W., Hosseinizadeh, A., et al. (2014). Trajectories of the ribosome as a brownian nanomachine. *Proc. Natl. Acad. Sci. U.S.A.* 111, 17492–17497. doi: 10.1073/pnas.1419276111

Davies, K. M., Blum, T. B., and Kühlbrandt, W. (2018). Conserved *in situ* arrangement of complex i and iii2 in mitochondrial respiratory chain supercomplexes of mammals, yeast, and plants. *Proc. Natl. Acad. Sci. U.S.A.* 115, 3024–3029. doi: 10.1073/pnas.1720702115

De la Rosa-Trevín, J., Quintana, A., Del Cano, L., Zaldivar, A., Foche, I., Gutierrez, J., et al. (2016). Scipion: a software framework toward integration, reproducibility and validation in 3D electron microscopy. *J. Struct. Biol.* 195, 93–99. doi: 10.1016/j.jsb.2016.04.010

Delarue, M., and Dumas, P. (2004). On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. *Proc. Natl. Acad. Sci. U.S.A.* 101, 6957–6962. doi: 10.1073/pnas.0400301101

Durand, P., Trinquier, G., and Sanejouand, Y. H. (1994). A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers* 34, 759–771. doi: 10.1002/bip.360340608

Elad, N., Clare, D. K., Saibil, H. R., and Orlova, E. V. (2008). Detection and separation of heterogeneity in molecular complexes by statistical

analysis of their two-dimensional projections. *J. Struct. Biol.* 162, 108–120. doi: 10.1016/j.jsb.2007.11.007

Eltsov, M., Grewe, D., Lemercier, N., Frangakis, A., Livolant, F., and Leforestier, A. (2018). Nucleosome conformational variability in solution and in interphase nuclei evidenced by cryo-electron microscopy of vitreous sections. *Nucleic Acids Res.* 46, 9189–9200. doi: 10.1093/nar/gky670

Förster, F., Pruggnaller, S., Seybert, A., and Frangakis, A. S. (2008). Classification of cryo-electron sub-tomograms using constrained correlation. *J. Struct. Biol.* 161, 276–286. doi: 10.1016/j.jsb.2007.07.006

Frank, J., and Ourmazd, A. (2016). Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-em. *Methods* 100, 61–67. doi: 10.1016/j.ymeth.2016.02.007

Fu, J., Gao, H., and Frank, J. (2007). Unsupervised classification of single particles by cluster tracking in multi-dimensional space. *J. Struct. Biol.* 157, 226–239. doi: 10.1016/j.jsb.2006.06.012

Galaz-Montoya, J. G., Flanagan, J., Schmid, M. F., and Ludtke, S. J. (2015). Single particle tomography in EMAN2. *J. Struct. Biol.* 190, 279–290. doi: 10.1016/j.jsb.2015.04.016

Guo, Q., Lehmer, C., Martínez-Sánchez, A., Rudack, T., Beck, F., Hartmann, H., et al. (2018). *In situ* structure of neuronal C9orf72 Poly-GA aggregates reveals proteasome recruitment. *Cell* 172, 696–705. doi: 10.1016/j.cell.2017.12.030

Harastani, M., Sorzano, C. O. S., and Jonić, S. (2020). Hybrid electron microscopy normal mode analysis with scipion. *Prot. Sci.* 29, 223–236. doi: 10.1002/pro.3772

Haselbach, D., Komarov, I., Agafonov, D. E., Hartmuth, K., Graf, B., Dybkov, O., et al. (2018). Structure and conformational dynamics of the human spliceosomal bact complex. *Cell* 172, 454–464.e11. doi: 10.1016/j.cell.2018.01.010

Himes, B. A., and Zhang, P. (2018). emclarity: software for high-resolution cryo-electron tomography and subtomogram averaging. *Nat. Methods* 15, 955–961. doi: 10.1038/s41592-018-0167-z

Hutchings, J., Stancheva, V., Miller, E. A., and Zanetti, G. (2018). Subtomogram averaging of copii assemblies reveals how coat organization dictates membrane shape. *Nat. Commun.* 9:4154. doi: 10.1038/s41467-018-06577-4

Iwasaki, W., Miya, Y., Horikoshi, N., Osakabe, A., Taguchi, H., Tachiwana, H., et al. (2013). Contribution of histone N-terminal tails to the structure and stability of nucleosomes. *FEBS Open Bio* 3, 363–369. doi: 10.1016/j.fob.2013.08.007

Jin, Q., Sorzano, C. O. S., De La Rosa-Trevín, J. M., Bilbao-Castro, J. R., Núñez-Ramírez, R., Llorca, O., et al. (2014). Iterative elastic 3D-to-2D alignment method using normal modes for studying structural dynamics of large macromolecular complexes. *Structure* 22, 496–506. doi: 10.1016/j.str.2014.01.004

Jonić, S. (2017). Computational methods for analyzing conformational variability of macromolecular complexes from cryo-electron microscopy images. *Curr. Opin. Struct. Biol.* 43, 114–121. doi: 10.1016/j.sbi.2016.12.011

Jonić, S., and Sorzano, C. (2016). Coarse-graining of volumes for modeling of structure and dynamics in electron microscopy: algorithm to automatically control accuracy of approximation. *IEEE J. Select. Top. Signal Process.* 10, 161–173. doi: 10.1109/JSTSP.2015.2489186

Kaplan, M., Ghosal, D., Subramanian, P., Oikonomou, C. M., Kjaer, A., Pirbadian, S., et al. (2019). The presence and absence of periplasmic rings in bacterial flagellar motors correlates with stator type. *Elife* 8:e43487. doi: 10.7554/eLife.43487

Katsevich, E., Katsevich, A., and Singer, A. (2015). Covariance matrix estimation for the cryo-em heterogeneity problem. *SIAM J. Imaging Sci.* 8, 126–185. doi: 10.1137/130935434

Kovacs, J. A., Chacón, P., Cong, Y., Metwally, E., and Wriggers, W. (2003). Fast rotational matching of rigid bodies by fast Fourier transform acceleration of five degrees of freedom. *Acta Crystallogr. D* 59, 1371–1376. doi: 10.1107/S0907444903011247

Kovacs, J. A., and Wriggers, W. (2002). Fast rotational matching. *Acta Crystallogr. D Biol. Crystallogr.* 58, 1282–1286. doi: 10.1107/S0907444902009794

Kovtun, O., Leneva, N., Bykov, Y. S., Ariotti, N., Teasdale, R. D., Schaffer, M., et al. (2018). Structure of the membrane-assembled retromer coat determined by cryo-electron tomography. *Nature* 561, 561–564. doi: 10.1038/s41586-018-0526-z

Kufareva, I., and Abagyan, R. (2011). "Methods of protein structure comparison," in *Homology Modeling*, eds A. J. W. Orry and R. Abagyan (New York, NY; Heidelberg; Dordrecht; London: Springer), 231–257. doi: 10.1007/978-1-61779-588-6_10

Leigh, K. E., Navarro, P. P., Scaramuzza, S., Chen, W., Zhang, Y., Castaño-Díez, D., et al. (2019). Subtomogram averaging from cryo-electron tomograms. *Methods Cell Biol.* 152, 217–259. doi: 10.1016/bs.mcb.2019.04.003

Lyumkis, D., Brilot, A. F., Theobald, D. L., and Grigorieff, N. (2013). Likelihood-based classification of cryo-em images using frealign. *J. Struct. Biol.* 183, 377–388. doi: 10.1016/j.jsb.2013.07.005

Ma, J. (2005). Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure* 13, 373–380. doi: 10.1016/j.str.2005.02.002

Mahalanobis, P. C. (1936). *On the Generalized Distance in Statistics*. Kolkata: National Institute of Science of India.

Mahamid, J., Pfeffer, S., Schaffer, M., Villa, E., Danev, R., Cuellar, L. K., et al. (2016). Visualizing the molecular sociology at the hela cell nuclear periphery. *Science* 351, 969–972. doi: 10.1126/science.aad8857

Mattei, S., Glass, B., Hagen, W. J. H., Kräusslich, H. G., and Briggs, J. A. G. (2016). The structure and flexibility of conical HIV-1 capsids determined within intact virions. *Science* 354, 1434–1437. doi: 10.1126/science.aah4972

Moebel, E., and Kervrann, C. (2020). A monte carlo framework for missing wedge restoration and noise removal in cryo-electron tomography. *J. Struct. Biol. X* 4:100013. doi: 10.1016/j.yjsbx.2019.100013

Mosalaganti, S., Kosinski, J., Albert, S., Schaffer, M., Strenkert, D., Salomé, P. A., et al. (2018). *In situ* architecture of the algal nuclear pore complex. *Nat. Commun.* 9:2361. doi: 10.1038/s41467-018-04739-y

Müller, C., Schlauderer, G., Reinstein, J., and Schulz, G. E. (1996). Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* 4, 147–156. doi: 10.1016/S0969-2126(96)00018-4

Nogales-Cadenas, R., Jonic, S., Tama, F., Arteni, A., Tabas-Madrid, D., Vázquez, M., et al. (2013). 3DEM loupe: analysis of macromolecular dynamics using structures from electron microscopy. *Nucleic Acids Res.* 41, W363–W367. doi: 10.1093/nar/gkt385

Park, D., Lara-Tejero, M., Waxham, M. N., Li, W., Hu, B., Galán, J. E., et al. (2018). Visualization of the type III secretion mediated salmonella-host cell interface using cryo-electron tomography. *Elife* 7:e39514. doi: 10.7554/eLife.39514

Penczek, P. A., Frank, J., and Spahn, C. M. (2006). A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation. *J. Struct. Biol.* 154, 184–194. doi: 10.1016/j.jsb.2005.12.013

Penczek, P. A., Kimmel, M., and Spahn, C. M. (2011). Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-em images. *Structure* 19, 1582–1590. doi: 10.1016/j.str.2011.10.003

Peng, L. M., Ren, G., Dudarev, S., and Whelan, M. (1996). Robust parameterization of elastic and absorptive electron atomic scattering factors. *Acta Crystallogr. A Foundat. Crystallogr.* 52, 257–276. doi: 10.1107/S0108767395014371

Pfeffer, S., Dudek, J., Schaffer, M., Ng, B. G., Albert, S., Plitzko, J. M., et al. (2017). Dissecting the molecular organization of the translocon-associated protein complex. *Nat. Commun.* 8:14516. doi: 10.1038/ncomms14516

Rapisarda, C., Cherrak, Y., Kooger, R., Schmidt, V., Pellarin, R., Logger, L., et al. (2019). *In situ* and high-resolution cryo-em structure of a bacterial type VI secretion system membrane complex. *EMBO J.* 38:e100886. doi: 10.15252/embj.2018100886

Riedel, C., Vasishtan, D., Siebert, C. A., Whittle, C., Lehmann, M. J., Mothes, W., et al. (2017). Native structure of a retroviral envelope protein and its conformational change upon interaction with the target cell. *J. Struct. Biol.* 197, 172–180. doi: 10.1016/j.jsb.2016.06.017

Scheres, S. H. (2012). Relion: implementation of a bayesian approach to cryo-em structure determination. *J. Struct. Biol.* 180, 519–530. doi: 10.1016/j.jsb.2012.09.006

Scheres, S. H., Melero, R., Valle, M., and Carazo, J. M. (2009). Averaging of electron subtomograms and random conical tilt reconstructions through likelihood optimization. *Structure* 17, 1563–1572. doi: 10.1016/j.str.2009.10.009

Schur, F. K., Obr, M., Hagen, W. J., Wan, W., Jakobi, A. J., Kirkpatrick, J. M., et al. (2016). An atomic model of HIV-1 Capsid-SP1 reveals structures regulating assembly and maturation. *Science* 353, 506–508. doi: 10.1126/science.aaf9620

Sorzano, C., Jonic, S., Núñez-Ramírez, R., Boisset, N., and Carazo, J. (2007). Fast, robust, and accurate determination of transmission electron

microscopy contrast transfer function. *J. Struct. Biol.* 160, 249–262. doi: 10.1016/j.jsb.2007.08.013

Sorzano, C. O., de la Rosa Trevín, J., Otón, J., Vega, J., Cuenca, J., Zaldívar-Peraza, A., et al. (2013). "Semiautomatic, high-throughput, high-resolution protocol for three-dimensional reconstruction of single particles in electron microscopy," in *Nanoimaging*, eds A. A. Sousa and M. J. Kruhlak (New York, NY; Heidelberg; Dordrecht; London: Springer), 171–193. doi: 10.1007/978-1-62703-137-0_11

Sorzano, C. O. S., Alvarez-Cabrera, A. L., Kazemi, M., Carazo, J. M., and Jonić, S. (2016). Structmap: elastic distance analysis of electron microscopy maps for studying conformational changes. *Biophys. J.* 110, 1753–1765. doi: 10.1016/j.bpj.2016.03.019

Sorzano, C. O. S., de La Rosa-Trevín, J. M., Tama, F., and Jonić, S. (2014). Hybrid electron microscopy normal mode analysis graphical interface and protocol. *J. Struct. Biol.* 188, 134–141. doi: 10.1016/j.jsb.2014.09.005

Sorzano, C. O. S., Jiménez, A., Mota, J., Vilas, J. L., Maluenda, D., Martínez, M., et al. (2019). Survey of the analysis of continuous conformational variability of biological macromolecules by electron microscopy. *Acta Crystallogr. F* 75, 19–32. doi: 10.1107/S2053230X18015108

Stölken, M., Beck, F., Haller, T., Hegerl, R., Gutsche, I., Carazo, J. M., et al. (2011). Maximum likelihood based classification of electron tomographic data. *J. Struct. Biol.* 173, 77–85. doi: 10.1016/j.jsb.2010.08.005

Suhre, K., Navaza, J., and Sanejouand, Y. H. (2006). Norma: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps. *Acta Crystallogr. D Biol. Crystallogr.* 62, 1098–1100. doi: 10.1107/S090744490602244X

Suhre, K., and Sanejouand, Y. H. (2004). ELNEMO: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.* 32, W610–W614. doi: 10.1093/nar/gkh368

Tagare, H. D., Kucukelbir, A., Sigworth, F. J., Wang, H., and Rao, M. (2015). Directly reconstructing principal components of heterogeneous particles from cryo-em images. *J. Struct. Biol.* 191, 245–262. doi: 10.1016/j.jsb.2015.05.007

Tama, F., and Brooks, C. L. III. (2006). Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. *Annu. Rev. Biophys. Biomol. Struct.* 35, 115–133. doi: 10.1146/annurev.biophys.35.040405.102010

Tama, F., Gadea, F. X., Marques, O., and Sanejouand, Y. H. (2000). Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins* 41, 1–7. doi: 10.1002/1097-0134(20001001)41:1<1::AID-PROT10>3.0.CO;2-P

Tama, F., Miyashita, O., and Brooks III, C. L. (2004b). Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-em. *J. Struct. Biol.* 147, 315–326. doi: 10.1016/j.jsb.2004.03.002

Tama, F., Miyashita, O., and Brooks, C. L. (2004a). Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J. Mol. Biol.* 337, 985–999. doi: 10.1016/j.jmb.2004.01.048

Tama, F., and Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Prot. Eng.* 14, 1–6. doi: 10.1093/protein/14.1.1

Tama, F., Wriggers, W., and Brooks, C. L. III. (2002). Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. *J. Mol. Biol.* 321, 297–305. doi: 10.1016/S0022-2836(02)00627-7

Tirion, M. M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* 77, 1905–1908. doi: 10.1103/PhysRevLett.77.1905

Wan, W., and Briggs, J. (2016). "Chapter 13–cryo-electron tomography and subtomogram averaging," in *The Resolution Revolution: Recent Advances in cryoEM, Volume 579 of Methods in Enzymology*, ed R. Crowther (Cambridge: Academic Press), 329–367. doi: 10.1016/bs.mie.2016.04.014

Wan, W., Kolesnikova, L., Clarke, M., Koehler, A., Noda, T., Becker, S., et al. (2017). Structure and assembly of the ebola virus nucleocapsid. *Nature* 551, 394–397. doi: 10.1038/nature24490

Wang, Y., Rader, A., Bahar, I., and Jernigan, R. L. (2004). Global ribosome motions revealed with elastic network model. *J. Struct. Biol.* 147, 302–314. doi: 10.1016/j.jsb.2004.01.005

Xu, M., Beck, M., and Alber, F. (2012). High-throughput subtomogram alignment and classification by fourier space constrained fast volumetric matching. *J. Struct. Biol.* 178, 152–164. doi: 10.1016/j.jsb.2012.02.014

Zhai, X., Lei, D., Zhang, M., Liu, J., Wu, H., Yu, Y., et al. (2020). LoTTor: an algorithm for missing-wedge correction of the low-tilt tomographic 3D reconstruction of a single-molecule structure. *Sci. Rep.* 10:10489. doi: 10.1038/s41598-020-66793-1

Zhang, L., and Ren, G. (2012). High-resolution single-molecule structure revealed by electron microscopy and individual particle electron tomography. *J. Phys. Chem. Biophys.* 2:2. doi: 10.4172/2161-0398.1000e103

Zhou, A., Rohou, A., Schep, D. G., Bason, J. V., Montgomery, M. G., Walker, J. E., et al. (2015). Structure and conformational states of the bovine mitochondrial atp synthase by cryo-em. *Elife* 4:e10180. doi: 10.7554/eLife.10180

Zlatanova, J., Bishop, T. C., Victor, J., Jackson, V., and van Holde, K. V. (2009). The nucleosome family: dynamic and growing. *Structure* 17, 160–171. doi: 10.1016/j.str.2008.12.016

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 © 2021 Harastani, Eltsov, Leforestier and Jonic. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 07 June 2021
doi: 10.3389/fmolb.2021.692230

Nascent Folding of Proteins Across the Three Domains of Life

Mateusz Chwastyk and Marek Cieplak*

Institute of Physics, Polish Academy of Sciences, Warsaw, Poland

We study the nascent behavior of three model coarse-grained proteins in six rigid all-atom structures representing ribosomes that come from three domains of life. The synthesis of the proteins is implemented as a growth process. The geometry of the exit tunnel is quantified and shown to differ between the domains of life: both in volume and the size of constriction sites. This results in different characteristic times of capture within the tunnel and various probabilities of the escape. One of the proteins studied is the bacterial YibK which is knotted in its native state. A fraction of the trajectories results in knotting and the probability of doing so is largest for the bacterial ribosomes. Relaxing the condition of the rigidness of the ribosomes should result in a better avoidance of trapping and better proper folding.

Keywords: ribosome exit tunnel, geometry, molecular dynamics, coarse-grained models, protein folding, proteins with knots

OPEN ACCESS

Edited by:
Joanna Trylska,
University of Warsaw, Poland

Reviewed by:
Patrícia FN Faísca,
University of Lisbon, Portugal
Edward Patrick O'Brien,
Pennsylvania State University (PSU), United States
Cristian Micheletti,
International School for Advanced Studies (SISSA), Italy

****Correspondence:***
Mateusz Chwastyk
chwastyk@ifpan.edu.pl

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *07 April 2021*
Accepted: *19 May 2021*
Published: *07 June 2021*

Citation:
Chwastyk M and Cieplak M (2021) Nascent Folding of Proteins Across the Three Domains of Life. Front. Mol. Biosci. 8:692230. doi: 10.3389/fmolb.2021.692230

INTRODUCTION

Ribosome is a biomolecular nanomachine that performs protein synthesis at its peptidyl-transferase center (PTC) as directed by an mRNA template. The schematic picture of the protein being created by the ribosome is presented in **Figure 1**. In terms of the evolution, the PTC has been recognized as the earliest part of the ribosome (Prosdocimi et al., 2020). Ribosome itself is an aggregate made of 2–6 RNA chains and around 50 proteins comprising altogether between 100,000 and 220,000 atoms. Six examples of the ribosomal structures, together with some parameters of their description, are listed in **Table 1**. Ribosomes are involved in the regulation of translation and they influence the folding process (Kaiser et al., 2011). The synthesis takes place at the rate that depends on the domain of life. For prokaryotes, it is about 20 amino acids (Young and Bremer, 1976) and for eukaryotes four amino acids (Boström et al., 1986; Ingolia et al., 2011). The PTC secretes the nascent protein into the exit tunnel that ends with a "mouth" that opens into the surrounding solvent. After the detachment from the center, the protein escapes the ribosome in less than 1 ms (Bui and Hoang, 2020; Nissley et al., 2020). The movement of the protein toward the solvent gets started at the PTC and then it is influenced mostly by diffusion, interactions with the walls of the tunnel, and the gain in the entropy associated with the escape. The walls of the tunnel are rough and its diameter varies between 10 and 20 Å (Voss et al., 2006; Cabrita et al., 2010; Frank and Gonzalez, 2010). It has been established (Melnikov et al., 2012; Dao Duc et al., 2019; Liutkute et al., 2020) that the very geometry of the tunnel depends on the domain of life. In particular, we find that the diameter in the bacterial ribosomes can reach even 30 Å. This happens at the branching points in the tunnel.

Here, we present results of a theoretical study in which we probe the impact of the nature of the ribosome on protein folding in a coarse-grained model. Specifically, we consider six ribosomes from three domains of life, bacteria, archea and eukarya, and elucidate the difference in behavior of three model bacterial proteins in the six ribosomal tunnels. The bacterial ribosomes considered correspond to structures PDB:5NJT (Beckert et al., 2017) and PDB:5AFI (Fischer et al., 2015) that originate from

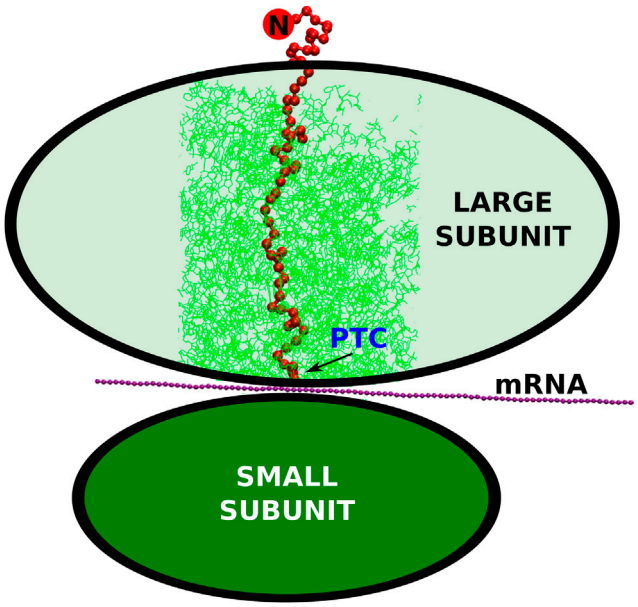


FIGURE 1 | The schematic representation of the small and large ribosomal subunits (the green ovals). The arrow points to the peptidyl-transferase center where the conversion of the nucleic acid genetic information (mRNA) into the polypeptide (the red string) takes place. The cylindrical cut-out, considered in our simulations, is marked by the thin, green lines. The N-terminus of the synthesized protein is marked by the red circle with the letter "N".

B. subtilis and *E. coli* respectively. The remaining structures and their origins are listed in **Table 1**. The first protein selected is the streptococcal protein G (Gronenborn et al., 1991) (PDB:1GB1) of 56 residues that is a common object of simulational studies. The second is YibK (Lim et al., 2003) (PDB:1J85) derived from *Haemophilus influenze*. YibK consists of 156 residues. The reason to consider it is that its native structure contains the deep trefoil knot between sites 75 and 119. It means that the backbone of the protein entangles themselves in to a knot. In the case of deeply knotted proteins, the transient state during the knotting process is often a slipknot. It is a conformation in which one of the protein termini adopts a hairpin-like conformation that threads a loop formed by the remainder of the chain (Faísca, 2015). The ribosomal action has been proposed (Chwastyk and Cieplak, 2015) to solve the puzzle of how the deeply knotted proteins form. An additional mechanism that may enhance knotting still further involves confinement generated by post-translational action of chaperonins (Takagi et al., 2003; Mallam and Jackson, 2011; Soler et al., 2016; Zhao et al., 2017; Especial et al., 2019; Chwastyk et al., 2021). It should be noted that the ribosome itself also provides a confining space that may favor formation of secondary structures (Elcock, 2006). The last protein we consider is Trp-Cage miniprotein (PDB:1L2Y) (Neidigh et al., 2002). It is composed of 20 amino acids and its small size allows it to fold near the PTC.

Theoretical studies of cotranslational folding, especially when knotting is involved, should start by considering simple coarse-grained structure-based models (see, e.g., ref (Sikora et al., 2009)) as they introduce a bias toward the native state. In our previous paper (Chwastyk and Cieplak, 2015), the mouth of the ribosome has been represented as an infinite repulsive plate which grows proteins by starting from the N-terminus. The plate has turned out to be positioning the planar knot loop of YibK (between sites 75–95) in a way that allows for formation of a C-terminal slipknot and then threading of the C-terminus through the loop on detachment. The effectiveness of this mechanism depends on the scheme to derive the contact map and it increases with time separation between the successive events of the emergence of new residues, t_w. The longest used was 5,000 τ where τ is of order 1 ns—the characteristic time scale of the CG simulation (Sikora et al., 2009). A better approach is to implement a steady growth at the PTC (Hoang and Cieplak, 2000; Krobath et al., 2013; Bui and Hoang, 2020).

Another simple model (Dabrowski-Tumanski et al., 2018) represents the exit tunnel as a smooth funnel-like potential. It is combined with an axial force that acts on the fully formed sequence placed near the PTC and pushes it toward the mouth. Necessarily, this force must induce an acceleration which is likely spurious. This approach was used for a large deeply knotted protein Tp0624 of 421 residues. The crucial ingredient in achieving high effectiveness in knotted folding was adding attractive centers at the mouth of the exit tunnel.

TABLE 1 | Six ribosomal structures considered in our analysis.

	PDB	Organelle	MM [kDa]	P_{ch}/RNA_{ch}	V_T (R = 4Å) [Å^3]	V_T (R = 3Å) [Å^3]	S_T/V_T [1/Å]	C_{at}/R_{at}
bacteria	5NJT	*B. subtilis* Beckert et al. (2017)	2,111.70	48/3	27145 ± 407	34920 ± 490	0.33	11971/133097
	5AFI	*E. coli* Fischer et al. (2015)	2,302.71	52/6	25628 ± 255	36295 ± 779	0.33	12330/152717
archaea	4V9F	*H. marismortui* Gabdulkhakov et al. (2013)	1,486.40	31/2	11333 ± 132	15188 ± 156	0.41	14499/103831
	4V6U	P. furiosus Armache et al. (2012)	2,602.99	63/5	16446 ± 216	24348 ± 362	0.38	13052/173979
eukarya	6EK0	H. sapiens Natchiar et al. (2017)	3,842.38	76/5	16822 ± 246	26838 ± 248	0.41	12702/219596
	5XY3	T. vaginalis Li et al. (2017)	1713.54	40/3	20043 ± 175	25213 ± 229	0.40	11680/107145

The bacterial ribosomes considered correspond to structures PDB:5NJT and PDB:5AFI that originate from B. subtilis *and* E.coli *respectively. The archaeal ribosomes were from* H. marismortui *and from* P. furiosus *(PDB:4V9F and PDB:4V6U) The last two eukarya ribosomes are eukaryotic and they come from* H. sapiens *(PDB:6EK0) and from* T. vaginalis *(PDB:5XY3). The second column shows the PDB code of the particular structure. The third column lists the source organism (italicized) and the references for the presented data. The fourth column lists the total structure molecular weight; the fifth column lists the number of unique proteins and, after the slash of the, nucleic acid chains. The sixth column lists the volume of the exit tunnel calculated with the probe radius* R = 4Å *and the seventh for* R = 3Å*. The eighth column lists the surface-area-to-volume ratio for the exit tunnels. The last column lists the fraction of the atoms within the cylindrical cut-out, considered in our simulations, to the total number of atoms that make the full structure of the ribosome. The two numbers differ by an order of magnitude.*

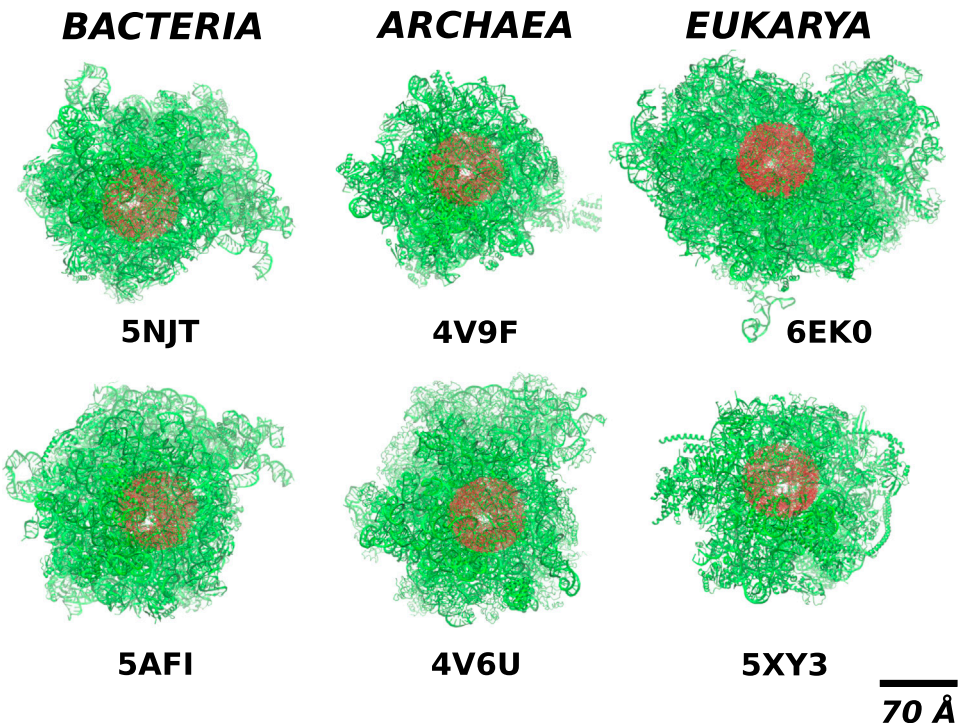


FIGURE 2 | The location of the cylindrical cut-out (the red shadow) of radius $R_R = 70$Å within the full structure of ribosomes (the green colour) from different domains of life. The white spot within the red blob shows the position of the exit tunnel.

However, we find that such centers generate trapping at the mouth.

Making comparisons between the domains of life, however, cannot be based on the ribosome models that are just generic. Here, we consider a coarse-grained model (Chwastyk et al., 2021) in which the growth takes place at the PTC and the tunnel has the shape that is determined by the structure file of the ribosome so that it is sensitive to the species (In ref. (Chwastyk et al., 2021), we have discussed the bacterial structure PDB:4V4J from *Thermus thermophilus*.). We first compare the geometries of the tunnels arising in the six ribosomes of **Table 1** and show how they differ across the domains of life. We then compare the folding processes and show that the bacterial ribosomes enhance the folding and knotting processes stronger than the other ribosomes. The relatively small average diameters of the archaea and eukarya ribosomal exit tunnels make the protein growing process difficult. The most efficient folding process happens within the ribosomal structure which is natural for a given protein.

The Geometry of the Ribosomal Exit Tunnels

The complexity of the full ribosomal structure makes it hard to include all of the atoms in the simulations. Thus we limit the number of relevant atoms to a cylindrical cut-out of radius $R_R = 70$ Å around the ribosomal exit tunnel as illustrated in **Figures 2**, **3**. The cylinder originates at the plane that goes through the PTC. It is oriented towards the mouth of the ribosome as shown in **Figure 2** against the background of the full ribosome. The structures around the ribosomal exit tunnel incorporated in our simulations are presented in **Figure 4**. Each cylinder is composed of rRNA and ribosomal proteins atoms. The specific number of nucleic acid and protein chains are listed in **Table 1**. Here, we consider these atoms to be fixed rigidly during the dynamics but an improved model should allow for their flexibility.

The location of the PTC was determined based on the data presented by Dao Duc et al. (2019)—it is found at the extension of the L22 and L4 proteins (Trylska, 2009; Dao Duc et al., 2019). In order to identify the tunnels and their volumes, we used the SPACEBALL algorithm (Chwastyk et al., 2014; Chwastyk et al., 2016a; Chwastyk et al., 2016b). The algorithm involves sending probing particles along main directions of a grid (at ten rotations of the ribosome) and checking which grid sites are accessible. Usually, one takes the radius of the probing particles to be 1.4 Å which is equal to the diameter of the water molecule. Here, we use primarily the radius of 4 Å in order not to penetrate the all-atom representation of the walls of the tunnel, but we also make comparisons to calculations done with the radius of 3 Å.

Our calculated volumes are somewhat smaller in comparison to the results obtained by Dao Duc et al. (2019). They got $(3.85 \pm 0.37)\times10^4$ Å^3 by considering 10 bacteria and $(2.78 \pm 0.13)\times10^4$ Å^3 for nine eukarya. When we decreased the radius of our probe to 3 Å we got very similar results, listed in **Table 1**. The difference in our algorithm is that we do not specify the position of the tunnel before undertaking the volume calculation. In addition, we make rotations of the grid. In the

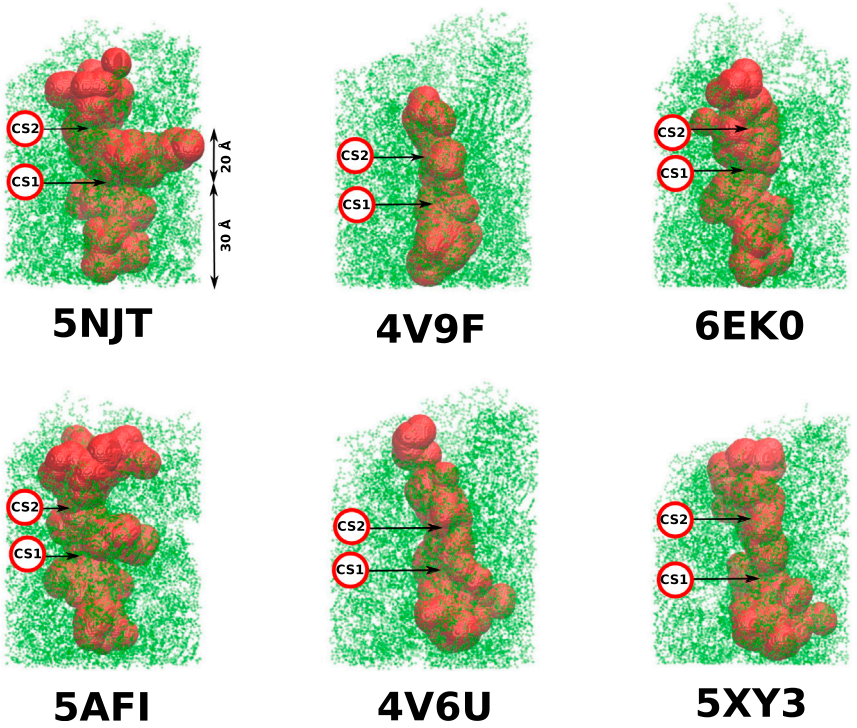


FIGURE 3 | Structures of the cylindrical cut-outs around the ribosomal exit tunnels from the three different domains of life. Each of the six panels shows the all-atom representation of the cylinder. The colors are meant just to generate a sense of perspective. The structures shown below the main picture represents the very exit of the cylinder. Up to five colors are used to indicate the vertical positions. The ball-like structure on the right of the cylinder represents the top view of the exit mouth.

first step, Dao Duc et al. (2019) search for the tunnel position by using MOLE 2.0 (Sehnal et al., 2013) program and then implement HOLLOW (Ho and Gruswitz, 2008) approach to determine the volume. The HOLLOW approach is very similar to our method but for well defined structure. This allows for the usage of smaller probes and thus to obtain larger volumes.

We observe that the bacterial ribosomes are associated with tunnels of the largest volume and are thus expected to obstruct the motion toward the exit the least. The SPACEBALL-derived tunnel spaces are shown in **Figure 4** in the red color. Despite the existence of pronounced side channels, the bacterial tunnels come with much fewer regions that are difficult to access. This was confirmed just by looking at the detected exit tunnel and by calculating the surface-area-to-volume ratio for the exit tunnels (see **Table 1**). The smallest value of this parameter was obtained for the bacterial exit tunnels and the largest one, for the eukaryotic tunnels.

The details of the cylindrical structures are shown in **Figure 3**. They are focused on the mouth regions and demonstrate that the openings are fairly irregular—they are not smooth funnel-like surfaces. The panels in **Figure 3** also show the vertical distances from the PTC plane to the most distant atoms. These distances vary between 94 Å and 109 Å and do not distinguish between the domains of life.

As in Dao Duc et al. (2019), we find that there are several constriction sites (CS) in the ribosomes. The first of these, CS1, is located around 30 Å from the PTC for each of the six ribosomes studied. The second, CS2, is at 20 Å farther away from CS1. The width of CS2 depends on the domain of life: it is narrow in the eukaryotic case (the radius is around 8 Å), a bit wider for the archaea ribosomes (the average radius is around 11 Å), and still wider for the bacterial case (the average radius is around 15 Å). The locations of the CS1 and CS2 are marked in **Figure 4**.

COTRANSLATIONAL FOLDING

Description of the Molecular Dynamics Model

The protein is modeled within the structure-based approach, as described in refs (Sulkowska and Cieplak, 2008; Sikora et al., 2009). with a chirality potential being responsible for the backbone stiffness. The contact interactions are selected by using the overlap criterion (Wołek et al., 2015) between the atoms of the residues as determined in the fully folded native state. The contacts correspond to the potential wells between the effective residues are located at the α-C atoms. The depth of the wells is denoted as ε. The remaining interactions are softly repulsive with the characteristic length of 4 Å. When one starts from an extended conformation and studies folding then

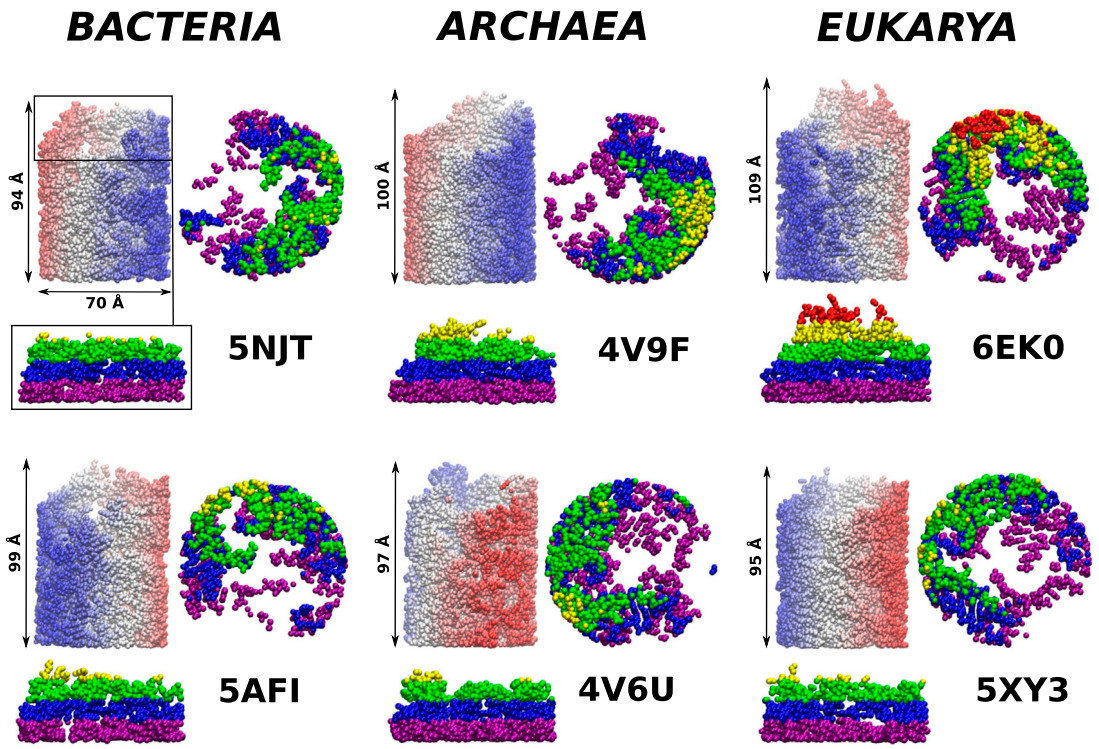


FIGURE 4 | The ribosomal exit tunnels (the red shape) detected by the SPACEBALL algorithm (Chwastyk et al., 2014; Chwastyk et al., 2016a; Chwastyk et al., 2016b) within the ribosomal structures from different domains of life. The arrows point to the constriction sites (CS1 and CS2). The green shadow represents the all atoms that are present in the cut-out part of the ribosomes. The PTC is always at the bottom of each panel.

the process is declared accomplished if all contacts get established. i.e., the distance between the residues involved becomes smaller than 1.5 σ, where σ denotes the width of the corresponding well. The temperature is controlled by random forces and the room temperature, T_R is around $0.35\varepsilon/k_B$. Each of the ribosomal atoms is a source of the soft repulsive potential that is cut at 4 Å and has the amplitude of ε.

Dynamically, the bottom of the cylinder is represented by a repulsive wall with the potential $\frac{3\sqrt{3}}{2}\varepsilon(\sigma_0/z)^9$, where z denotes the distance away from the plate and $\sigma_0 = 4 \cdot 2^{-1/6}$. This wall prevents making any backward steps. The models that incorporate the sequential growth consider it taking place either at the PTC or, effectively, at the mouth.

The most common practice is to place a fully synthesized chain near the PTC (Frank and Gonzalez, 2010; Elcock, 2006; Dabrowski-Tumanski et al., 2018; Nissley et al., 2020) and then to monitor folding (Nilsson et al., 2015; Kudva et al., 2018; Bock et al., 2018). The exit through the mouth can be helped computationally by switching from a CG simulation to a steered all-atom molecular dynamics approach with a steady motion of a pulling cantilever (Nissley et al., 2020). Another method is to apply a constant force (Dabrowski-Tumanski et al., 2018). Here, we incorporate the sequential growth at the PTC. Each of the amino acids emerges with some time interval after the previous one created earlier. The direction of the amino acid motion is given by the repulsive potential accelerating the created bead toward the exit of the tunnel. Since the mRNA is translated from the 5' to 3' ends, the proteins are synthesized from the N terminus to the C terminus, so the N terminus emerges first, as presented in **Figure 1**. This method is similar to the one considered in ref (Bui and Hoang, 2020). There are several differences, however: 1) all protein and RNA atoms of the walls provide repulsion (not just the α-C atoms), 2) the backward motion is prevented by the repulsion from the bottom wall, 3) the growth is implemented in a quasi-continuous fashion, 4) in one variant of the model, used for the YibK protein, we introduce electrostatics-mimicking contacts at the mouth of the ribosome, similarly to Dabrowski-Tumanski et al. (2018). The strength of these contacts is the same as in the case of the intramolecular contacts. We declare that the contacts can be created between two amino acids with opposed electric charge. We have found that the presence of such contacts do not impact the knotting process, but impedes the dissociation of the protein from the ribosome, so the results presented here have been obtained without these attractive contacts.

We represent our results as in **Figure 5**. It shows the histograms of distances of the N-terminal points of the proteins awya from the PTC. The bin size was chosen so that the first bin corresponds to the trajectories that are stuck between the PTC and CS1, the second one—between CS1 and CS2 and the others—the trajectories that resulted in proteins leaving the ribosome.

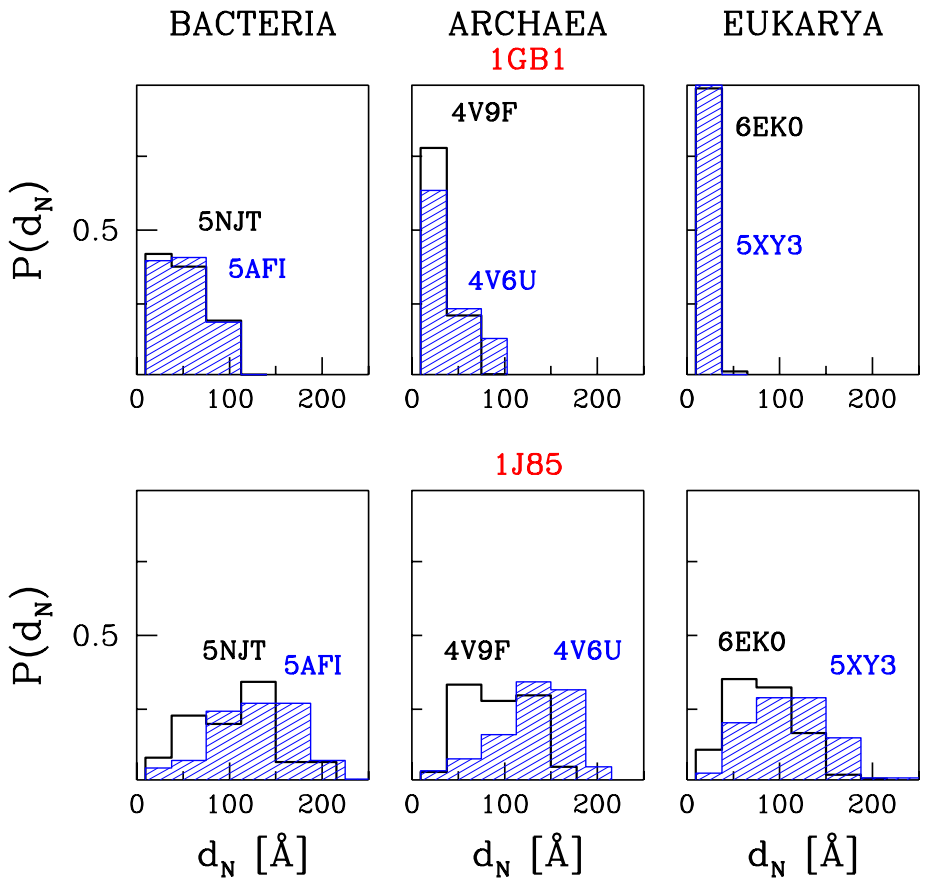


FIGURE 5 | The histograms of the distances between PTC and the N-terminus after time t_E of simulations at $T_R = 0.35\varepsilon/k_B$ and with $t_w = 100\tau$. The results are for three different domains of life. The upper panels are for 1GB1, with $t_E = 50000\tau$, and the lower ones for 1J85, with $t_E = 150000\tau$.

Protein G

The upper panels of **Figure 5** are for protein G. The temperature is equal to T_R and t_w is set to 100 τ, similarly to ref. (O'Brien et al., 2011). The panels show that the most difficult part of the movement of the growing protein is to pass through constriction CS1. For the two eukaryotic ribosomes, almost all of the trajectories get stuck at CS1. In the case of the two archaea ribosomes, this happens with 60–80% of the trajectories and around 20% get stuck at CS2. However, the bacterial ribosomes offer an easier passage for protein G: 40% of the trajectories stop at CS1, another 40% at CS2. Thus around 20% reach the mouth of the ribosome. Folding takes place after the exit and the median folding time of protein G in the 5NJT and 5AFI ribosomes is 7747τ and 7566τ respectively. An illustration of the various stages of the process is given in **Figure 6**. The folding process is considered to be accomplished when all native contacts are established for the first time (the distance between the corresponding α-C atoms is smaller than 1.5 σ where σ is the length parameter associated with the potential well. We considered 90 folding trajectories for each of the ribosomes.

When t_w is increased to 500τ, the number of trajectories that get stuck at the CS1 becomes similar for the two eukaryotic ribosomes and for the 4V6U archaea ribosome. However, it gets halved for the bacterial ribosomes while it increases from 80 to 100% for 4V9F (archaea). A further increase in t_w results in an increased probability of stopping at CS1 because there is more time to penetrate the exit tunnel nooks. The eukarya ribosomes do not allow for proteins leaving the ribosomal exit tunnel. We expect that when the walls are made flexible the protein will have a better chance of negotiating the constrictions at any t_w, but this remains to be demonstrated. These results are based on 60 trajectories for each ribosome.

Protein YibK

We now consider the knotted YibK protein. The knot ends are at LEU-75 and LYS-119 in the native state. We examine the impact of the full ribosomal structure on the knotting process. Unfortunately, it was impossible to reach $t_w = 5000\tau$, as in ref (Chwastyk and Cieplak, 2015). that this time gives the highest value of the percentage-wise success of reaching the properly knotted folded conformation. t_w longer than 100τ decreases the number of trajectories with the full exit of the protein. To be able to examine the knotting process we did a numerical trick

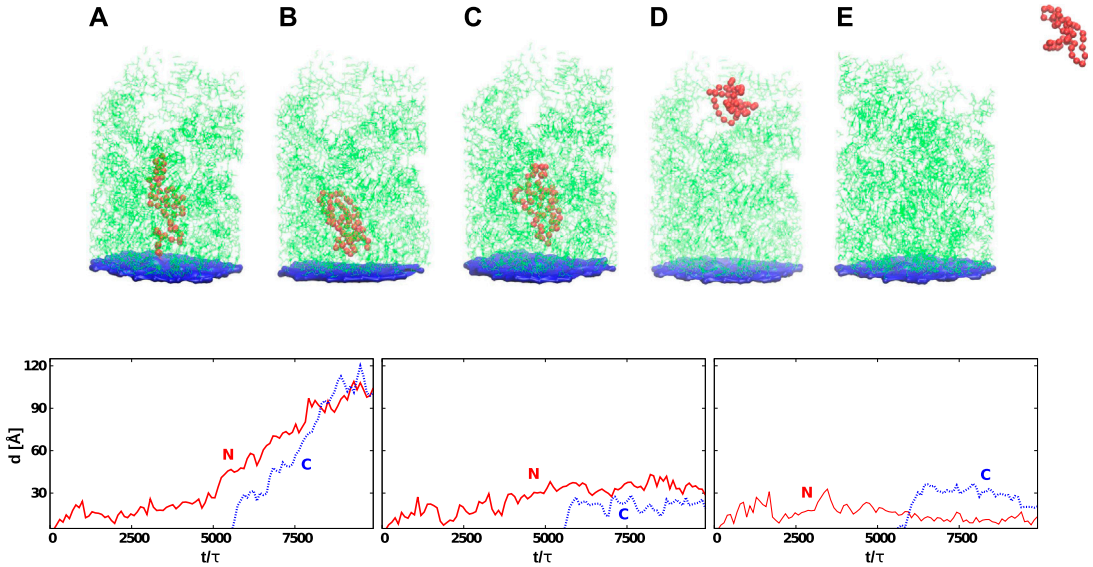


FIGURE 6 | Top: Six snapshots of protein G (the red color) at various stages of its development that start at its synthesis **(A)** and, in many cases, end with the detachment **(E)**. The average time needed by the ribosome to release the protein depends on t_w. The exit tunnel (in green) corresponds to the eukaryotic structure PDB: 6EK0. The blue plane represents the origin of the repulsive potential. The PTC is located at the center of the plane. In **(B)**, the protein is stuck permanently at the CS1. In **(C)**, the jamming is at CS2. In **(D)**, the protein has reached the mouth of the tunnel. **Bottom:** Distances between the PTC and the N-terminus (solid-red line) or C-terminus (dotted-blue line) as a function of time. The left panel corresponds to the situation presented in **(D)** (protein is leaving the ribosome). The graph in the center corresponds to the situation shown in **(C)** (the protein gets stuck at CS2). The last graph shows the situation corresponding to **(B)** (the protein is stuck between PTC and CS1). In all cases, the C-terminus gets released after the synthesis of the 56-residue protein which, for t_w = 100τ, corresponds to 5,600 τ.

and we allowed for a fast growth (with $t_w = 100\tau$) at the beginning of the protein synthesis and then switched to a slower process with $t_w = 1000\tau$ when the process reached LEU-75. This trick generates between 63 to 70 trajectories out of 100, depending on the ribosome, that manage to escape from the tunnel.

For this set of successful trajectories, about 10% resulted in formation of non-native temporal knots with one knot end at LYS-119. Another ~10% resulted in temporal knots with improperly located knot ends. We expect, however, that an extension of the simulation will eventually move the ends to the proper locations. The most interesting situation is observed for the 5AFI ribosome. Only in this case, 1% of 1,000 trajectories resulted with the properly folded and knotted conformations. The knotting mechanism was based on a slipknot transient conformation, the same as observed in ref. (Chwastyk and Cieplak, 2015). In addition, 10% of the trajectories resulted in one temporal knot end located at LYS-119 and 5% of the trajectories with temporal knot elsewhere. Again, an extension of the simulations is expected to move the knot ends to their native locations. When we increased the temperature from T_R to 0.45 ε/k_B the total number of trajectories that escape the tunnel got decreased, but the success in folding and knotting was similar to the one at T_R.

Trp-Cage Miniprotein

In the case of the very short protein (1L2Y) which we also examined, all of the 100 trajectories got stuck at the CS1. Nevertheless, there is enough space between PTC and CS1 for this short protein to accomplish correct folding. This happens for all trajectories and the average folding time is 1979τ. The average here is over the trajectories and the six ribosomes.

CONCLUSION

We have demonstrated the existence of differences in the dynamical behavior of nascent proteins across the domains of life. These differences result from the dissimilarities in the ribosomal geometries. The tight eukarya ribosomal exit tunnels impede the protein movement towards the exit. The wider constrictions sites found in the two other domains of life allow for proteins for an easier squeezing through away from the PTC. It is important to mention that the percentage-wise success of reaching the properly knotted and folded conformations was found to be the highest for 5AFI ribosome that originates from the *E. coli* bacterium. The YibK protein is also bacterial (*Haemophilus influenzae*). The two bacteria have one common ancestor (de Rosa and Labedan, 1998) which suggests that the 5AFI ribosome is evolution-related to the one that makes YibK. It is thus natural to consider YibK in this ribosome and expect the best folding results in this case.

If the longer sequences manage to exit the mouth and start making globular structures then this process, in principle, should be helped by attractive contacts at the mouth by generating an extracting mechanism. However, we find such contacts to be

disruptive to the process because they also capture the extracted protein at the mouth. The reason why the attractive patches at the mouth appear to promote knotting in Dabrowski-Tumanski et al. (2018). may be related to the fact that the bulk driving force used also eliminates the capture. This point needs to be elucidated further.

In our current model, the ribosomal molecules are considered to be rigid. Relaxing this condition should result in better escape rates and possibly also in a stronger knotting success. Moreover, it will allow for usage of the all-atom protein model, because right now the side chains block the protein that attempt to squeeze through away from the PTC. The flexibility can be introduced either by connecting the atoms to fixed centers by harmonic springs or by coarse-graining the RNA and protein chains of the ribosome (Kudva et al., 2018). We hope to address these issues in future research. We also plan to use the Dynamical Structure-Based model (Mioduszewski and Cieplak, 2018) to consider the exit times for selected intrinsically disordered proteins and study their behavior in the tunnel.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

Both authors developed the model and wrote the paper. MCh performed the simulations.

FUNDING

This work has been supported by the National Science Centre (NCN), Poland, under grant No. 2018/31/B/NZ1/00,047. This work results within the collaboration of the COST Action CA17139 Eutopia. The computer resources were supported by the PL-GRID infrastructure.

REFERENCES

Armache, J.-P., Anger, A. M., Márquez, V., Franckenberg, S., Fröhlich, T., Villa, E., et al. (2012). Promiscuous Behaviour of Archaeal Ribosomal Proteins: Implications for Eukaryotic Ribosome Evolution. *Nucleic Acids Res.* 41, 1284–1293. doi:10.1093/nar/gks1259

Beckert, B., Abdelshahid, M., Schäfer, H., Steinchen, W., Arenz, S., Berninghausen, O., et al. (2017). Structure of the Bacillus Subtilis Hibernating 100S Ribosome Reveals the Basis for 70S Dimerization. *EMBO J.* 36, 2061–2072. doi:10.15252/embj.201696189

Bock, L. V., Kolář, M. H., and Grubmüller, H. (2018). Molecular Simulations of the Ribosome and Associated Translation Factors. *Curr. Opin. Struct. Biol.* 49, 27–35. doi:10.1016/j.sbi.2017.11.003

Boström, K., Wettesten, M., Borén, J., Bondjers, G., Wiklund, O., and Olofsson, S. O. (1986). Pulse-chase Studies of the Synthesis and Intracellular Transport of Apolipoprotein B-100 in Hep G2 Cells. *J. Biol. Chem.* 261, 13800–13806. doi:10.1016/s0021-9258(18)67090-5

Bui, P. T., and Hoang, T. X. (2020). Protein Escape at the Ribosomal Exit Tunnel: Effect of the Tunnel Shape. *J. Chem. Phys.* 153, 045105. doi:10.1063/5.0008292

Cabrita, L. D., Dobson, C. M., and Christodoulou, J. (2010). Protein Folding on the Ribosome. *Curr. Opin. Struct. Biol.* 20, 33–45. doi:10.1016/j.sbi.2010.01.005

Chwastyk, M., Cieplak, M., and Faisca, P. F. N. (2021). Ribosome- and Chaperonin-Assisted Knotting of Proteins. (submitted).

Chwastyk, M., and Cieplak, M. M. (2015). Cotranslational Folding of Deeply Knotted Proteins. *J. Phys. Condens. Matter* 27, 354105. doi:10.1088/0953-8984/27/35/354105

Chwastyk, M., Jaskólski, M., and Cieplak, M. (2014). Structure-based Analysis of Thermodynamic and Mechanical Properties of Cavity-Containing Proteins - Case Study of Plant Pathogenesis-Related Proteins of Class 10. *FEBS J.* 281, 416–429. doi:10.1111/febs.12611

Chwastyk, M., Jaskólski, M., and Cieplak, M. (2016). The Volume of Cavities in Proteins and Virus Capsids. *Proteins* 84, 1275–1286. doi:10.1002/prot.25076

Chwastyk, M., Panek, E. A., Malinowski, J., Jaskólski, M., and Cieplak, M. (2016). Properties of Cavities in Biological Structures – A Survey of the Protein Data Bank. *Front. Mol. Biosci.* 7, 591381.

Dabrowski-Tumanski, P., Piejko, M., Niewieczerzal, S., Stasiak, A., and Sulkowska, J. I. (2018). Protein Knotting by Active Threading of Nascent Polypeptide Chain Exiting from the Ribosome Exit Channel. *J. Phys. Chem. B* 122, 11616–11625. doi:10.1021/acs.jpcb.8b07634

Dao Duc, K., Batra, S. S., Bhattacharya, N., Cate, J. H. D., and Song, Y. S. (2019). Differences in the Path to Exit the Ribosome across the Three Domains of Life. *Nucl. Acids Res.* 47, 4198–4210. doi:10.1093/nar/gkz106

de Rosa, R., and Labedan, B. (1998). The Evolutionary Relationships between the Two Bacteria *Escherichia coli* and Haemophilus Influenzae and Their Putative Last Common Ancestor. *Mol. Biol. Evol.* 15, 17–27. doi:10.1093/oxfordjournals.molbev.a025843

Elcock, A. H. (2006). Molecular Simulations of Cotranslational Protein Folding: Fragment Stabilities, Folding Cooperativity, and Trapping in the Ribosome. *Plos Comput. Biol.* 2, e98. doi:10.1371/journal.pcbi.0020098

Especial, J., Nunes, A., Rey, A., and Faísca, P. F. (2019). Hydrophobic Confinement Modulates thermal Stability and Assists Knotting in the Folding of Tangled Proteins. *Phys. Chem. Chem. Phys.* 21, 11764–11775. doi:10.1039/c9cp01701a

Faísca, P. F. N. (2015). Knotted Proteins: A Tangled Tale of Structural Biology. *Comput. Struct. Biotechnol. J.* 13, 459–468. doi:10.1016/j.csbj.2015.08.003

Fischer, N., Neumann, P., Konevega, A. L., Bock, L. V., Ficner, R., Rodnina, M. V., et al. (2015). Structure of the *E. coli* Ribosome-EF-Tu Complex at Å Resolution by C S-Corrected Cryo-EM. *Nature* 520, 567–570. doi:10.1038/nature14275

Frank, J., and Gonzalez, R. L., Jr. (2010). Structure and Dynamics of a Processive Brownian Motor: the Translating Ribosome. *Annu. Rev. Biochem.* 79, 381–412. doi:10.1146/annurev-biochem-060408-173330

Gabdulkhakov, A., Nikonov, S., and Garber, M. (2013). Revisiting theHaloarcula marismortui50S Ribosomal Subunit Model. *Acta Crystallogr. D Biol. Cryst.* 69, 997–1004. doi:10.1107/s0907444913004745

Gronenborn, A., Filpula, D., Essig, N., Achari, A., Whitlow, M., Wingfield, P., et al. (1991). A Novel, Highly Stable Fold of the Immunoglobulin Binding Domain of Streptococcal Protein G. *Science* 253, 657–661. doi:10.1126/science.1871600

Ho, B. K., and Gruswitz, F. (2008). HOLLOW: Generating Accurate Representations of Channel and interior Surfaces in Molecular Structures. *BMC Struct. Biol.* 8, 49. doi:10.1186/1472-6807-8-49

Hoang, T. X., and Cieplak, M. (2000). Molecular Dynamics of Folding of Secondary Structures in Go-type Models of Proteins. *J. Chem. Phys.* 112, 6851–6862. doi:10.1063/1.481261

Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011). Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* 147, 789–802. doi:10.1016/j.cell.2011.10.002

Kaiser, C. M., Goldman, D. H., Chodera, J. D., Tinoco, I., Jr., and Bustamante, C. (2011). The Ribosome Modulates Nascent Protein Folding. *Science* 334, 1723–1727. doi:10.1126/science.1209740

Krobath, H., Shakhnovich, E. I., and Faísca, P. F. N. (2013). Structural and Energetic Determinants of Co-translational Folding. *J. Chem. Phys.* 138, 215101. doi:10.1063/1.4808044

Kudva, R., Tian, P., Pardo-Avila, F., Carroni, M., Best, R. B., Bernstein, H. D., et al. (2018). The Shape of the Bacterial Ribosome Exit Tunnel Affects Cotranslational Protein Folding. *eLife* 7, e36326. doi:10.7554/eLife.36326

Li, Z., Guo, Q., Zheng, L., Ji, Y., Xie, Y.-T., Lai, D.-H., et al. (2017). Cryo-EM Structures of the 80S Ribosomes from Human Parasites *Trichomonas Vaginalis* and *Toxoplasma Gondii*. *Cell Res* 27, 1275–1288. doi:10.1038/cr.2017.104

Lim, K., Zhang, H., Tempczyk, A., Krajewski, W., Bonander, N., Toedt, J., et al. (2003). Structure of the YibK Methyltransferase fromHaemophilus Influenzae (HI0766): A Cofactor Bound at a Site Formed by a Knot. *Proteins* 51, 56–67. doi:10.1002/prot.10323

Liutkute, M., Samatova, E., and Rodnina, M. V. (2020). Cotranslational Folding of Proteins on the Ribosome. *Biomolecules* 10, 97. doi:10.3390/biom10010097

Mallam, A. L., and Jackson, S. E. (2011). Knot Formation in Newly Translated Proteins Is Spontaneous and Accelerated by Chaperonins. *Nat. Chem. Biol.* 8, 147–153. doi:10.1038/nchembio.742

Melnikov, S., Ben-Shem, A., Garreau de Loubresse, N., Jenner, L., Yusupova, G., and Yusupov, M. (2012). One Core, Two Shells: Bacterial and Eukaryotic Ribosomes. *Nat. Struct. Mol. Biol.* 19, 560–567. doi:10.1038/nsmb.2313

Mioduszewski, Ł., and Cieplak, M. (2018). Disordered Peptide Chains in an α-C-based Coarse-Grained Model. *Phys. Chem. Chem. Phys.* 20, 19057–19070. doi:10.1039/c8cp03309a

Natchiar, S. K., Myasnikov, A. G., Kratzat, H., Hazemann, I., and Klaholz, B. P. (2017). Visualization of Chemical Modifications in the Human 80S Ribosome Structure. *Nature* 551, 472–477. doi:10.1038/nature24482

Neidigh, J. W., Fesinmeyer, R. M., and Andersen, N. H. (2002). Designing a 20-residue Protein. *Nat. Struct. Biol.* 9, 425–430. doi:10.1038/nsb798

Nilsson, O. B., Hedman, R., Marino, J., Wickles, S., Bischoff, L., Johansson, M., et al. (2015). Cotranslational Protein Folding inside the Ribosome Exit Tunnel. *Cel Rep.* 12, 1533–1540. doi:10.1016/j.celrep.2015.07.065

Nissley, D. A., Vu, Q. V., Trovato, F., Ahmed, N., Jiang, Y., Li, M. S., et al. (2020). Electrostatic Interactions Govern Extreme Nascent Protein Ejection Times from Ribosomes and Can Delay Ribosome Recycling. *J. Am. Chem. Soc.* 142, 6103–6110. doi:10.1021/jacs.9b12264

O'Brien, E. P., Christodoulou, J., Vendruscolo, M., and Dobson, C. M. (2011). New Scenarios of Protein Folding Can Occur on the Ribosome. *J. Am. Chem. Soc.* 133 (3), 513–526. doi:10.1021/ja107863z

Prosdocimi, F., Zamudio, G. S., Palacios-Pérez, M., Torres de Farias, S., and V. José, M. (2020). The Ancient History of Peptidyl Transferase center Formation as Told by Conservation and Information Analyses. *Life* 10, 134–151. doi:10.3390/life10080134

Sehnal, D., Svobodová Vařeková, R., Berka, K., Pravda, L., Navrátilová, V., Banáš, P., et al. (2013). MOLE 2.0: Advanced Approach for Analysis of Biomacromolecular Channels. *J. Cheminform.* 5, 39. doi:10.1186/1758-2946-5-39

Sikora, M., Sułkowska, J. I., and Cieplak, M. (2009). Mechanical Strength of 17 134 Model Proteins and Cysteine Slipknots. *Plos Comput. Biol.* 5, e1000547. doi:10.1371/journal.pcbi.1000547

Soler, M. A., Rey, A., and Faísca, P. F. N. (2016). Steric Confinement and Enhanced Local Flexibility Assist Knotting in Simple Models of Protein Folding. *Phys. Chem. Chem. Phys.* 18, 26391–26403. doi:10.1039/c6cp05086g

Sulkowska, J. I., and Cieplak, M. (2008). Selection of Optimal Variants of Go-like Models of Proteins through Studies of Stretching. *Biophys. J.* 95, 3174–3191.

Takagi, F., Koga, N., and Takada, S. (2003). How Protein Thermodynamics and Folding Mechanisms Are Altered by the Chaperonin Cage: Molecular Simulations. *Proc. Natl. Acad. Sci.* 100, 11367–11372. doi:10.1073/pnas.1831920100

Trylska, J. (2009). Simulating Activity of the Bacterial Ribosome. *Quart. Rev. Biophys.* 42, 301–316. doi:10.1017/s0033583510000028

Voss, N. R., Gerstein, M., Steitz, T. A., and Moore, P. B. (2006). The Geometry of the Ribosomal Polypeptide Exit Tunnel. *J. Mol. Biol.* 360, 893–906. doi:10.1016/j.jmb.2006.05.023

Wołek, K., Gómez-Sicilia, À., and Cieplak, M. (2015). Determination of Contact Maps in Proteins: a Combination of Structural and Chemical Approaches. *J. Chem. Phys.* 143, 243105. doi:10.1063/1.4929599

Young, R., and Bremer, H. (1976). Polypeptide-chain-elongation Rate in *Escherichia coli* B/r as a Function of Growth Rate. *Biochem. J.* 160, 185–194. doi:10.1042/bj1600185

Zhao, Y., Dabrowski-Tumanski, P., Niewieczerzal, S., and Sulkowska, J. I. (2017). The Exclusive Effects of Chaperonin on the Behavior of Proteins with 52 Knotknot. *Plos Comput. Biol.* 14, e1005970. doi:10.1371/journal.pcbi.1005970

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 © 2021 Chwastyk and Cieplak. 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 Molecular Biosciences

REVIEW
published: 07 June 2021
doi: 10.3389/fmolb.2021.676976

From System Modeling to System Analysis: The Impact of Resolution Level and Resolution Distribution in the Computer-Aided Investigation of Biomolecules

Marco Giulini[1,2], *Marta Rigoli*[1,2], *Giovanni Mattiotti*[1,2], *Roberto Menichetti*[1,2], *Thomas Tarenzi*[1,2], *Raffaele Fiorentini*[1,2] *and Raffaello Potestio*[1,2]*

[1]*Physics Department, University of Trento, Trento, Italy,* [2]*INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, Trento, Italy*

OPEN ACCESS

Edited by:
Fabio Trovato,
Freie Universität Berlin, Germany

Reviewed by:
Adam Liwo,
University of Gdansk, Poland
Philippe Derreumaux,
UPR9080 Laboratoire de Biochimie Théorique (LBT), France
Will Noid,
Pennsylvania State University (PSU), United States

***Correspondence:**
Raffaello Potestio
raffaello.potestio@unitn.it

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *06 March 2021*
Accepted: *06 May 2021*
Published: *07 June 2021*

Citation:
Giulini M, Rigoli M, Mattiotti G, Menichetti R, Tarenzi T, Fiorentini R and Potestio R (2021) From System Modeling to System Analysis: The Impact of Resolution Level and Resolution Distribution in the Computer-Aided Investigation of Biomolecules. Front. Mol. Biosci. 8:676976. doi: 10.3389/fmolb.2021.676976

The ever increasing computer power, together with the improved accuracy of atomistic force fields, enables researchers to investigate biological systems at the molecular level with remarkable detail. However, the relevant length and time scales of many processes of interest are still hardly within reach even for state-of-the-art hardware, thus leaving important questions often unanswered. The computer-aided investigation of many biological physics problems thus largely benefits from the usage of coarse-grained models, that is, simplified representations of a molecule at a level of resolution that is lower than atomistic. A plethora of coarse-grained models have been developed, which differ most notably in their granularity; this latter aspect determines one of the crucial open issues in the field, i.e. the identification of an optimal degree of coarsening, which enables the greatest simplification at the expenses of the smallest information loss. In this review, we present the problem of coarse-grained modeling in biophysics from the viewpoint of system representation and information content. In particular, we discuss two distinct yet complementary aspects of protein modeling: on the one hand, the relationship between the resolution of a model and its capacity of accurately reproducing the properties of interest; on the other hand, the possibility of employing a lower resolution description of a detailed model to extract simple, useful, and intelligible information from the latter.

Keywords: modeling, coarse-graining, molecular dynamics, proteins, biophysics

1 INTRODUCTION

Among the many revolutions that have spangled the 20th Century, the advent and diffusion of the computer is certainly one of the most momentous. Computing machines have impacted human life and society in practically all compartments, such as communication, work, information, education, health, and entertainment. The scientific environment is certainly one of the main leaders of this revolution, but it has been largely affected by it as well: in fact, computers have not only changed the way we do science, they also created new ways of doing science that were simply unthinkable before. Besides the "trivial" usage of computers in speeding up regular calculations (that is, to carry out the

job of Los Alamos' *human computers*[1] in a faster and more human-friendly manner), a novel technique arose that rapidly became pervasive of practically all scientific fields, as well as a field *per se*: computer simulations.

Among the synonyms of *simulation* we can find words such as *copy*, *facsimile*, *imitation*, *counterfeit*, and *fake*. Computer simulations are indeed all these things: while aiming at reproducing, as faithfully as possible, the real object of study, its properties, and its dynamics, they necessarily are but *the shadow of a dream*—the fictitious dance of a projection of the object. And yet, precisely in this intangible nature lies their power.

Simulations constitute a bridge between the experimental investigation of a system and its abstract, theoretical study. While the former relies on direct observation, probing, and quantitative measurement, the latter describes the system or phenomenon of interest in terms of quantities and relations among them, and carries out the investigation making use of mathematical manipulations. The computational approach takes from both: it presupposes a representation of the system in terms of rather idealized fundamental constituents, whose nature is closer to abstract Platonic entities rather than physical, "Aristotelian" ones. Such representation enables the investigation down to a level of detail that is practically and even fundamentally inaccessible to experiments; however, its usability in the study and comprehension of Nature presupposes that a one-to-one relation can be established between the constitutive elements of a real system and those of its *model*. The validity of the latter depends on the capacity of the modeler of identifying the essential features of a system and endowing the model with them; a model is just as good as the pieces of which it is made.

The field of application of this *computational microscope* (Lee et al., 2009) spans several orders of magnitude in space and time. Depending on the specific property or phenomenon of interest, various models can be employed that describe reality (or rather a part of it) in a relatively small length and time scale interval; no single model can be employed to study whatever system, for two reasons: our limited computational capacity, and the intrinsic limitations of the model.

At present, the most successful description we possess of the constituents of matter and their interactions is provided by the Standard Model of particle physics: even though the latter is an incomplete and effective[2] theory, it still provides the most powerful and predictive (Hanneke et al., 2011; Aoyama et al., 2015) framework for the investigation of physical reality; that is to say, this theory constitutes the sharpest conceptual device we currently have at hand to rationalize observed phenomena and predict new ones. Nonetheless, a straightforward and brute-force application of such model to the study of systems larger than a small atomic nucleus is practically unfeasible: in fact, the associated computational cost makes it impossible to simulate, in terms of relativistic quantum fields, even the smallest molecule for a physically interesting time scale. Hence the first of the two aforementioned limitations.

All models beyond the most fundamental one (if any) are affected by both shortcomings. Certainly there will be systems too large or processes too slow to be studied by means of any derived representation; additionally, all these non-fundamental, effective representations will have a range of validity beyond which the model does not make sense. Non relativistic quantum mechanics works well for slow, low-energy particles, but the *resolution* of the processes it can reproduce is limited from below; additionally, it is too complex to study systems composed by more than a few atoms. Fortunately, within appropriate ranges of time, size, and energy, further effective theories can be constructed, that allow one to incorporate quantum mechanical properties in classical potentials: this process, epitomized by the Born-Oppenheimer approximation, fills the gap between quantum and classical mechanics, and between the small world and the not-so-small (e.g., molecular) world.

In general, then, the larger the scale of the system, and the longer the time scales of the processes of interest, the harder it is to perform simulations at a given level of resolution. This limitation originates from the increasing duration of the simulation and, in turn, the necessity to employ larger and larger memory and computing power. However, even when sufficient computational resources are at hand, another issue lies before us, which is the capability to make sense of the simulation. A detailed description of a large macromolecule, e.g., one in which each atom is described as a point-like particle, is certainly sufficient to reproduce several properties that would not involve quantum mechanical features explicitly, but it might as well be *excessively detailed* for the purpose. A simplified representation of the system and its interactions might be sufficient to reproduce the process of interest.

A further reduction of resolution is thus possible, in which the system is not described in terms of atoms, but rather of effective interaction sites each of which is representative of a group of several atoms. A model whose resolution is lower than atomistic is commonly referred to as a *coarse-grained* (CG) model. CG models range all possible resolutions from a few atoms per site up to the continuum, and a plethora of strategies have been developed to parametrize them so as to reproduce one or more properties of the system of interest. In fact, exactly as any effective theory can be trusted in a limited range of length and times scales only, so it is for any particular CG model.

This apparently trivial observation opens up a crucial issue, whose practical and philosophical implications have just started to be studied (see **Figure 1**), namely the identification of the level of model detail that is the most appropriate for the study of a given phenomenon. In fact, the construction of a model is implicitly dictated by its purpose, and its usage implicitly complies (or should comply) with the range of validity in which the model is effective. Insofar, the decision of the model resolution has largely been based on intuition, and quantitative investigation of the appropriate level of detail is really just in its infancy.

A second, even more subtle issue is the definition of the appropriate resolution *distribution*, that is, whether each

[1]https://www.atomicheritage.org/history/human-computers-los-alamos.

[2]The SM is incomplete as it does not incorporate the general-relativistic theory of gravity; furthermore, it is an effective theory as it holds for energies lower than the Planck scale (see e.g., Burgess (2020)).

FIGURE 1 | In the construction of a model we are confronted with several questions, whose apparent philosophical quality entails a rather practical nature. In particular, we ask ourselves: Can one always coarse-grain? Is there an optimal level of resolution? Is a single level of resolution meaningful? How to identify the optimal resolution level or distribution? And how can we implement it in practice? (In the picture: "The thinker", A. Rodin, 1904).

system part should be represented with the same level of detail, or rather a modulation of the latter can be implemented so as to attribute higher resolution (and more computational resources) to a given region, while reducing the accuracy elsewhere. This task actually requires solving two problems: first, one has to determine what level of resolution can be employed, and where; second, one has to devise a model that guarantees the appropriate degree of accuracy to each region of the system, such that the various regions at different resolution can interact with one another seamlessly.

Besides the questions related to the construction of computational models of physical systems a further one lies, that tackles the issue of system representation from a different perspective, namely that of employing modeling strategies for the *analysis* of the system. Computer simulations of large systems are becoming increasingly more feasible, which bears with it two major consequences: on the one hand, the steady growth in the amount of data to make sense of, even for a single run; on the other hand, the increase in the complexity of the systems and processes that can be tackled, which naturally requires a richer and often system-specific toolbox of analysis instruments. These are essential to safely navigate the sea of data produced, and land to the shores of the system's *understanding*; the latter, however, can only be achieved through a process of *reduction and synthesis*, by which the vast amount of numbers crunched and spat out by the computer are distilled into a few, intelligible and interpretable parameters, their time evolution, and their relations.

This is indeed what is being done since the dawn of thermodynamics, as systems composed by bazillions of particles are eventually described and studied in terms of a handful of quantities (temperature, pressure, volume, chemical potential, compressibility, specific heat...). The necessity behind this procedure is the human incapacity of making sense of $\sim 10^{23}$ degrees of freedom; the reason behind the success of such a drastic program, which brings down that number of coordinates to less than ten, is the fact that indeed a full, *meaningful* characterization of the system is intrinsically achievable in terms of those few variables and no more than that.

In the case of a system as simple as a gas or a liquid, the identification of those parameters that are relevant and sufficient for a complete description and understanding of the system is straightforward and largely intuitive. When the object of study is a macromolecule, however, things might be more subtle: one can wonder if it is possible to devise an algorithmic procedure aimed at the identification of those variables in terms of which a simplified representation of the system can be achieved, which maximizes the insight about it while at the same time retaining the lowest number of descriptors. Questions such as this hold the promise of discriminating, in an unsupervised manner, the signal from the noise in the outcome of a computer simulation.

The scope of this review is to present and discuss in some detail the questions raised insofar. The extension and richness of the field of modeling and coarse-graining forces us to renounce at any expectation of exhaustiveness: we however hope to provide the readers with a sufficiently broad and organized overview to grasp and appreciate the variety and diversity of models and methods that have been developed in the context of computer simulations of macromolecules. Our focus will lie on applications to proteins. This choice has two reasons: first, any attempt at including more than this class of systems might have easily doubled the length of the manuscript, as the field of biological and artificial soft matter modeling is just as broad as the list of systems itself; second, all the issues we discuss find in proteins a most evident, remarkable, and interesting playground. Many problems that we pose make little to no sense in other contexts: for example, it is relatively uninspiring (even though

not devoid of insight (Harmandaris et al., 2006; Ohkuma and Kremer, 2017)) to wonder about a modulation of the model resolution in a long homopolymer, or to question whether an intrinsically optimal level of detail exists in the representation of a lipid bilayer. On the contrary, these questions can have as many different answers as the proteins they are applied to, due to the diversity of size, structure, function, and properties that these molecules exhibit.

The remainder of the paper is structured as follows. In **Section 2** we introduce some fundamental concepts upon which the procedures of modeling and coarse-graining are constructed. Albeit non-standard and universally accepted terms are introduced, these are sufficiently intuitive and serve the purpose of removing some of the potential ambiguities that such a broad and rich field entails. In **Section 3** we discuss the most fundamental models one commonly finds in the context of soft and biological matter simulation, namely atomistic models. We enter the field of coarse-graining in **Section 4**, where we recapitulate the main general ideas, and illustrate examples of the models and methods that have been developed. Specifically, in **Section 5** we focus on those models where the degree of detail is uniform through out the system, while in **Section 6** we consider those strategies that make use of two or more resolution levels in the same model. In **Section 7** we shift from the idea of modeling to that of *filtering*, that is, reading a simulation with a lower level of detail so as to discriminate those structural characteristics that entail the largest amount of information about the system properties. Finally, in **Section 8** we summarize with a few concluding thoughts.

2 REPRESENTATION, MODEL AND FILTER

In order to carry out some kind of computer-aided quantitative investigation of a macromolecular system, e.g., a molecular dynamics simulation, it is necessary to provide a *representation* of the system. By this term we refer to a collection of mathematical entities conceptually associated to a corresponding physical entity: for example, the physical entity "atomic nucleus" is associated to a point in three dimensions whose position in space is determined through its coordinates in a (usually three-dimensional) Cartesian space. This point is the mathematical entity associated to the physical atomic nucleus, and a collection of such points, one for each atom of the molecule and its environment, constitutes the representation of the system we feed the computer with.

As such, the representation is a static object. This does not mean that it cannot be informative *per se*: indeed, knowing the position of the molecule's atoms allows us to establish geometrical relationships among them, from which, in turn, we can infer properties that connect shape and function. A particularly intuitive example is provided by the crystallographic or NMR techniques that turn real molecules in a set of atomic coordinates. However, in order to make a step forward from structure to function it is necessary to expand the set of properties the representation is endowed with and, most importantly, to confer to it the capacity of actively producing dynamic information. To this end, we have to "dress" the representation with *interactions*, thus enabling it to evolve in time or to sample its accessible phase space. When properties and interactions are specified for a given representation of the system, we dub it a *model*. Models are thus mathematical idealisations of a system that, by means of an appropriate processing of their properties and interaction, can produce nontrivial information (e.g., time series, correlation and response functions, conformational sampling...).

Once a representation or a model are given, though, one can apply to them the same procedure that leads from the real-life system to the idealized representation. More specifically, given a representation that we think of as the more fundamental one, thus dubbed *first level representation* (FLR), it is possible to establish a quantitative relationship from its mathematical entities to those of a *second level representation* (SLR), typically given in smaller number than those of the first one. As an example, all the atoms constituting a protein in the first level representation can be associated to one single point for each amino acid, so that the second level representation constitutes a *simplified* description of the first. If this procedure is applied to all configurations obtained in a molecular dynamics simulation, the result is a trajectory generated with the model defined at the first level, but described in terms of the second level representation. Hence, the SLR cannot produce nontrivial information by itself, but it can return a *subset* of the information produced by the underlying model; because of this property we refer to such SLR overlaid on a model as a *filter*[3]. These ideas are illustrated in **Figure 2**.

For a given system, one can provide representations at different levels of resolution: restricting our considerations to particle-based representations of proteins, we can let each of these particles represent an atom, part of an amino acids, an entire amino acid, a group of amino acids, an entire protein and so on. When the size of the system is such that a particle-based description of it does not make sense any more, continuum or quasi-continuum representations come into play, such as finite elements representations, where the surface of a protein is described in terms of a triangular tessellation, or descriptions involving density fields. Each of these representations is informative in its own right, in that it can highlight different structural features of great importance—atom proximity, binding pocket geometry, solvent-accessible surface area, overall shape, and so on. As already highlighted, however, the amount of information they can deliver is limited to what can be extracted from the structure alone; to gain further insight, conformational sampling, time correlation and energetics are required, which can only be achieved through simulation and, in turn, rely on a set of interactions. Each of the aforementioned representations thus constitutes the basis for a wide range of models, differing by complexity, accuracy, computational requirements, and so on. These models are employed to investigate the behavior and properties of systems at various level of detail, ranging from all-atom descriptions, where each atom is explicitly accounted for, to very coarse and qualitative pictures where an entire protein is treated as a featureless sphere. Evidently, the choice of a model over another

[3]As it will be detailed in the following sections, filters and SLRs are referred to in the literature as *mappings* and *mapped representations*, respectively.

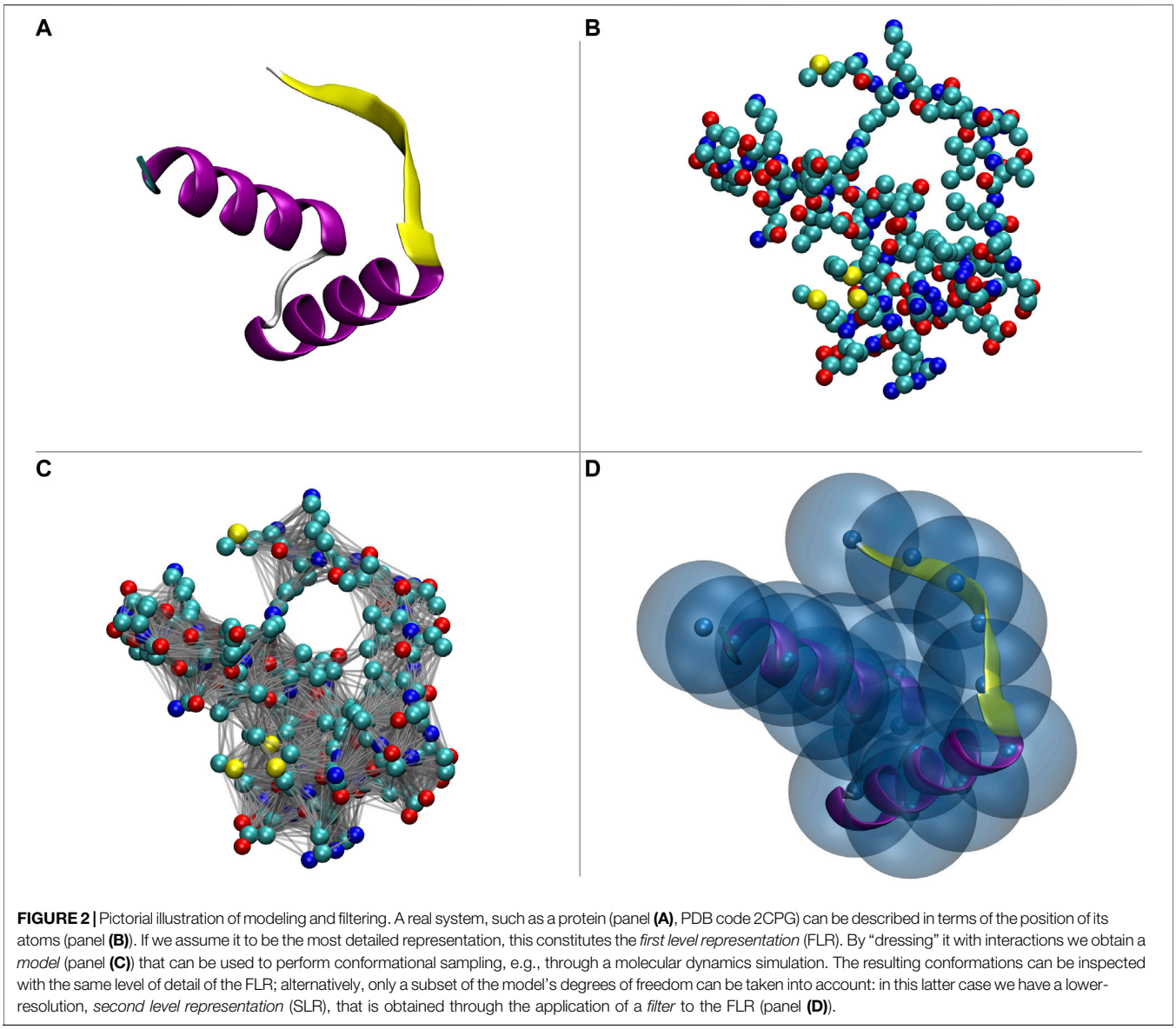


FIGURE 2 | Pictorial illustration of modeling and filtering. A real system, such as a protein (panel **(A)**, PDB code 2CPG) can be described in terms of the position of its atoms (panel **(B)**). If we assume it to be the most detailed representation, this constitutes the *first level representation* (FLR). By "dressing" it with interactions we obtain a *model* (panel **(C)**) that can be used to perform conformational sampling, e.g., through a molecular dynamics simulation. The resulting conformations can be inspected with the same level of detail of the FLR; alternatively, only a subset of the model's degrees of freedom can be taken into account: in this latter case we have a lower-resolution, *second level representation* (SLR), that is obtained through the application of a *filter* to the FLR (panel **(D)**).

depends on the problem one is interested in: the resolution of the model determines the lowest-level, most fundamental causes it can produce, and with it the processes and properties it can generate.

Filters, on the other hand, have not been developed so far with the same intensity as models, in spite of the fact that their usage is ubiquitous. We cannot think of making sense of an all-atom molecular dynamics simulation by examining all $3N$ coordinates in each frame at a time: a process of *synthesis*[4] is necessary in order to extract, from such a large amount of data, the relevant *and intelligible* bit that we can make use of. This process is often carried out quantitatively, e.g., by defining a specific reaction coordinate that allows one to discriminate between two distinct conformers of a molecule. In such a case, one has to know in advance what to look at in order to construct this coordinate; rather frequently, however, a qualitative approach is the first and possibly unique one, and it takes the form of a *visual inspection* of the MD trajectory. Albeit very sophisticated, as it passes through immensely complex neural networks (our brains), the information is eventually reduced to simple notions such as "open" and "closed".

More quantitative examples of filters are available, such as simplified representations of a protein in terms of quasi-rigid domains, where a group of amino acids is treated as a unique block whose internal dynamics is neglected. To determine the structure of these domains (i.e., which amino acids belong to which domain) on the basis of their dynamical properties it is necessary to make use of a model defined at a higher resolution with respect to the blocks themselves; however, once their

[4]In a currently very popular language, one might call this a feature extraction process.

identity is fixed, the trajectory can be studied *filtering out* the movement internal to the blocks and focusing on the relative displacements among them. This procedure of feature extraction enables one to derive, from a large amount of data inherent in the output of a model defined at the fist level of representation, a smaller and more manageable amount of information defined at the second level of representation.

In the following, we will provide an overview of the most common models employed in the computer-aided investigation of proteins, and discuss what impact the level and distribution of the detail of the underlying representation has on the capacity of the model to generate information; subsequently, we will discuss how filters can be employed to rationalize the impressive amount of data produced in a single MD run and separate the signal from the noise.

3 ALL-ATOM MODELING

The models of reference in the computational study of molecular systems are the so-called *all-atom* models. They are defined at an atomic scale resolution, meaning that each atomic nucleus is represented as a material point-like particle.

In all-atom models, each particle interacts with the surrounding ones through classical potentials. This is justified by the Born-Oppenheimer approximation (Born and Oppenheimer, 1927), which allows one to eliminate the electron degrees of freedom by taking the quantum-mechanical expectation value over the electronic wave function, under the assumption of an effective decoupling of nuclear dynamics and electronic ground state. As a result, the interaction energy, which is quantum in nature, can be approximated by a classical potential energy surface that depends only on the position of the atomic nuclei, ignoring the evolution of the electronic distributions. Moreover, interactions in all-atom models are based on the assumption that the contribution of each atom to the Born-Oppenheimer potential energy can be approximated by a sum of few-body terms, each shaped into a simple, empirical and semi-empirical functional form. These energy terms, which collectively take the name of *force field*, can generally be divided in two types: those describing bonded interactions, and those describing interactions between particles close in space but not connected by any chemical bonds (non-bonded interactions). The former are associated to the presence and distortion of chemical bonds, and are modeled as a sum of contributions with a dependence on bond lengths, bond angles, and dihedral angles; non-bonded interactions, instead, are described in terms of Van der Waals, electrostatic, and hydrogen bond potentials.

Force field parameters are obtained from experimental data and quantum-level calculations performed on specific sets of systems. Bond lengths and corresponding stiffness values, as well as angle parameters, are commonly determined from crystallographic or spectroscopic data; Van der Waals terms from small molecules liquid density, heat of evaporation, or solvation free energies; partial atomic charges from quantum-mechanical calculations (González, 2011). As no unique parametrization strategy exists, a plethora of atomistic force fields have been developed through the years, all having a strikingly similar functional form but different coefficients. In the case of proteins, examples of common atomistic force fields are Amber (Maier et al., 2015) and CHARMM (Huang and MacKerell, 2013); recently, improved versions of these force fields for both folded and intrinsically disordered proteins have also been developed (Huang et al., 2017; Robustelli et al., 2018), in addition to force field types designed for amyloid assembly (Nguyen et al., 2021). In any of these force fields, each amino acid type in a defined protonation state is described through the same set of parameters, irrespective of its position along the protein sequence; exceptions are the N- and C-termini, which usually require *ad hoc* parameterisations according to the capping groups. Moreover, particles within each residue are generally not distinguished on the basis of the sole chemical element, but according to the *atom type*. This distinction is much stronger than the one based on the atomic number, since atom types differ also in their hybridization state and the local electronic environment of the atoms they are covalently bonded to. The definition of atom types in a force field is of fundamental importance, since it determines the specificity of the interactions.

Even within the limits of validity imposed by the aforementioned approximations, these models are of tremendous importance to perform an *in silico* exploration of a macromolecule's energy landscape, with the aim of bridging the gap between structure, dynamics, and function. In this regard, the conformational sampling method of choice in the biophysics community is molecular dynamics (MD), through which successive configurations of the system are generated by numerically integrating Newton's equation of motion, thus allowing the calculation of both equilibrium and time-dependent properties.[5]

Atomistic MD has brought a significant progress in a wide range of biological applications in the last decades, due to the advancement of novel algorithms and high-performance computing. The gap between timescales resolved in simulations and in experiments has been significantly reduced due to the concurrent advances in the corresponding techniques; particularly significant is the recent diffusion of graphic processing units (GPUs) for MD calculations (Stone et al., 2010; Lindert et al., 2013; Sweet et al., 2013), and the consequent GPU implementation of popular molecular modeling software packages (Lee et al., 2018; Kutzner et al., 2019; Phillips et al., 2020). Groundbreaking was the development in the last decades of the supercomputer Anton (Shaw et al., 2008; Shaw et al., 2009; Shaw et al., 2014), specifically

[5]Monte Carlo (MC) is another popular simulation technique, which, however, found less space in the all-atom investigation of biomolecular systems; its main disadvantages, when compared to MD, are the inefficiency for exploring the configurational space of large biomolecules, the slowness of the convergence rate, and the lack of information about the time evolution of structural events (Adcock and McCammon, 2006).

designed for running atomistic MD simulations with extremely high efficiency (Pan et al., 2016; Masureel et al., 2018; Pan et al., 2019).

Making use of standard resources, computational scientists can nowadays access micro-to millisecond timescales with atomic detail, for systems comprising several hundreds of thousands of particles. This is sufficient to characterize many critical biological processes, such as ligand-binding events and the folding of small proteins (Kubelka et al., 2004; Freddolino et al., 2008). However, several phenomena are still inaccessible with all-atom MD; this is the case of large-scale structural rearrangements, whose characteristic time scale typically impairs an exhaustive exploration of the accessible conformational space. To alleviate this limitation, diverse enhanced sampling techniques have been developed, including metadynamics and replica exchange MD, which boost the conformational sampling by "helping" the system overcome high free energy barriers; excellent reviews on these topics can be found in Bernardi et al. (2015) and Yang et al. (2019).

The advances in the field of atomistic simulation are paralleled by the increase in the amount of information they generate. MD trajectories, which consist of the set of three Cartesian coordinates per atom per simulation time step, can easily result in an enormous quantity of raw data: appropriate tools are required to separate the most meaningful information buried in the high dimensional space of the simulation output from the rest, thus addressing specific questions about the phenomenon under investigation. Indeed, this challenge led to the development and application of several dimensionality reduction algorithms (Sittel and Stock, 2018; Tribello and Gasparotto, 2019) to the analysis of all-atom MD trajectories. Still, no standard procedure or unique technique exists that can allow the blind and automated determination of the fundamental degrees of freedom of the system. Nowadays, the common approach consists in combining several analysis tools in a system-specific fashion, in order to reduce data complexity and facilitate the understanding. The specific analyses performed are strictly connected to the molecule simulated, to the technique used to generate the dynamics, and to the type of information one is interested in.

In order to investigate the behavior of a protein in terms of its structural stability, it is standard procedure to compute the root mean square deviation (RMSD) and the root mean square fluctuation (RMSF) on the positions of a subset of particles, typically the C_α atoms. The former quantity indicates the global evolution in time of the atomic position, and gives indications on the drift from a given conformation; the latter instead is usually time-averaged on a residue basis, and can help to identify the relative flexibility of protein segments. In addition, secondary structure content can be monitored by applying analysis algorithms such as STRIDE (STRuctural IDEntification) (Frishman and Argos, 1995) or DSSP (Define Secondary Structure of Proteins) (Kabsch and Sander, 1983), which assign a secondary structure conformation to each residue on the basis of the hydrogen bond pattern of its backbone. In the case of the STRIDE method, the secondary structure assignment includes torsion angle potential calculations, as well as statistical propensities extrapolated from experimentally determined structures. The information provided by DSSP and STRIDE is useful to follow the evolution of the secondary structures in time, and eventually to detect changes associated to partial unfolding or disorder-to-order transitions (Lin et al., 2019).

The combination of RMSD, RMSF, and secondary structure analysis can give details on specific regions of the protein that are more stable than others. As an example, in recent works (Spagnolli et al., 2019; Spagnolli et al., 2020), one of us investigated the stability of atomistic models of infectious prion proteins by combining the aforementioned analysis techniques in a synergistic approach. Due to difficulties in the wet lab procedure, experimentally solved structures of prion proteins are still unavailable; therefore, testing the stability of models via MD simulation represents an extremely important step toward the 3D structure elucidation.

If the protein under investigation undergoes large conformational changes, it can be appropriate to group, or cluster, the configurations explored during the simulation on the basis of their structural similarities. To this aim, various clustering methods have been developed in the past 30 years, each presenting different algorithmic characteristics and computational performances (Shao et al., 2007). The 2D RMSD matrix, which includes the deviations between any pair of trajectory frames, typically defines the distance between the conformations to cluster. Based on this measure, a variety of open source tools are available to perform the clustering; among these, it is worth mentioning the widely used MDAnalysis package (Michaud-Agrawal et al., 2011), which employs python libraries to perform the calculations.

Clustering can also be combined with principal component analysis (PCA), as in Wolf and Kirschner (2013) and Wolf and Kirschner (2013), where the two approaches are applied to the trajectory of a bacterial ribosomal domain. The advantages of applying PCA prior clustering analysis lie in the remarkable dimensionality reduction, which results in a simplification of the clustering operation, and in a better visualization of the clusters when plotted in the most represented PCA space.

While clustering allows one to easily identify the variety of conformations sampled during an MD simulation, it can hardly give information on the dynamics of transitions between them. Kinetically relevant states and their rates of interconversion can instead be estimated from Markov state models (Chodera and Noé, 2014). Starting from large sets of individual short MD trajectories, this approach has been used to tackle biological problems happening at relatively long time-scales, such as protein folding, protein-ligand binding, or large conformational changes.

Tools from information theory can also be employed to analyze atomistic trajectories. For instance, cross-correlation or mutual information (Lange and Grubmüller, 2006) can shed light on concerted movements between protein regions; while the former captures only linear correlations between residues, the latter can detect also the non-linear ones. Both cross-correlation and mutual information can be used to build a network representation of the protein, where each residue is defined as

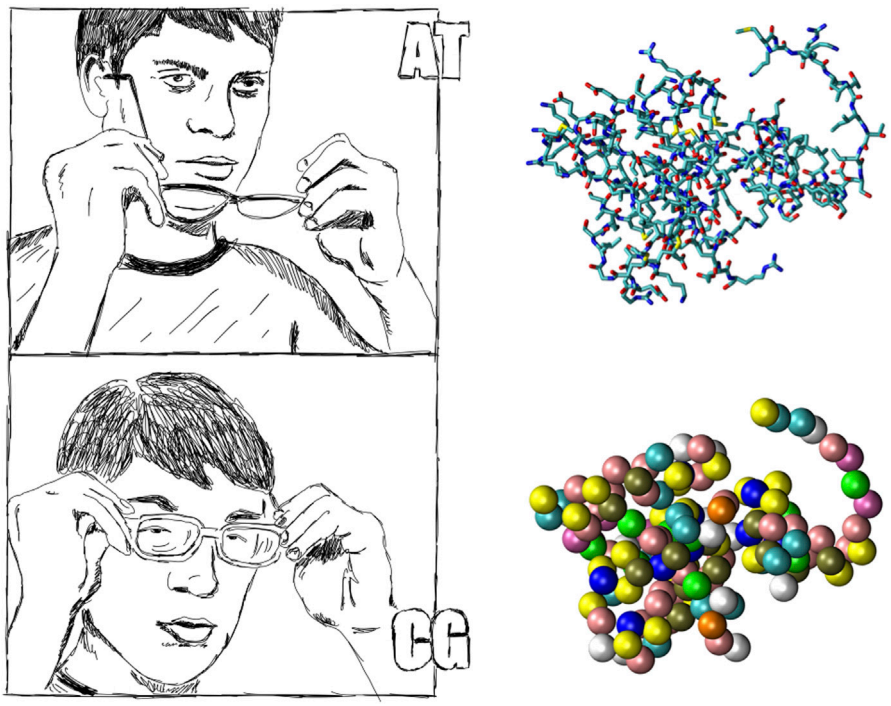


FIGURE 3 | The process of filtering is akin to wearing a pair of spectacles while enjoying a perfect eyesight. The loss of detail is generally considered a defect, however it does have the advantage of simplicity and parsimony - if it is done properly. Artwork by R. Potestio.

a node and the graph edges between neighboring elements are weighted according to the correlations extracted from an MD simulation. The resulting network can be analyzed through well-established techniques of graph theory, such as node centrality and edge betweenness (Böde et al., 2007). These analyses proved valuable in the study of allostery, as described in the work of (Bowerman and Wereszczynski, 2016). Here, the authors simulated at atomistic level the enzyme thrombin, and calculated correlations among residues in terms of cross-correlation, mutual information, and non-linear generalized correlation. The latter is then employed to construct a graph and obtain information about allosteric pathways and hotspots.

In spite of the advancements in simulation tools and analysis techniques, however, the simulation of large macromolecules and slow biophysical processes still remains out of reach for detailed, atomistic models. Additionally, such a high level of detail in the description of the system can even represent a limitation in the comprehension of the system of interest and its properties. To overcome these limitations simpler representations are employed, which offer an increased efficiency at the expenses of a lower degree of resolution. These *coarse-grained models* are the object of the following sections.

4 COARSE-GRAINED MODELING: GENERAL FRAMEWORK

In 1975, Levitt and Warshel published a paper in which they employed an extremely crude representation of a protein in terms of few sites endowed with simple interactions to gain insight in the process of protein folding (Levitt and Warshel, 1975); while the specific results obtained have been later questioned (Hagler and Honig, 1978), this work represents a milestone as the first, pioneering attempt to investigate fundamental biophysical problems making use of minimalistic models of the system instead of extremely accurate ones. Since then, biomolecular CG modeling has steadily grown to become an essential tool in the computational investigation of biological matter: only considering proteins, a whole zoo of CG models and techniques has been developed, which aim at capturing the physicochemical behavior of a large variety of molecules over a wide range of characteristic length and time scales (Saunders and Voth, 2012; Kmiecik et al., 2016; Singh and Li, 2019; Nguyen et al., 2021). Given this extreme diversity, we deem it useful to briefly recapitulate the main concepts underlying the development or choice of a CG model.

As discussed in **Section 2**, the first ingredient required in the construction of a CG model for a biomolecular system is the selection of a SLR, obtained by superimposing a *mapping* to the fundamental representation—in our case, the atomistic one. The mapping constitutes the observational filter connecting the detailed description of the protein's instantaneous configuration to its low resolution counterpart, meaning that, in the latter, only a limited amount of the original degrees of freedom is explicitly employed. One can think of this process as putting on a pair of "coarse-graining glasses" whose effect is that of blurring an already neat and defined image (see **Figure 3**).

Critically, inherent to the CG mapping is the definition of the elemental units composing the newly-introduced representation: in particle-based CG pictures (Noid, 2013; Kmiecik et al., 2016), such units are the effective interaction sites, or "beads", obtained by lumping together subsets of the system's constituent atoms. Depending on the chosen resolution level, each site can be representative of small to medium-sized chemical moieties (Monticelli et al., 2008; Bereau and Deserno, 2009; Darrè et al., 2015), single or groups of amino acids (Clementi et al., 2000; Atilgan et al., 2001; Micheletti et al., 2004; Zhang et al., 2017), up to entire molecular structures (Chu and Voth, 2006; Sept and MacKintosh, 2010; Dama et al., 2013). In all these cases, the mapping is formally expressed as the functional relation $\mathbf{R} = \mathbf{M}(\mathbf{r})$ between the effective sites' coordinates $\mathbf{R}$ and the atomistic ones $\mathbf{r}$ (Noid et al., 2008a; Rudzinski and Noid, 2011).[6] In continuous or quasi-continuous CG representations (Oliver et al., 2013; Welch et al., 2020), the elemental units can instead be identified with the finite volume elements employed to decompose the protein's macroscopic structure, and the mapping can be considered as the specific discretization mesh prescription employed in numerical calculations.

The selection of the level of resolution employed to describe a system is *per se* an already highly nontrivial problem, as this process naturally introduces a lower bound in the length scales the CG representation is in principle able to resolve. Indeed, for a CG observer it will be impossible to capture fluctuations of the system taking place below the size of the average radius of a bead—or the distance between two beads—in particle-based pictures, or smaller than the size of the discretization mesh in continuous ones. In turn, this implies that a CG representation characterized by a specific minimum length scale, when employed to *inspect* the system, can only enable the investigation of emergent properties or phenomena occurring *at or above* such scale: it is technically impossible for the CG filter to grasp the rotation of a specific protein side chain around its main axis, if it is depicted as a point-like particle. This limitation has to be explicitly accounted for when designing the low-resolution representation of a system.

Subsequently, for the simplified representation to acquire predictive power—that is, for the CG mapping to become a *model*—interactions among its effective degrees of freedom must be introduced. In the case of continuous CG representations, this amounts at providing, as input parameters, the appropriate material properties of the protein, e.g., shear viscosity and shear/bulk moduli, which determine the overall stress tensor of the system and consequently its hydrodynamic behavior (Oliver et al., 2013). Particle-based CG models, on the other hand, require the definition of the interaction potential—more precisely, a free energy—acting among the point-like effective sites that constitute the molecular structure (Rudzinski and Noid, 2011; Noid, 2013). As during the last decades substantial effort has been devoted to the development and application of particle-based CG protein models, we here showcase the main approaches behind the parameterization of the associated constitutive interactions. Our objective in this and the following sections is to provide the reader with an idea of the diversity of the available models, their properties, and their applications in a qualitative and non-exhaustive manner; for much more detailed and technical presentations the interested reader is referred to the excellent reviews that have been recently presented in the literature (Noid, 2013; Kmiecik et al., 2016; Singh and Li, 2019; Nguyen et al., 2021).

Depending on the nature of the ingredients employed in the construction of the CG potential, particle-based models are usually divided in three main classes: knowledge-based, top-down, and bottom-up models (Noid, 2013).[7]

In the knowledge-based approach, the parameters of the CG potential are identified through statistical analyses performed over one or more experimentally resolved, static protein structures. To some extent, knowledge-based methods thus directly translate bioinformatic or "frequentist" information about the occurrence of specific local properties—such as side-chain affinities (Tanaka and Scheraga, 1976; Miyazawa and Jernigan, 1996; Bahar and Jernigan, 1997; Davtyan et al., 2012), backbone torsional angles (Betancourt, 2008; Kim et al., 2013), or hydrogen bond capabilities (O'Meara et al., 2015)—to the forces acting among the system's CG effective sites. Top-down models, on the other hand, typically hinge on simple functional forms for the CG potential, the *a priori* choice of which is dictated by physicochemical intuition, and fine-tune their constituent parameters so as to reproduce a set of experimentally-measured meso-to macroscopic observables for the system at hand, including structural and/or thermodynamic ones (Monticelli et al., 2008; Coluzza, 2011; Najafi and Potestio, 2015; Dignon et al., 2018; Perego and Potestio, 2019; Dignon et al., 2019).

The two aforementioned CG strategies do not explicitly rely on the existence of a more fundamental model of the protein, in our case an all-atom force field; the problem of distilling the interactions among CG sites directly from those governing the microscopic constituents, so that the former become emergent properties of the latter, is addressed in bottom-up methodologies.

Bottom-up CG'ing stems from a rigorous statistical mechanics framework in which the high-resolution detail of a system is explicitly integrated out in favor of a lower-resolution representation (Rudzinski and Noid, 2011). This process results in an effective interaction among the CG sites, the potential of mean force (PMF), which in principle provides a complete, *faultless* description of the system as observed through the "CG glasses", see **Figure 3**. The price one pays for the simplification is that the PMF is intrinsically many-body in nature: even if the energetic landscape of the original microscopic system comprises only pair potentials among its

[6]Note that this definition does not straightforwardly apply to the case of adaptive resolution models, *vide infra*.

[7]We stress that this sharp distinction is getting more and more smeared as the field of CG modeling steadily evolves: indeed, complementary ingredients extracted from each of the three aforementioned classes are often combined together in parametrizing the CG potential of a protein.

constituent atoms, once the resolution reduction is performed a whole hierarchy of interactions appears that involve, in addition to pairwise terms, triplets of CG sites, quadruplets, and so on (Dijkstra et al., 1999). These many-body components can play a key role in generating, and comprehend the origin of, the correct large-scale behavior of a system (D'Adamo et al., 2015; Menichetti et al., 2017), such as, in the case of proteins, secondary structure motifs (Kolinski et al., 1993; Derreumaux, 1999; Bereau and Deserno, 2009; Liwo, 2013; Sieradzan et al., 2017); at the same time, however, their presence makes the exact determination of a PMF largely unfeasible in practice, except for very simple microscopic models (Diggins et al., 2018).

The ultimate goal of bottom-up strategies has thus become the construction of increasingly accurate approximations to the correct result, achieved by relying on a wide variety of different theoretical techniques. Approaches exist that allow the explicit calculation of the set of interactions composing the low-resolution potential, including the aforementioned many-body terms, through a systematic decomposition of the PMF in terms of Kubo cluster cumulants (Liwo et al., 2014; Sieradzan et al., 2017; Liwo et al., 2020). Other methods focus instead on reproducing a subset of the system's structural observables (Lyubartsev and Laaksonen, 1995; Soper, 1996; Tschöp et al., 1998; Mullinax and Noid, 2009a; Lyubartsev et al., 2010; Rudzinski and Noid, 2015) or aim at approximating the MB-PMF by means of variational approaches, either directly (Shell, 2008; Shell, 2016), or matching the many-body mean forces—that is, the gradient of the MB-PMF (Izvekov and Voth, 2005; Noid et al., 2008a; Noid et al., 2008b). In the case of structure-based techniques, the implicit assumption is that, if the model generates a set of important properties whose values are quantitatively in line with those of the high-resolution model, this is a sign of CG interactions reasonably approximating the MB-PMF; conversely, if the CG potential optimally reproduces the MB-PMF as in variational approaches, one can expect it to give rise to observables that match the AA ones.

Recently, many of these strategies have benefited from the introduction of machine learning protocols that aim at easing the construction and usage of bottom-up CG force fields; notable examples are, in the case of force-based methods, DeepCG for liquid water (Zhang et al., 2018) and CGnets for small peptides (Wang et al., 2019). Despite having been so far applied to relatively small systems, the promising results obtained suggest that machine learning techniques will soon grow to become a cornerstone in the parameterization of accurate CG potentials for biologically relevant macromolecules.

Irrespective of the parameterization workflow, CG interactions pose nontrivial conceptual challenges. In principle, the selection of a filter with a specific level of resolution allows one to *observe* all phenomena in the system that occur at a length scale equal to or larger than the characteristic size of the elemental CG units; in the construction of a CG model, though, it is the choice of the interactions that limits its ability to *reproduce* such phenomena. If the CG potential accurately reproduces the MB-PMF, all thermodynamical properties and observables of the system can be obtained, even if they originate from processes that take place at a scale below the resolution level of the model (Wagner et al., 2016; Lebold and Noid, 2019a; Lebold and Noid, 2019b; Dannenhoffer-Lafage et al., 2019). However, in practical applications it is not possible to calculate *all* many-body contributions that appear in the PMF, let alone embodying them into computationally manageable functional forms. In the construction of the CG potential one is thus doomed to rely on a limited basis set of interaction terms, commonly consisting in few-body ones, which leaves out high-order contributions; consequently, some effects of the removed DoF's will not appear. *With a limited expansion of the MB-PMF, we thus expect that a reduction in the resolution level will correspond to a decrease in the spectrum of properties and phenomena that the model is able to predict.*

In some occasions it is reasonable to suppose that a particularly well-chosen representation of the system might lead to a substantial simplification of the interactions, e.g., by making many-body terms small or even negligible (D'Adamo et al., 2015): if this were the case, the MB-PMF could be expressed through simple interactions among few constituents, thus making the model simple to parametrize and understand. Alternatively, if the many-body nature of the PMF cannot be reduced or neglected, more complex interactions have to be incorporated, as it is done in the case of density-dependent potentials (Allen and Rutledge, 2008; Sanyal and Shell, 2016; Wagner et al., 2017; Sanyal and Shell, 2018; Rosenberger et al., 2019; DeLyser and Noid, 2019; Shahidi et al., 2020).

It is thus of paramount importance, for a successful usage of coarse-graining methods, to identify which is the simplest model—in terms of representation and interactions—that is capable of accurately reproducing the properties of interest. This problem will be addressed in **Sections 5 and 6** of this work, where we will present an overview of examples taken from the literature. Specifically, in **Section 5** we will focus on the interplay between resolution level and range of observable phenomena: we will discuss how, in general, by decreasing the former we limit the latter, but at the same time we gain access to larger length and time scales. The discussion will be restricted to the case of homogeneous CG representations, that is, models in which roughly the same level of resolution is employed throughout the whole protein structure.

The homogeneity constraint will be subsequently relaxed in **Section 6**, where we will focus on examples of hybrid models in which different levels of resolution are concurrently coupled in the description of the protein and/or the surrounding solvent. This approach is particularly suited for the study of phenomena localized on a well-defined region of the molecular structure: in such a case, the high-resolution level is typically dictated by the characteristic length scale of the phenomenon of interest, while the rest of the system can be described with a coarser, and consequently computationally less expensive, degree of detail.

5 COARSE-GRAINED MODELING: RESOLUTION LEVEL

In **Section 4** we discussed how a reduction in the resolution level of a CG model can in turn lead to a limitation in the amount of

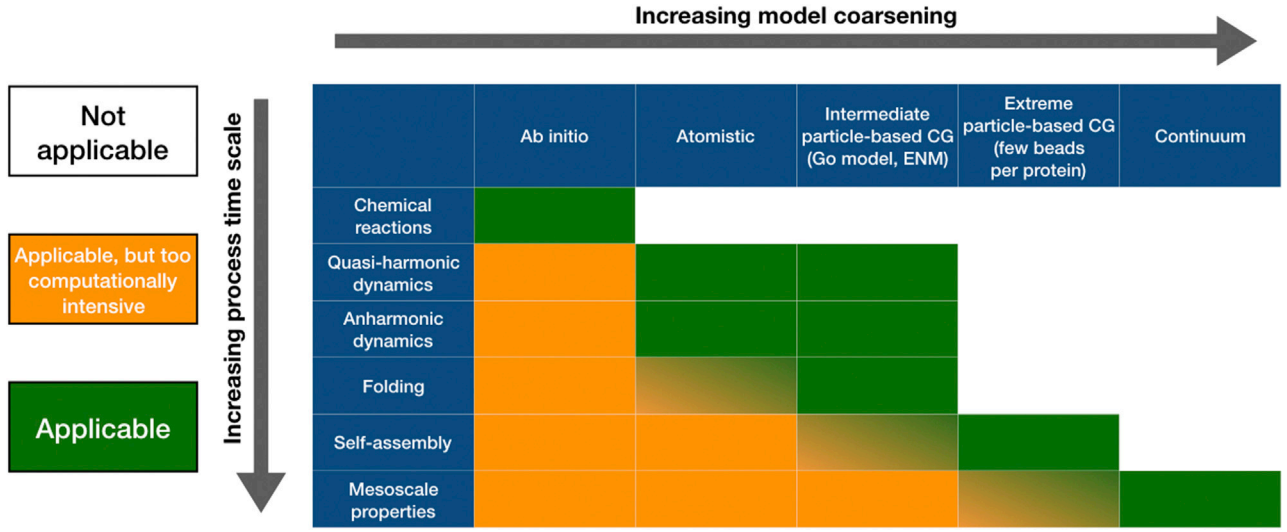


FIGURE 4 | A schematic illustration of the relation between a model's accuracy and its capacity of reproducing long time-scale phenomena. In principle, an extremely accurate model *might* reproduce all phenomena that take place at a characteristic length and time scale that lies above that of its fundamental constituents; however, practical limitations make its usage impossible beyond a certain limit. The coarser the model, the longer the time scale that can be achieved, at the expenses of a shorter and shorter list of processes that it can manage to produce.

emergent properties or phenomena the model is able to reproduce. Critically, this is not a mere consequence of the increase in the model's smallest length scale; it also stems from our incapacity to parametrize CG potentials capable of comprehensively capturing the increasing amount of microscopic detail that gets integrated out.

In this section we will explicitly investigate this tradeoff by relying on a subset of protein CG models extracted from the literature (Saunders and Voth, 2012; Kmiecik et al., 2016; Singh and Li, 2019; Nguyen et al., 2021). We will restrict our analysis to homogeneous models and present them in order of decreasing resolution, moving from the most detailed CG representations all the way down to continuum pictures. For each resolution step, we will mention what phenomena the corresponding model is appropriate to investigate.

Before starting the discussion, however, a couple of remarks are in order. Firstly, we stress that the following list is not exhaustive. The CG models selected in this work serve the only purpose of providing the reader with a representative landscape of the possible descriptions employed to investigate biomolecular systems across a wide range of time and length scales.

Secondly, although a decrease in the resolution of the CG model should in principle correspond to an increase in the characteristic sizes of the analyzed systems and phenomena it can produce—thanks to the reduction in the associated computational cost—in the presented applications this will not always be the case. This is a consequence of our choice of ranking CG models according to their resolution level rather than their chronological appearance: more detailed and thus resource-intensive CG models, when recent, benefited from the modern explosion experienced in accessible computational power, enabling their applications to system sizes that only a few years earlier would have been unconceivable to address at such a high level of detail. The inverse correlation between resolution and accessible time/length scales should thus be interpreted as a trend rather than a rule.

To construct our hierarchy of CG models, in the following we will resort to a division in four main categories that account for similarities in the underlying CG'ing philosophies. Specifically, in Section 5.1 we will discuss explicit solvent CG models, whose elemental units aim at preserving, in a reasonable although approximate manner, the chemical features of the original microscopic components of the protein as well as those of the solvent in which the protein is immersed. Subsequently, in Section 5.2 this relatively high degree of local detail will be significantly reduced through the introduction of implicit solvent CG models. It is in this context that a decrease in resolution, combined with further simplifications in the associated interactions, more severely implies a bottleneck in the landscape of phenomena a specific model can capture. The residue-based decomposition of a protein that is common to both explicit and implicit solvent CG models will be then loosened in Section 5.3, where we will discuss Ultra-CG ones. Here, a single effective site becomes representative of group of residues, a small protein, up to an entire molecular complex. Finally, in Section 5.4 the particle-based CG'ing scheme will be abandoned in favor of protein models in the continuum. A schematic representation of this resolution-based hierarchy of CG force fields, providing information about which emergent phenomena each class of models can provide insights on, is presented in **Figure 4**.

5.1 Explicit Solvent Coarse-Grained Models

The uppermost rung of a hierarchy of CG models arranged in order of decreasing resolution is occupied by particle-based ones

that account for the solvent environment in a simplified *but explicit* manner. Immersed in this CG solvent, an ensemble of beads is then employed to describe a protein, each bead being meant to encapsulate a small chemical moiety comprising few constituent atoms, thus resorting to a rather moderate level of CG'ing. Notable examples in this class of models include the popular SIRAH (Darrè et al., 2015; Machado et al., 2019) and MARTINI (Marrink et al., 2007; Monticelli et al., 2008) force fields, in which, within a relatively "granular" solvent, a quite conspicuous number of effective interaction sites is employed to represent a single amino acid composing the protein structure. Particular attention is further paid to approximately capturing the "local" chemical features and flexibility of amino acid side chains, so that several beads can be employed in their description. Overall, this fairly high level of detail can limit the computational speedup generated by these models, especially due to the presence of the solvent; at the same time, it often allows an almost one-to-one reconstruction, or backmapping, of the microscopic structure starting from the CG one (Darrè et al., 2015).

Interactions among the CG sites are parametrized to account for the average properties of the atoms they enclose, and include bonded as well as non-bonded contributions; in both SIRAH and MARTINI, the former are tailored so as to reproduce (a subset of) structural features, such as bond distances and the bending and dihedral angles between consecutive units. Different philosophies lie instead at the core of the determination of the non-bonded potentials: while SIRAH aims at providing an accurate description of the system electrostatics and sterics (Darrè et al., 2015; Machado et al., 2019), MARTINI mainly targets experimental free energies of partitioning of small chemical fragments between a polar and an apolar phase (Marrink et al., 2007; Monticelli et al., 2008). In both cases, the result of this overall parameterization protocol is a "dictionary" of CG building blocks, one per amino acid, that can be combined together to model the protein structure of interest and investigate its behavior.

The resolution level and chemical specificity characterizing the fundamental units of SIRAH and MARTINI enables their application to the investigation of large-scale conformational and/or thermodynamic properties of a system, as well as to problems in which the local detail, down to a sub-residue level, can play a crucial role on the system emergent phenomena: among these we mention the rearrangement of side chains; hydrogen bonding; and protein-solvent, protein-protein, or protein-substrate interactions. Despite the similar length scales characterizing the elemental units composing the two models, however, already at this limited degree of CG'ing the delicate interplay between resolution level and effective interactions has a considerable impact on the spectrum of observable phenomena. Restricting ourselves to one significant example, SIRAH was shown to be able to preserve the stability of proteins comprising α-helix as well as β-sheet elements in absence of explicit topological biases (Darrè et al., 2015). On the contrary, MARTINI requires secondary structure motifs to be enforced *a priori*, thus preventing its application in studies involving folding or general conformational rearrangements (Monticelli et al., 2008; Poma et al., 2017; Souza et al., 2020). While this limitation is commonly associated to the relatively low resolution at which the protein backbone is treated in MARTINI (one bead per peptide), it should rather be considered a direct consequence of the particular choice in the parametrization of the interactions: in fact, effective models exist that rely on MARTINI-like CG representations and are capable of stabilizing secondary structure elements without introducing ad hoc constraints (Alemani et al., 2010; Spiga et al., 2013).

This crucial difference naturally introduces a distinction in the class of phenomena on which the two models can provide insight. Specifically, SIRAH has been largely applied in the study of problems where structural properties, in combination with the local chemical detail, are pivotal: these include the dynamic behavior of disordered proteins (Ramis et al., 2019), the impact of post-translational modifications on protein structural stability (Garay et al., 2019), and the prediction of protein-protein binding free energies (Patel and Ytreberg, 2018). On the other hand, the parametrization of MARTINI is based on polar/apolar phase partitioning, which makes it particularly suited in the analysis of protein-membrane systems. Applications include the insertion and assembly of membrane proteins and protein-protein complexes in lipid bilayers (Bond and Sansom, 2006; Periole et al., 2012), the investigation of the effect of protein crowding on transmembrane diffusion (Javanainen et al., 2013), and the simulation of proteins in realistic membrane environments (Corradi et al., 2018). Recently, the model was also shown capable of predicting protein-ligand binding affinities with no prior knowledge of binding pockets or pathways (Souza et al., 2020).

5.2 Implicit Solvent Coarse-Grained Models

Explicit solvent CG models are required when, although by relying on a blurred microscope, an attempt of tackling all of the intricacies of a system's local chemical maze is conducted. On the contrary, their level of resolution can be considered excessive when dealing with phenomena that take place at larger length scales, such as protein folding, conformational rearrangements, or self-assembly. Consider for example the case in which a net attraction/repulsion between pairs of amino acids constitutes the driving force of the macroscopic process; for this to emerge from the CG model, a much lower resolution than that of SIRAH or MARTINI might be sufficient, e.g., removing the solvent and describing each amino acid as an effective interaction unit.

In principle, such a procedure should come at the price of introducing a more complex (free-)energetic landscape among the elemental sites to compensate for the additional reduction in detail. This, however, is largely unfeasible in practice, and one typically relies on further approximations, such as the derivation of the low-resolution potential through the truncation of formal statistical mechanics series expansions (Liwo et al., 2014), or its *a priori* definition in terms of extremely simplified functional forms (Derreumaux, 1999; Voegler Smith and Hall, 2001; Bereau and Deserno, 2009; Cheon et al., 2010).

This discussion brings us to the second class of CG models within our hierarchical ladder, that is, *implicit solvent* ones. Here, as the name suggests, the solvent degrees of freedom are

integrated out from the description, and one only accounts for the effect they *on average* exert on the proteins under investigation. Such proteins, on the other hand, are still decomposed in terms of the their constituent residues, albeit in an increasingly simpler form as the structural coarsening progresses. It is in this context that the correlation between resolution level, CG interactions, and range of observable phenomena becomes particularly strong: a decrease in the first is usually not balanced by an increase in the second, which in turn can result in a reduction of the third.

Among implicit solvent CG models, the more detailed ones aim at preserving the "chemical identity" of each amino acid. Since such information is inherently contained in the side chain, this directly translates into the usage of one or more explicit CG beads representing it and accounting for its chemical features, in addition to the effective sites that are employed to describe the peptide backbone. In analogy with the case of the explicit solvent models discussed in Section 5.1, the desired outcome is again a protocol in which the fundamental units embodying each amino acid type can be joined together to assemble the specific system under investigation. Examples of such *intermediate resolution* CG force fields are OPEP (Derreumaux, 1999; Maupetit et al., 2007; Sterpone et al., 2013), the one by Bereau and Deserno (BD) (Bereau and Deserno, 2009), PRIME (Voegler Smith and Hall, 2001; Cheon et al., 2010), AWSEM (Davtyan et al., 2012; Wu et al., 2018), and UNRES (Liwo et al., 2014; Sieradzan et al., 2017; Liwo et al., 2020).

The first model, OPEP, is characterized by a high degree of structural detail (Derreumaux, 1999; Maupetit et al., 2007; Sterpone et al., 2013). All the heavy atoms composing the protein backbone as well as the amide hydrogens are retained as CG sites, while a single bead describes the side chain of each amino acid—except for proline, which is represented by all its heavy atoms. Interactions among these fundamental units are then parametrized via a combination of structural, thermodynamic and knowledge-based approaches, and comprise conventional bonded and nonbonded contributions—e.g., harmonic or Lennard-Jones potentials—as well as terms that account for hydrogen bond capabilities and ion pair interactions. Interestingly, while the original version of the model neglected the solvent degrees of freedom, hydrodynamic interactions were later incorporated in OPEP by coupling it with a Lattice Boltzmann representation of the solvent (Sterpone et al., 2015). As for BD and PRIME, they lean on a similar CG mapping prescription to describe each amino acid, namely three beads for the backbone and one for the associated side chain. Notable differences exist, however, in the derivation of their constitutive interactions. In particular, in analogy with OPEP, BD is again defined in terms of a conventional basis set for the bonded and non-bonded interactions, whose fundamental parameters are tuned by combining structural and knowledge-based protocols (Bereau and Deserno, 2009). In addition to terms accounting for connectivity, steric repulsion, side chain affinities, and hydrogen bond capabilities, the BD force field aims at favoring the correct α/β secondary structure ratio through the presence of additional bonded potentials mimicking dipolar-like interactions that tend to stabilize β-sheet components. Furthermore, BD was later generalized to protein-lipid systems (Bereau et al., 2014). PRIME, on the other hand, resorts to a very crude interaction network in which extremely simplified potentials such as hard-sphere and square-well functions describe steric repulsion and bonding/attractive interactions among the effective sites, respectively (Voegler Smith and Hall, 2001). This choice enables the usage of discontinuous molecular dynamics (Rapaport, 1978; Bellemans et al., 1980), further speeding up simulations. Originally blind to the side chain chemical detail, PRIME was later generalized via a knowledge-based approach so as to capture their specificity (Cheon et al., 2010). In AWSEM, three CG sites, respectively located on the peptide C_α, C_β, and oxygen atoms, are employed to represent a single protein amino acid (Davtyan et al., 2012; Chen et al., 2016; Wu et al., 2018). Bonded potentials among the AWSEM CG units are then complemented with a complex network of nonbonded interactions: these include hydrogen-bonding terms, bioinformatic terms biasing the formation of local structures, [8] nonlocal terms describing contacts—either direct or water/protein-mediated—among distal residues along the sequence, and burial terms that aim at accommodating an amino acid into its preferential environment—e.g., the protein bulk or surface. The corresponding parameters are tuned via a combination of structural and knowledge-based approaches. AWSEM further enables the simulation of membrane proteins by relying on an implicit membrane potential (Kim et al., 2014). Finally, UNRES maps each amino acid onto three CG sites, namely the C_α atom, the center of the peptide bond, and the side chain, the latter being described as an ellipsoid of revolution (Liwo et al., 2014). Only the last two elements, however, are explicit effective interaction sites, while the C_α sites only serve the purpose of tracing the protein geometry. Interactions among the UNRES building blocks are then parametrized through a rigorous bottom-up procedure: the potential of mean force of the system is expanded in a truncated series of Kubo-cluster cumulants, which enable the derivation of the multi-body interactions acting among the CG sites in a systematic manner (Liwo et al., 2014; Sieradzan et al., 2017; Liwo et al., 2020).

The computational speedup generated by OPEP, BD, PRIME, AWSEM and UNRES enables their application to problems whose characteristic time and length scales were, until recently, prohibitively large to be effortlessly accessed by conventional all-atom simulations. Specifically, OPEP was largely employed in the context of folding and structure prediction of isolated proteins, protein-ligand and protein-protein complexes (Wei et al., 2004; Shen et al., 2014; Lamiable et al., 2016; Kynast et al., 2016), in aggregation studies (Lu et al., 2012; Nasica-Labouze and Mousseau, 2012), as well as to investigate the structure of long, intrinsically disordered amyloid monomers (Nguyen and Derreumaux, 2020). Moreover, the introduction in OPEP of a Lattice Boltzmann solvent paved the way for its exploitation to

[8]We exclude from the discussion the AAWSEM force field developed by Wolynes and coworkers (Chen et al., 2016; Chen et al., 2017) in which the local structure biasing terms are obtained through explicit all-atom MD simulations of fragments of the protein under investigation, thus rendering the model non transferable.

analyze hydrodynamic effects on biomolecular systems, including the behavior of proteins under shear flow (Sterpone et al., 2018) or the impact of molecular crowding on the dynamics of protein suspensions (Sterpone et al., 2014). Applications of the BD model encompass the investigation of folding processes, including the analysis of the interplay between secondary and tertiary structures in cooperative folding (Bereau et al., 2010), as well as peptide aggregation phenomena (Bereau and Deserno, 2009). Given the additional computational gain provided by its discontinuous potentials, PRIME was instead extensively exploited to investigate the behavior of large-scale systems, especially in the context of aggregation of fibrils in presence or absence of fibrillation seeds or inhibitors (Nguyen and Hall, 2004; Cheon et al., 2011; Wang et al., 2017; Wang and Hall, 2018). In addition to protein folding (Jin et al., 2020), applications of AWSEM include the investigation of protein-protein association (Zheng et al., 2012) and fibrillar aggregation processes (Zheng et al., 2016; Chen et al., 2020), as well as the analysis of the static and dynamic behavior of intrinsically disordered proteins (Wu et al., 2018; Lin et al., 2019). The incorporation of an implicit membrane potential in AWSEM enabled it to provide insight on the folding behavior of transmembrane proteins (Lu et al., 2018) and protein assemblies (Truong et al., 2015). Finally, while UNRES was originally applied to perform protein structure prediction via energy minimization (Liwo et al., 1999), subsequent MD-based studies include the investigation of folding processes (Liwo et al., 2005), self-assembly of protein complexes (Sieradzan et al., 2012), fibrillar aggregation (Rojas et al., 2010; Rojas et al., 2017), as well as conformational transitions in molecular chaperones (Gołaś et al., 2012).

The power of the intermediate resolution CG models lies in their transferability, that is, the possibility of employing them to provide insight on the behavior of systems that were not directly involved in the models' parameterization. It follows that particular care must be taken as far as meso-to macroscopic properties are concerned; while these can be explicitly included in the construction of the effective potential, for the latter to be transferable the introduced restraints should be flexible enough so as not to bias the model predictions toward very specific outcomes, associated to particular systems. This requirement is especially evident in the case knowledge-based approaches, in which abstraction of the interaction parameters from the stable conformation of a specific protein—conformation that is here interpreted as the emergent property participating in the parameterization of the CG potential—is achieved by performing a statistical analysis over an ensemble of structures (Tanaka and Scheraga, 1976; Miyazawa and Jernigan, 1996; Bahar and Jernigan, 1997; Betancourt, 2008; Kim et al., 2013; O'Meara et al., 2015). It is thus possible, and indeed often advantageous, to design transferable implicit solvent CG models tackling well-defined large-scale problems; at the same time, one should make their constitutive ingredients as general as possible, so as to enable the characteristic phenomenon of the system of interest to arise from the model, without the need of imposing it a priori. On the other hand, one might need implicit solvent CG models that are more severely bound to a subset of known macroscopic properties associated to a specific biomolecule. In this case, the model could be asked, e.g., to reproduce the experimentally resolved tertiary structure of a particular system. The emergent property now directly represents an input of the CG'ing protocol.

One could clearly resort to standard CG'ing strategies and develop a dedicated effective model in which these conditions are satisfied (Izvekov and Voth, 2005; Rudzinski and Noid, 2011; Shell, 2016); this often lengthy parameterization procedure, however, should at least in principle be repeated from the ground up every time a new system is investigated, for which the same kind of external piece of information is available. It is therefore highly desirable to construct CG models that rely on more "intuitive" interaction potentials and are easily generalizable to arbitrary systems through a minimal fine-tuning. The particular choice of the phenomenological potential will play a pivotal role in defining the class of phenomena the model can *additionally* provide insight on. The simplification of the interaction network typically goes on par with an additional reduction in the resolution level and chemical detail, with every amino acid composing the molecule being now described as a single interaction site.

A notable example of this second class of implicit solvent CG models is represented by *structure-based* ones, such as Gō-like models (GLM) (Hills and Brooks, 2009; Takada, 2019) or elastic network models (ENM) (Sanejouand, 2013; Togashi and Flechsig, 2018). Here, the external macroscopic input involved in the construction of the effective CG potential is the static, either stable or metastable, three-dimensional spatial conformation assumed by the protein of interest. Both GLM and ENM describe the interaction among the elemental CG units in terms of very general functional forms, tailored to *reproduce* the target structure but easily applicable to arbitrary ones; the complexity and richness of the basis set, however, significantly decreases while moving from GLMs to ENMs, generating a crucial impact on the spectra of phenomena these two classes of models can respectively capture.

Gō models originally represented a protein as a self-avoiding walk on a lattice (Taketomi et al., 1975). Large-scale structural information enters GLMs through attractive interactions occurring between pairs of sites that, although distant along the protein sequence, are in direct contact in the native conformation. Despite this extremely crude description, such models are capable of driving a protein to spontaneously fold toward its native state (Go, 1983). The lattice formulation was later extended to the continuum enabling the use of MD simulations (Clementi et al., 2000). Here, a protein is represented by sites located on its C_α atoms and interacting via simple potentials whose functional form is borrowed from standard all-atom force fields. The folded conformation is enforced in both bonded and non-bonded interaction terms: the former are parametrized by setting the equilibrium structural parameters equal to the distances and bond/dihedral angles of the protein native state; non-bonded contributions are instead conceptually akin to the lattice version of the model, so that general, unspecific attractive (resp. repulsive) interactions occur between residues that form (resp. do not form) a native contact. In both cases the strength of the interaction parameters is

independent of the residues' chemical detail; this condition, together with the C_α mapping prescription, was later relaxed in subsequent generalizations, which relied on more chemically-realistic functional forms for the interactions (Karanicolas and Brooks, 2002), as well as quasi-atomistic descriptions of the biomolecule (Whitford et al., 2009).

Due to their extreme simplicity and flexibility, GLMs have a long and successful history in the field of protein folding (Hills and Brooks, 2009; Hu et al., 2017; Takada, 2019). Furthermore, the original native-centric standpoint was later extended to account for the presence of multiple (meta)stable conformational basins, allowing transitions between them at greatly reduced computational cost (Best et al., 2005; Okazaki et al., 2006). Applications in this context range from the investigation of conformational rearrangements of "simple" proteins (Lu and Wang, 2008), all the way up to, e.g., large-scale molecular motors (Hyeon et al., 2006; Kanada et al., 2013).

Sticking to a structure-based CG'ing protocol but further reducing the complexity of the interaction basis set one encounters elastic network models (Sanejouand, 2013; Togashi and Flechsig, 2018). ENMs stem from the pioneering observation, made by Monique Tirion (Tirion, 1996), that the low-frequency dynamics of globular proteins, *in the vicinity of their native conformation*, can be accurately reproduced by replacing the system's complex interaction network by a set of Hookean springs of *equal strength* connecting neighboring atoms up to a given cutoff distance. CG equivalents of this original version of the model have been subsequently developed, which typically retain one or two atoms per amino acid (Atilgan et al., 2001; Micheletti et al., 2004). Structural information is more strictly enforced in ENMs compared to GLMs, preventing the study of processes such as folding or tertiary structure rearrangements. As for the latter, however, it was shown that ENMs are able to capture at least the essential, early-stage behavior of a protein's conformational changes, further cementing their role as a fundamental building block in the edifice of mesoscopic CG modeling of biomolecules (Tama and Sanejouand, 2001; Petrone and Pande, 2006). Moving away from protein simulations, the simplicity of ENMs allowed their application to the investigation of the low-energy fluctuations of complex systems where all-atom as well intermediate resolution CG models would prove computationally too demanding, ranging from macromolecular motors such as ribosomes (Tama et al., 2003) up to entire viral capsids (Tama and Brooks, 2005; Grime et al., 2016).

5.3 Ultra Coarse-Grained Models

The class of models presented in Sections 5.1 and 5.2, although characterized by a gradual decrease in the level of detail, always rely on a residue-based decomposition of a protein, in which only one or few effective interaction centroids describe *each* amino acid composing the biomolecule. To push the applicability of particle-based CG models to the investigation of phenomena occurring at even larger time and length scales, one possibility is that of resorting to ultra coarse-graining (UCG) methods. Here, each CG site becomes representative of larger chemical entities, be that few residues, whole proteins or even entire molecular complexes (Chu and Voth, 2006; Sept and MacKintosh, 2010; Zhang et al., 2017). Several examples of UCG models, ranging from more "chemically accurate" to more heuristic ones, have been presented in the literature. While more traditional applications typically focus on single proteins (Zhang et al., 2017; Zhang et al., 2020), UCG methods have provided impressive insights into the behavior of overwhelmingly complicated macromolecular structures (Saunders and Voth, 2012; Hagan and Zandi, 2016), including actin filaments (Chu and Voth, 2006), bacterial flagella (Arkhipov et al., 2006a), and viral capsids (Arkhipov et al., 2006b; Nguyen et al., 2009; Grime et al., 2016).

As pointed out in Dama et al. (2013), from a conceptual point of view UCG models pose notable additional challenges compared to their more detailed counterparts, which are, as it is the case for the previously discussed studies, often overlooked in the construction of the UCG effective interaction potential of a system. Specifically, as the structural coarsening progresses, several internal states of the system can end up being mapped onto the same CG configuration. For instance, let us consider the case of a macromolecular complex, a whole protein of which is represented as a single UCG site. If the protein undergoes a conformational rearrangement between two states that leave the CG site coordinates unaltered, both states contribute to the energetic landscape of a single CG macrostate and, as far as the model is concerned, they are indistinguishable. At the same time, the rearrangement could play a key role in the generation of the macroscopic phenomenon of interest, and it would thus be desirable to construct a UCG model able to discriminate the two conformational basins. To tackle the problem of constructing CG models for systems possessing internal states, Voth and coworkers have recently developed an extremely elegant Theory of Ultra Coarse-Graining (UCGT) in a series of works (Dama et al., 2013; Davtyan et al., 2014; Dama et al., 2017), to which we refer the interested reader. While applications of this theory have been, to our knowledge, so far limited to relatively high-resolution CG representations of liquids, UCGT represents an extremely promising framework for the development of accurate UCG models of biologically relevant macromolecules.

5.4 Continuous Models

Particle-based CG models share the fundamental common feature of tracking the dynamics of a system through each of its mesoscopic constituent degrees of freedom, or effective interaction sites. As an extreme act of coarse-graining, such a scheme can be completely abandoned in favor of representations that treat the whole macromolecular body as a continuous medium subject to the laws of hydrodynamics.

In this perspective, starting from the observation that a protein in its folded, globular conformation behaves as a viscoelastic solid (Wang and Zocchi, 2011), Harris *et al.* introduced the Fluctuating Finite Elements Analysis (FFEA) scheme for macromolecular simulations (Oliver et al., 2013). In FFEA, fluctuations of a biomolecule around its native conformation are described through hydrodynamic observables, with the evolution of the system in response to stress being obtained by means of finite element analysis.

Notably, in addition to elastic and viscous factors, the effect of thermal noise is directly incorporated into the protocol by means of an appropriate stress tensor. The absence of an atomistic level of detail clearly sets a dramatically large lower bound to the length scales achievable by FFEA; on the other hand, this method represents a promising opportunity for pushing the analysis of biological systems to truly meso-to macroscopic scales. Originally applied to the prediction of the dynamics of globular proteins close to their native states (Oliver et al., 2013), FFEA was later employed to analyze the behavior of complex macromolecular systems such as molecular chaperones (Solernou et al., 2018), conformational transition of molecular motors (Richardson et al., 2014; Hanson et al., 2015; Richardson et al., 2020; Hanson et al., 2021), and the effect of the application of stretching and torsional forces on the structural stability of antibodies (van der Heijden et al., 2020).

A straightforward application of FFEA to the case of highly elongated biomolecules such as thin, rod-like structures is difficult because of the variety of length scales characterizing the conformational variability of these systems. To tackle the problem, Welch *et al.* recently introduced the Kirchhoff biological rod algorithm (KOBRA), a fluctuating rod model designed to perform continuum simulation of slender molecules (Welch et al., 2020). In KOBRA, a thin system is represented as an elastic material curve subject to thermal noise, whose dynamic equations of motion are solved based on a discretization in terms of straight rods connecting a set of nodes. The first application of the model on elongated protein complexes showed promising results; furthermore, the coupling of KOBRA with FFEA suggests the possibility of generating mesoscopic, continuous models of biomolecular systems comprising globular as well as slender components.

6 COARSE-GRAINED MODELING: RESOLUTION DISTRIBUTION

The first historical applications of hybrid multiscale models of biomolecules trace back to the 1970s, with the works of Warshel and Karplus (1972) and, a few years later, of Warshel and Levitt (1976). These works, coupling a quantum mechanical and a classical description, led the foundation for the quantum mechanics/molecular mechanics (QM/MM) methodologies (Amaro and Mulholland, 2018; Magalhães et al., 2020), whose relevance was recognized by the attribution of the Nobel prize in Chemistry to Karplus, Warshel, and Levitt in 2013. The development of QM/MM approaches opened the way for the coupling of lower resolution levels for the investigation of phenomena happening at increasingly larger length scales. In this section, we present examples of such coupling schemes and their range of applications. These include processes of ligand binding studied with hybrid atomistic/coarse-grained resolutions, or protein conformational changes reproduced by the integration of CG scales at different levels of detail. In addition, examples of a dual description of the solvent (atomistic/CG or atomistic/continuum) are reported: they allow the construction of a larger simulation box, representing a computationally efficient solution for finite-size effects. Importantly, all these cases require the definition of the resolution domains during the phase of simulation set-up, on the basis of some previous knowledge of the system. This issue is overcome by the use of coarse-graining as an informative tool, as explained in **Section 7**.

6.1 Coupling Quantum Mechanical–Classical Atomistic Models

In QM/MM, a computationally expensive quantum mechanical approach is used to simulate only a subset of atoms, where a classical force field may fail: a typical example is the active site of an enzyme, where a chemical reaction takes place. The other components of the system, including the rest of the protein and the solvent, are treated in a less computation-intensive way by molecular mechanics. The classical, atomistic description of the largest part of the system allows the simulation of full proteins in their natural environment, either the solvent or the lipid bilayer; however, the time-consuming quantum mechanical calculations—even though restricted to a small number of residues—limit the time scale spanned, which typically covers a few hundreds of picoseconds.

Despite this limitation, which can nonetheless be alleviated by the application of enhanced sampling techniques (Yang et al., 2019), the QM/MM method is having an increasing impact on the study of biomolecules (Lonsdale et al., 2013; Lonsdale et al., 2014; Tyzack et al., 2016), mostly enzyme-ligand complexes. For instance, QM/MM simulations proved useful to compute binding free-energy profiles and barriers for enzyme-catalyzed reactions (Barnes et al., 2013), and to characterize binding kinetics (Haldar et al., 2018). Moreover, QM/MM plays a key role in drug design for the discovery of covalent inhibitors (Lodola et al., 2008; Ranaghan et al., 2014), small organic molecules that steadily inactivate the target protein by forming a covalent bond.

Although enzymatic reactions have been the primary target of QM/MM studies, the approach proved to be effective also for the investigation of proton transfer events, where the excess positive charge is propagated through a network of hydrogen bonds dynamically connecting water molecules, protein residues, and/or cofactors. Since this process involves breaking and forming covalent bonds and charge delocalization, a QM description is required at least in the region where the transfer takes place. Recent applications include the study of ion channels, such as the Cl^-H^+antiporter ClC-ec1 (Chiariello et al., 2020). Here, DFT-based QM/MM simulations and well-tempered metadynamics (Barducci et al., 2008) free energy calculations were performed, contributing to explain the transport inhibition in ClC anion/proton exchangers. Another less obvious field of application of QM/MM simulations is the computational study of metallodrugs—namely coordination and organometallic complexes, typically containing platinum, silver, gold, vanadium, or iron ions (Palermo et al., 2016). The variety of coordination modes, bond breaking and formation, ligand exchange reactions, charge-transfer, and polarization effects in

these molecules requires a QM description of the drug and its binding site on the biomolecular target. A widely studied case is the mechanism of action of cysplatin, one of the most effective and broadly used chemotherapeutic agents (Calandrini et al., 2015).

A recent methodological progress in QM/MM simulations is the development of a Hamiltonian adaptive multiscale scheme (Boereboom et al., 2016). As solvent molecules diffuse in and out of the reactive region, they are gradually included into (and excluded from) the QM computation. This was later implemented along with state-of-the-art path integral simulation techniques, which allow for the calculation of quantum statistical properties, and ring-polymer and centroid molecular dynamics, which allow the calculation of approximate quantum dynamical properties (Kreis et al., 2017). In another definition of QM/MM adaptive scheme, the boundaries of the QM region change during the simulation, adapting themselves to the reaction site (Mones et al., 2015). These advancements pave the way for further methodological developments in the QM/MM field.

6.2 Coupling Atomistic–Coarse-Grained Models and Beyond

Biological phenomena do not always involve the breaking/formation of chemical bonds, which require an accurate description of the electronic structure. It is the case, for instance, of non-covalent protein-protein and ligand-protein interactions, including the vast majority of drug discovery applications, where the designed drug is not supposed to undergo any chemical reaction once accommodated in the protein binding site. If the domains involved in the interaction are known in advance, e.g., from experimental evidence or previous computational analysis, one can additionally exploit the inherently multiscale nature of the problem to build a hybrid atomistic/coarse-grained (AA/CG) set-up, where the atomistic detail is retained only in the region of interest (in the above example, the binding site of a receptor). The rest of the macromolecule is instead treated at a lower, coarse-grained resolution, bringing the immediate advantage of a reduced computational cost.

This general idea gave rise to a variety of approaches, where the details of each method (namely, the resolution distribution and the parameterization of interactions) are specifically designed to tackle the system under investigation. Examples range from the multi-resolution model of a polyamide melt (Gowers and Carbone, 2015), where only the amide groups involved in the formation of the hydrogen bonds are maintained at atomistic resolution, to multimeric complexes including both proteins and nucleic acids, as in Villa et al. (2004), Villa et al. (2005), and Dans et al. (2010). In the latter case, a multi-resolution simulation of the lac repressor protein from *E. coli* and a 107-bp-long DNA segment is performed, where the protein and the two bound operators are described atomistically, while the DNA loop is modeled as an elastic ribbon connecting the terminal base pairs of the DNA operators.

In most of AA/CG applications the size of the atomistic region is larger than a single chemical moiety, but substantially smaller than the protein itself. This is the case of ligand-binding multiscale studies, where an atomistic resolution is required for only a few protein residues. In the work by Fogarty and coworkers (Fogarty et al., 2016), an ENM representation of the hen egg-white lysozyme is coupled with an atomistic description of the active site, with and without the inhibitor di-N-acetylchitotriose. The same model has been employed by Fiorentini and coworkers (Fiorentini et al., 2020) with the aim of assessing the accuracy of a hybrid AA/CG description of the protein for binding free energy calculations.

A hybrid method specifically designed for the study of ligand-protein interactions is the so-called Molecular Mechanics/Coarse-Grained approach (MM/CG) (Neri et al., 2005; Neri et al., 2006; Leguèbe et al., 2012; Tarenzi et al., 2019). In its first version (Neri et al., 2005), MM/CG is validated on cytoplasmic enzymes, whose catalytic site is represented atomistically, while the rest of the protein is described at a CG resolution according to a Gō-like model. Despite the absence of explicit solvent, the method showed a good agreement between the RMSF of the MM/CG simulations and the fully atomistic ones, and a good overlap between the subspaces of the most relevant eigenvectors computed with MM/CG and atomistic MD.

The MM/CG method was then applied to membrane receptors of pharmacological relevance. In particular, with the introduction of a surface potential surrounding the transmembrane region of the protein and mimicking the interaction with the lipid bilayer (Leguèbe et al., 2012), the MM/CG was specifically tailored for predicting binding poses in low-resolution models of membrane proteins, such as homology models of G-protein-coupled receptors (GPCRs). The paucity of experimental structural information and the low sequence identity between members of the family lead to models with inaccurate side chain orientations, which may introduce biases in fully atomistic simulations: in such cases, coarse-graining part of the receptor allows atomistic residues in the binding site to relax more easily to the biologically functional conformation. Atomistic water molecules in the extracellular side, hydrating the binding site, are confined by a repulsive potential. This approach has been widely tested on bitter taste receptor GPCRs (Schneider et al., 2018; Fierro et al., 2019), and recently implemented in a webserver pipeline (Schneider et al., 2020).

In the latest implementation of the method (Open Boundary MM/CG) (Tarenzi et al., 2019), the multi-resolution model of the protein is coupled to an adaptive resolution description of the solvent through the Hamiltonian adaptive resolution (H-AdResS) scheme (Potestio et al., 2013) (see Section 6.3 for further details). Water is modeled with atomistic accuracy in the two hemispheres capping the intracellular and extracellular parts of the receptor, and free diffusion is ensured with a surrounding reservoir of CG water molecules. The improved hydration model leads to the simulation of a rigorous statistical ensemble and enables accurate binding free energy calculations for a drug design purpose (Korshunova and Carloni, 2021).

We conclude this subsection mentioning multi-resolution models where the two or more resolutions concurrently

employed are coarse-grained, that is, lower than atomistic. These approaches aim at reproducing the large-scale conformational dynamics of large biomolecules in a particularly efficient manner, and are especially easy to rationalize. Proteins have been modeled as networks of a small number of CG sites, fewer than the total number of residues (Doruker et al., 2002; Kurkcuoglu et al., 2004; Eom et al., 2007), and unevenly distributed along the primary structure. The partitioning among resolution levels can be performed on the basis of previous knowledge of the working of the system functions: this is the case of the multiscale network model (Jang et al., 2009): here, the fine-grained region is constituted by specific functional sites represented at the residue level as an ENM; the remaining regions are described at a lower resolution, including only a subset of the C_α atoms as interaction sites.

In other approaches, the choice of the level of resolution and its distribution along the protein structure is not so obvious. This is the case of the essential dynamics coarse-graining (ED-CG) (Zhang et al., 2008; Zhang et al., 2009), where residues undergoing collective dynamics are represented by pseudo-nodal points. Such CG sites are determined through a variational approach, with the objective of reproducing the protein's essential dynamics.

This last example illustrates in a rather clear manner the relationship between the filter or mapping on one hand, and the resulting model on the other. The definition of the CG sites is not determined by their chemical structure or identity (as it is the case for residue-to-bead mappings), but rather it is a consequence of their *emergent properties*, such as the internal flexibility. This, in turn, implies a non-uniform assignment of atoms to beads, as different parts of the protein can show different degrees of a given property, so that each CG site represents an arbitrary number of atoms—or, alternatively, the resolution of the model varies non-uniformly along the structure in terms of mass, number of atoms, or chemical identity. More importantly, the mapping is not assigned by the modeler from the top down: it is identified by the system itself as the solution to a minimization process. This is to say, a cost function is defined whose argument is the mapping and whose minimum is the *optimal* mapping. The idea that the system "informs the modeler" about which representation of itself is the most appropriate (given certain criteria) represents a crucial step forward in the process of modeling, and bears important consequences on the interpretation of its outcomes. **Section 7** of this paper is devoted to exploring these ideas.

6.3 Multiscale Schemes Tailored for Solvent Description

We conclude this section with an overview of those multi-resolution approaches that have been applied to liquids and diffusive systems, rather than to bonded structures whose parts have a fixed resolution.

Needless to say, the most important liquid in biology is water. Water molecules often play a direct role in biological processes, such as ligand binding and enzymatic catalysis, by establishing stable non-bonded interactions with protein residues and/or ligands. At the same time, bulk solvent undoubtedly represents the computationally most expensive component of the simulation box. Several multiscale schemes have been designed in order to tackle this duality; they are based on the common idea that water in the hydration shell of a biomolecule requires an atomistic description, while bulk water can be described at a coarser resolution. However, such schemes pose the problem of enabling proper diffusion of solvent molecules across regions at different resolution, while keeping the overall thermodynamic equilibrium under control. This issue is tackled, e.g., in Szklarczyk et al. (2015) through the so-called "flexible boundaries for multiresolution solvation" (FBMS). Here, the spatial partitioning between atomistic and coarse-grained solvent is enforced by means of half-harmonic distance restraints, which attract atomistic molecules to the surface of the solute and repel the CG beads. A restraint-free region at intermediate distances enables the formation of a buffer layer, where the atomistic and CG solvents can mix freely.

An alternative is given by those methods that allow solvent molecules to smoothly change their resolution on the fly when transitioning between an atomistic region and a CG region; these include the adaptive resolution scheme (AdResS) (Praprotnik et al., 2005) and the Hamiltonian adaptive resolution scheme (H-AdResS) (Potestio et al., 2013). In both cases, solvent molecules are free to diffuse between regions at different resolution, without constraints; in so doing, they pass through a hybrid resolution layer, where interactions between molecules are governed by an interpolation of atomistic and CG forces (in case of AdResS) or potentials (in the case of H-AdResS). The interpolation scheme is defined by a position-dependent transition function, which smoothly couples the two domains. Moreover, tailored correction forces can be automatically calculated and applied to the molecules in the hybrid region, in order to ensure a uniform density profile across the simulation box.

The relevance of such approaches in the context of biomolecular simulations has been assessed by studying ubiquitin at fully atomistic resolution in a multi-resolution AdResS solvent (Fogarty et al., 2015), and atomistic proteins atox1 and cyclophilin J in an H-AdResS solvent (Tarenzi et al., 2017). In both works, each CG water molecule is represented as a single bead located on the molecule's center of mass. CG particles interact through a potential derived from Iterative Boltzmann Inversion (Reith et al., 2003; Rosenberger et al., 2016), which reproduces the centre-of-mass radial distribution function of the atomistic solvent; the protein is placed at the center of the atomistic region, which is shaped as a sphere. A study on the effect of the high-resolution region radius on the solute and the hydration solvent was also performed. In Kreis et al. (2016), a similar approach is employed, however, the shape of the high resolution region is self-adjusting during the course of the simulation, following the conformational changes of an atomistic polypeptide during folding. The adaptive resolution representation of the solvent served also for the calculation of solvation free energies of side-chain analogues, using the AdResS (Fiorentini et al., 2017) or H-AdResS (Korshunova and Carloni, 2021) scheme. Further applications of adaptive resolution simulation methods include the coupling of atomistic water

with a supramolecular, MARTINI-style model (Zavadlav et al., 2019), or even an ideal gas representation (Kreis et al., 2015), which can also provide an innovative solution for solvation free energy calculation, by pulling the solute from the atomistic solvent region to the CG one (Heidari et al., 2019).

A natural evolution of particle-based multiscale approaches toward even coarser resolutions is the coupling of atomistic solvent and continuum representations, which aims at extending the range of applicability of such models to system sizes beyond those reachable by particle-based models alone. Several examples exist of simulation schemes including atomistic and continuum descriptions for the simulations of water (Brünger et al., 1984; Beglov and Roux, 1994; Im et al., 2001; Lee et al., 2004; Deng and Roux, 2008; Wagoner and Pande, 2011; Wagoner and Pande, 2013; Petsev et al., 2015). Attempts of triple-scale simulation of liquid water have also been performed, by concurrently coupling atomistic, CG, and continuum models (Delgado-Buscalioni et al., 2009; Zavadlav et al., 2018). Applications of an atomistic/continuum representation of the solvent for biomolecular studies have been performed in Wagoner and Pande (2018), where the boundary between the explicit/continuum solvent models can adapt itself in response to the conformational fluctuations of the atomistic peptide simulated; and in Hu et al. (2019), for a multi-resolution simulation of protein diffusion in water under a steady shear flow.

7 ON CHOOSING THE OPTIMAL RESOLUTION LEVEL AND DISTRIBUTION, AND ON MODELING AS AN ANALYSIS TOOL

In the previous sections we showed how coarse-graining techniques model soft matter systems, proteins in particular, using a plethora of simplified representations, each one characterized by its level of detail. In addition, several methods have been developed to concurrently employ, in the same simulation setup, models at different resolution, so as to provide a small subregion with an accurate description and the remainder of the system with a computationally efficient one. In both cases, the level of detail and its distribution is usually determined *a priori* on the basis of various characteristics (chemical identity, biological function, intuition), depending on the usage one does of the model. Recently, however, interest has grown around the idea of allowing the system itself to decide the "best" coarse-grained description of it. Clearly, the notion of "best" is relative, and it necessarily has to answer to the question *best for what?*

In this final section we report on the recent attempts to find the optimal resolution of a biomolecule, namely the "most appropriate" number and selection of degrees of freedom to describe it, together with their spatial distribution. These two concepts are deeply intertwined and several studies suggest the existence of a link among the optimal resolution, the distribution of detail assigned in the coarse-grained model, and the relevant properties of the system of interest. This connection has its roots in the philosophy behind bottom-up CG modeling, which assumes that the properties of a system should emerge from the behavior of a statistical mechanics-based, simplified model obtained through the (exact) integration of a subset of its degrees of freedom. Usually, this concept of "behavior" refers to the time evolution of the CG system and its conformational space sampling, which enable one to comprehend and understand it. Here, we argue that the process of simplification (mapping) itself can provide hints to non-trivial features of the high-resolution model. This hypothesis has immediate consequences, such as the conversion of coarse-graining methods into analysis tools, a change of paradigm that could constitute a valuable instrument for the analysis of high-resolution, fully atomistic representations of biomolecules.

In bottom-up CG modeling, the choice of the CG mapping has proved to be critical for the properties of interest to emerge systematically (Mullinax and Noid, 2009b; Rudzinski and Noid, 2011). This idea is pushed forward by Rudzinski and Noid (Rudzinski and Noid, 2014), who quantitatively rationalize how the quality of the modeling is influenced by the quality of the mapping. Specifically, the authors group the configurations sampled in a MD simulation into n (m) distinct molecular states of the high-resolution (low-resolution) system; as the low-resolution macrostates clearly depend on the choice of the mapping scheme, Rudzinski and Noid posit that the most informative CG representation should generate a bijective correspondence between atomistic and CG molecular states. This approach allows, in principle, to estimate the optimal level of resolution as well as its distribution. It is thus the system itself that informs the modeler about its low-resolution description that maximizes the consistency with the high-resolution behavior.

This promising paradigm is at the heart of a recent work by Fiorentini and coworkers (Fiorentini et al., 2020), in which a protein-ligand system is considered and the relationship between the binding free energy and the chosen level of resolution is quantified. The authors consider several hybrid atomistic–coarse-grained representations of the protein by treating a variable number of amino acids at the all-atom level. The resulting values of binding free energy are compared with the atomistic reference, showing that the accuracy of the dual-resolution model does not necessarily increase with the spatial extension of the atomistic region. This result suggests the existence of a system-specific, optimal number of amino acids that should be modeled with high detail in such hybrid schemes.

In general, then, the idea has started to emerge that a macromolecular system admits one or more *optimal* reduced models, that is, simplified representations in terms of which it (viz. its high-resolution model) can be *observed* with a marginal loss of information in spite of a loss of detail. Furthermore, it appears more and more evident that such an optimal representation cannot, in general, be uniform: the degree of fidelity with which the original, high-resolution structure is reproduced in the simplified model can vary from point to point, in parallel with the system's chemical, mechanical, dynamical, and functional properties.

Foley and coworkers (Foley et al., 2015; Foley et al., 2020) have pioneered the analysis of the CG model spectrum in a formal and systematic way. In Foley et al. (2015) they considered a one-bead-per-residue Gaussian network model (GNM) of proteins as the reference, high-resolution representation; then, taking advantage of the exact integrability of GNMs, they performed a systematic *decimation* of the system's beads to investigate how reduced models at varying degrees of resolution manage to reproduce fluctuations and correlations of the original model. In so doing, they showed that the information loss that is inherent in the process of coarse-graining is not a monotonic function of the resolution, as an optimal value of the latter was found for which the information content per CG bead (quantified by an appropriate measure) exhibits a maximum. These works thus highlighted the relation between the informativeness of a representation and its resolution *level*.

The impact of resolution *distribution* was later studied by Koehl and coworkers, also in this case making use of ENMs: the *Decimate* (Koehl et al., 2017) algorithm progressively reduces the resolution of a biomolecule by creating a hierarchy of increasingly simplified models, in the spirit of the renormalization group theory. As expected, such CG mappings show an uneven distribution of detail: in the case of globular proteins, for example, optimal models tend to concentrate atoms on the surface of the molecule, thus heavily coarse-graining the inner region—whose mechanical properties require fewer degrees of freedom to be aptly reproduced. A related approach is employed in a work by Diggins et al. (2018): here, the authors identify the CG beads that produce a coarse-grained ENM whose Hamiltonian interaction matrix is as close as possible, measured according to an appropriate distance, to the high-resolution, atomistic ENM. The proposed selection of atoms proves to outperform a random assignment in terms of several observables, such as the intra-block dynamics fraction.

Most of the mentioned approaches can be grouped under the umbrella of methods to optimize the SLRs of a biomolecule in order to improve the capability of the reduced models to faithfully reproduce the atomistic properties of interest. We now summarize the existing methods that, acting as pure filters, focus only on the choice of the SLR without considering the parametrization of the effective interactions.

The first prominent attempts at finding the most informative reduced description of a biomolecule can be ascribed to Voth and coworkers, who employed the χ^2 residual of essential dynamics to estimate the optimal number and partitioning of coarse-grained sites for large protein complexes (ED-CG) (Zhang et al., 2009; Zhang and Voth, 2010; Sinitskiy et al., 2012). In particular, in Sinitskiy et al. (2012) this χ^2 is subject to a constrained minimization, in which the addition of a CG site to a simplified description of a molecule is accepted only if there is a substantial gain in information about the system. Related works (Li et al., 2016a; Li et al., 2016b; Wu et al., 2020) by Xia and colleagues start from the ED-CG method to develop several protocols for the determination of the optimal representations of biomolecules. In Li et al. (2016a) the authors introduce the stepwise optimization with boundary constraint (SOBC) algorithm to enhance the numerical performances of ED-CG (Zhang et al., 2009; Zhang and Voth, 2010) on large proteins. Subsequently (Li et al., 2016b) they propose to maximize the ENM pairwise fluctuations between atoms that are mapped to different CG sites (fluctuation maximisation). The resulting reduced models, once equipped with simple, harmonic interactions, are capable of matching the large-scale fluctuations of the corresponding fine-grained counterparts. More recently, Wu et al. (2020) adopt a combination of ED-CG and internal clustering validation indices to estimate the proper number of sites to coarse-grain proteins. Their results suggest that the appropriate number of C_α atoms to be preserved in a simplified model should lie between one half and one fourth of the total.

Multiple examples of the application of CG'ing methods to analyze simulation data of biomolecules rely on quasi-rigid domain decomposition (Hinsen, 1998; Aleksiev et al., 2009; Potestio et al., 2009). Polles et al. (2013) employed a quasi-rigid domain decomposition of several viral capsids to single out their fundamental mechanical blocks; once validated on a dataset of known viruses, this method is used to formulate predictions about structures whose mechanical subunits had not been characterized yet. Following a similar approach Morra et al. (2012) studied MD trajectories of three representatives of the heat shock protein 90 (Hsp90) family, simulated with and without substrates. They observed that, when the protein is partitioned in as few as three quasi-rigid domains, the relative rigid-like movements of the latter can account for a significant fraction of the system fluctuations, thus allowing to pinpoint two *optimal axes* for rigid rotations of the domains. In turn, the position of these hinges was shown to correspond to two interfaces: while the biological importance of one of them had already been assessed, the other one was hitherto unknown, thus highlighting a potentially druggable functional site.

These remarkable results prove that it is possible to exploit CG methodologies to perform a detailed analysis of the fundamental aspects of an atomistic system. Nevertheless, it is important to notice how these approaches rely on the examination of *mechanical* properties of the system of interest; although they certainly represent simple, intuitive variables to look at, such features do not seem to be as fundamental as the underlying problem they are applied to. Examples of more profound approaches exist that aim at optimizing the SLR of biomolecules in a systematic way (Delvenne et al., 2010; Chen and Habeck, 2017; Boninsegna et al., 2018; Wang and Gómez-Bombarelli, 2019). Delvenne et al. (2010) rank SLRs according to the quality of the corresponding partitioning induced on the protein graph. Chen and Habeck (2017) propose a Bayesian procedure that extracts the optimal SLR from a single macromolecule or cryo-EM map. Boninsegna et al. (2018) combine time-averaged diffusion maps (Banisch and Koltai, 2017) and Markov State Models (Bowman et al., 2013) to select groups of atoms that are mutually close (coherent) over a conformational basin. Wang and Gómez-Bombarelli (2019) employ a variational autoencoder to learn a set of latent CG variables (that is, a SLR) from the atomistic configuration

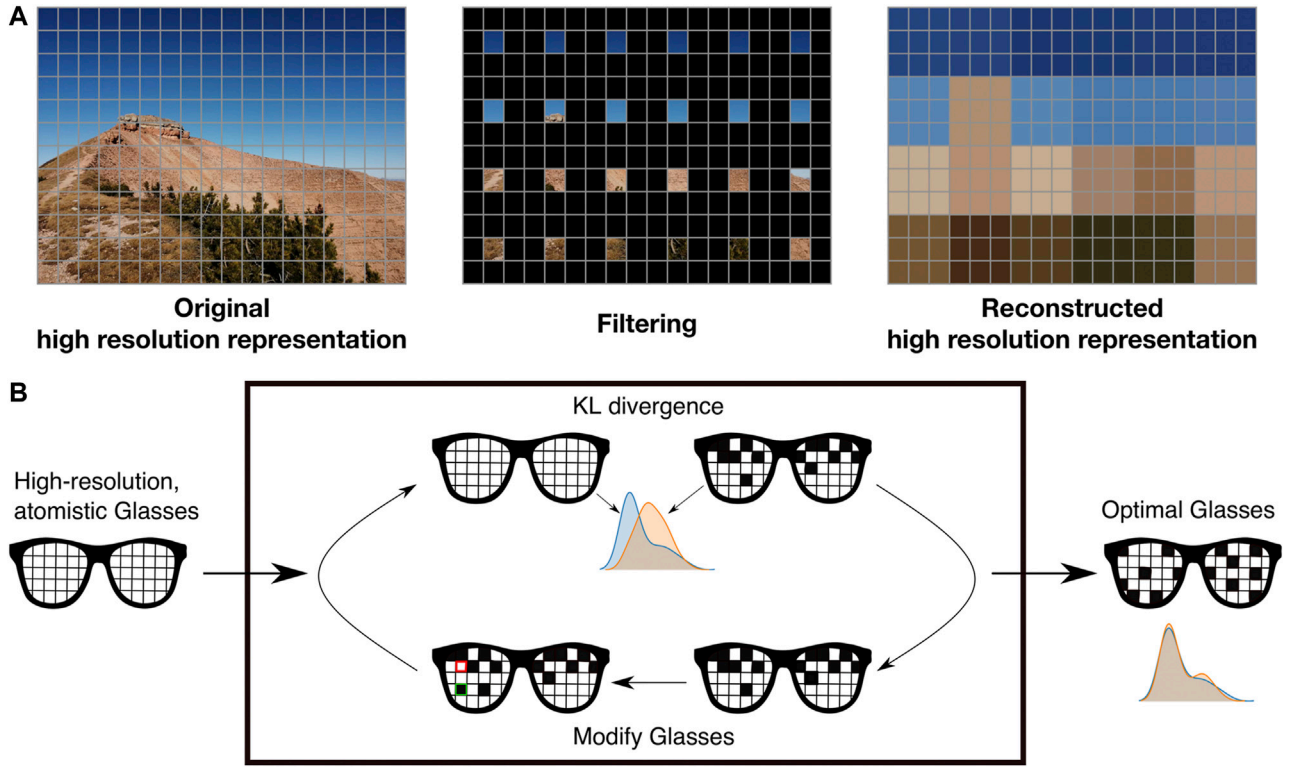


FIGURE 5 | (A) The mapping entropy is defined as the Kullback-Leibler distance between the atomistic, fully detailed probability distribution ***(left panel)*** and its blurred reconstruction ***(right panel)***, obtained from a reduced description obtained neglecting a subset of degrees of freedom ***(middle panel)***. **(B)** The algorithmic implementation of the method introduced by Giulini and coworkers (Giulini et al., 2020) iteratively refines the CG mapping until convergence: the resulting, optimal SLR generates a reconstructed FLR that adheres as much as possible to the original one.

(FLR): in the decoding process the SLR aims at reconstructing the original FLR in a deterministic procedure.

In the spirit of searching for a more significant and informative metric to link the FLR with the space of associated SLRs, some of us proposed a method (Giulini et al., 2020; Errica et al., 2021) that aims at optimizing the choice of the CG mapping through the minimization of the information loss between the description given by the all-atom model and its reduced representation. More specifically, the approach relies on the calculation of the *mapping entropy* S_{map} (Shell, 2008; Rudzinski and Noid, 2011; Foley et al., 2015), which is the "distance", in a Kullback-Leibler sense, between the all-atom Boltzmann distribution and its projection onto the CG space. **Figure 5A** illustrates a comparison of these distributions.

Given a reference all-atom MD simulation and a CG mapping, this protocol optimizes the latter until a (necessarily local) minimum of S_{map} is reached (see **Figure 5B**). CG mappings obtained from independent minimisations of S_{map} share features that are connected to the relevant biological properties of proteins. Moreover, the resolution is not uniformly assigned across the structures, but rather it is distributed to preserve the maximum amount of information about the original, atomistic description.

Since the calculation of S_{map} can be computationally time-consuming, some of us (Errica et al., 2021) have proposed a machine learning model to accelerate the assessment of the quality of a coarse-grained mapping. This improvement allows one to estimate the correct density of states of the system (expressed in terms of the mapping entropy) by means of the Wang Landau sampling scheme, a calculation that would be computationally intractable without such machine learning-based acceleration.

In conclusion, all the works showcased here reflect the emergence of a profound need in the computational biophysics community: that of a strategy to build a faithful simplified representation of a molecular system in an entirely unsupervised manner. In standard coarse-graining recipes, such reduced descriptions must be equipped with proper effective interactions in order to *generate data*. However, the impressive development of techniques to enhance the performances of atomistic simulations is making this necessity less and less pressing. In contrast, the huge amount of high-resolution data produced at each MD run these days might benefit from the capacity of CG models to serve as powerful instruments to *make sense of the data*.

8 DISCUSSION

In this review we have presented a broad, though certainly incomplete, overview of the ideas and motivations behind coarse-grained modeling. The construction of a model of a physical system, simple enough to be employed and understood while detailed enough to enable nontrivial insight, is one of the core activities of science in general. In the study of soft and biological matter, this need becomes particularly pressing and complex, as the advantages of generality, symmetry, and universality one enjoys in areas such as particle physics or statistical mechanics of critical systems lose ground in favor of specificity, peculiarity, and non-transferability; these latter characteristics, however, are those that confer to soft matter its spectacular spectrum of properties.

In such a varied and diverse scenario, one needs a comparably large toolbox of models and analysis techniques to crack the code of the relation among the constituents of a system, their arrangement and relations, and the emergent properties. The term "coarse-grained models" encompasses indeed such a variety, providing descriptions of the same system at different levels of resolution and detail, and serving as instruments to *produce* a given behavior as well as techniques to *analyze* it.

During the past few decades the vast majority of the effort has been put in the usage of coarse-grained models *in lieu* of more detailed, but also computationally more expensive descriptions; the recent impressive advancements of computer science are releasing pressure from this need, and all-atom simulations can now be performed of systems whose size and time scales were yesterday achievable by low-resolution models only.

However, the feasibility of large-scale all-atom simulations is not really putting coarse-grained models out of their job, but rather it is making them change employment: indeed, the extraordinary amount of data generated by such simulations is, in general, all but trivial to understand, and appropriate methods of analysis are required to make this information intelligible. The knowledge acquired in the development of effective low-resolution models thus proves especially useful in discriminating the signal from the noise.

To conclude, based on the presented analysis of the development of protein modeling throughout the decades, we foresee that a bright future lies ahead of coarse-graining: there will always be an inpatient necessity of simple models to investigate complex phenomena, as the curiosity of researchers is bound to lie beyond the capacity of their tools; complementarily, as more and more systems will be viable for accurate and detailed simulations, the need will grow for algorithmic, unsupervised methods to climb the mountain of data, reach its top and say *we understand*.

AUTHOR CONTRIBUTIONS

RP conceived and outlined the work, and wrote **Sections 1, 2, and 8**. TT, MR and RF wrote **Section 3**. RM wrote **Sections 4 and 5**. TT wrote **Section 6**. MG and RP wrote **Section 7**. All authors actively participated in the writing and revision of the whole manuscript.

FUNDING

This project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant No. 758588).

ACKNOWLEDGMENTS

The authors are indebted with Robinson Cortes-Huerto for a critical and insightful reading of the manuscript, as well as with the Reviewers for their supportive comments and constructive suggestions.

REFERENCES

Adcock, S. A., and McCammon, J. A. (2006). Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* 106, 1589–1615. doi:10.1021/cr040426m

Aleksiev, T., Potestio, R., Pontiggia, F., Cozzini, S., and Micheletti, C. (2009). Pisqrd: a Web Server for Decomposing Proteins into Quasi-Rigid Dynamical Domains. *Bioinformatics*. 25, 2743–2744. doi:10.1093/bioinformatics/btp512

Alemani, D., Collu, F., Cascella, M., and Dal Peraro, M. (2010). A Nonradial Coarse-Grained Potential for Proteins Produces Naturally Stable Secondary Structure Elements. *J. Chem. Theor. Comput.* 6, 315–324. doi:10.1021/ct900457z

Allen, E. C., and Rutledge, G. C. (2008). A Novel Algorithm for Creating Coarse-Grained, Density Dependent Implicit Solvent Models. *J. Chem. Phys.* 128, 154115. doi:10.1063/1.2899729

Amaro, R. E., and Mulholland, A. J. (2018). Multiscale Methods in Drug Design Bridge Chemical and Biological Complexity in the Search for Cures. *Nat. Rev. Chem.* 2, 1–12. doi:10.1038/s41570-018-0148

Aoyama, T., Hayakawa, M., Kinoshita, T., and Nio, M. (2015). Tenth-order Electron Anomalous Magnetic Moment: Contribution of Diagrams without Closed Lepton Loops. *Phys. Rev. D.* 91, 033006. doi:10.1103/physrevd.91.033006

Arkhipov, A., Freddolino, P. L., Imada, K., Namba, K., and Schulten, K. (2006a). Coarse-grained Molecular Dynamics Simulations of a Rotating Bacterial Flagellum. *Biophysical J.* 91, 4589–4597. doi:10.1529/biophysj.106.093443

Arkhipov, A., Freddolino, P. L., and Schulten, K. (2006b). Stability and Dynamics of Virus Capsids Described by Coarse-Grained Modeling. *Structure*. 14, 1767–1777. doi:10.1016/j.str.2006.10.003

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001). Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophysical J.* 80, 505–515. doi:10.1016/s0006-3495(01)76033-x

Bahar, I., and Jernigan, R. L. (1997). Inter-residue Potentials in Globular Proteins and the Dominance of Highly Specific Hydrophilic Interactions at Close Separation 1 1 Edited by B. Honig. *J. Mol. Biol.* 266, 195–214. doi:10.1006/jmbi.1996.0758

Banisch, R., and Koltai, P. (2017). Understanding the Geometry of Transport: Diffusion Maps for Lagrangian Trajectory Data Unravel Coherent Sets. *Chaos: Interdiscip. J. Nonlinear Sci.* 27, 035804. doi:10.1063/1.4971788

Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* 100, 020603. doi:10.1103/PhysRevLett.100.020603

Barnes, T. A., Goodpaster, J. D., Manby, F. R., and Miller, T. F., III (2013). Accurate Basis Set Truncation for Wavefunction Embedding. *J. Chem. Phys.* 139, 024103. doi:10.1063/1.4811112

Beglov, D., and Roux, B. (1994). Finite Representation of an Infinite Bulk System: Solvent Boundary Potential for Computer Simulations. *J. Chem. Phys.* 100, 9050–9063. doi:10.1063/1.466711

Bellemans, A., Orban, J., and Van Belle, D. (1980). Molecular Dynamics of Rigid and Non-rigid Necklaces of Hard Discs. *Mol. Phys.* 39, 781–782. doi:10.1080/00268978000100671

Bereau, T., and Deserno, M. (2009). Generic Coarse-Grained Model for Protein Folding and Aggregation. *J. Chem. Phys.* 130, 06B621. doi:10.1063/1.3152842

Bereau, T., Wang, Z. J., and Deserno, M. (2014). More Than the Sum of its Parts: Coarse-Grained Peptide-Lipid Interactions from a Simple Cross-Parametrization. *J. Chem. Phys.* 140, 03B615_1–11220. doi:10.1063/1.4867465

Bereau, T., Bachmann, M., and Deserno, M. (2010). Interplay between Secondary and Tertiary Structure Formation in Protein Folding Cooperativity. *J. Am. Chem. Soc.* 132, 13129–13131. doi:10.1021/ja105206w

Bernardi, R. C., Melo, M. C. R., and Schulten, K. (2015). Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems. *Biochim. Biophys. Acta* 1850, 872–877. doi:10.1016/j.bbagen.2014.10.019

Best, R. B., Chen, Y.-G., and Hummer, G. (2005). Slow Protein Conformational Dynamics from Multiple Experimental Structures: the helix/sheet Transition of Arc Repressor. *Structure*. 13, 1755–1763. doi:10.1016/j.str.2005.08.009

Betancourt, M. R. (2008). Knowledge-based Potential for the Polypeptide Backbone. *J. Phys. Chem. B*. 112, 5058–5069. doi:10.1021/jp076906+

Böde, C., Kovács, I. A., Szalay, M. S., Palotai, R., Korcsmáros, T., and Csermely, P. (2007). Network Analysis of Protein Dynamics. *Febs Lett.* 581, 2776–2782. doi:10.1016/j.febslet.2007.05.021

Boereboom, J. M., Potestio, R., Donadio, D., and Bulo, R. E. (2016). Toward Hamiltonian Adaptive QM/MM: Accurate Solvent Structures Using many-body Potentials. *J. Chem. Theor. Comput.* 12, 3441–3448. doi:10.1021/acs.jctc.6b00205

Bond, P. J., and Sansom, M. S. P. (2006). Insertion and Assembly of Membrane Proteins via Simulation. *J. Am. Chem. Soc.* 128, 2697–2704. doi:10.1021/ja0569104

Boninsegna, L., Banisch, R., and Clementi, C. (2018). A Data-Driven Perspective on the Hierarchical Assembly of Molecular Structures. *J. Chem. Theor. Comput.* 14, 453–460. doi:10.1021/acs.jctc.7b00990

Born, M., and Oppenheimer, R. (1927). Zur quantentheorie der molekeln. *Ann. Phys.* 389, 457–484. doi:10.1002/andp.19273892002

Bowerman, S., and Wereszczynski, J. (2016). Detecting Allosteric Networks Using Molecular Dynamics Simulation. *Methods Enzymol.* 578, 429–447. doi:10.1016/bs.mie.2016.05.027

Bowman, G. R., Pande, V. S., and Noé, F. (2013). *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. New York City, NY: Springer Science & Business Media. Vol. 797

Brünger, A., Brooks, C. L., III, and Karplus, M. (1984). Stochastic Boundary Conditions for Molecular Dynamics Simulations of St2 Water. *Chem. Phys. Lett.* 105, 495–500. doi:10.1016/0009-2614(84)80098-6

Burgess, C. P. (2020). *Introduction to Effective Field Theory*. Cambridge, UK: Cambridge University Press. doi:10.1017/9781139048040

Calandrini, V., Rossetti, G., Arnesano, F., Natile, G., and Carloni, P. (2015). Computational Metallomics of the Anticancer Drug Cisplatin. *J. Inorg. Biochem.* 153, 231–238. doi:10.1016/j.jinorgbio.2015.10.001

Chen, M., Lin, X., Lu, W., Onuchic, J. N., and Wolynes, P. G. (2017). Protein Folding and Structure Prediction from the Ground up II: AAWSEM for α/β Proteins. *J. Phys. Chem. B*. 121, 3473–3482. doi:10.1021/acs.jpcb.6b09347

Chen, M., Lin, X., Zheng, W., Onuchic, J. N., and Wolynes, P. G. (2016). Protein Folding and Structure Prediction from the Ground up: The Atomistic Associative Memory, Water Mediated, Structure and Energy Model. *J. Phys. Chem. B*. 120, 8557–8565. doi:10.1021/acs.jpcb.6b02451

Chen, X., Chen, M., Schafer, N. P., and Wolynes, P. G. (2020). Exploring the Interplay between Fibrillization and Amorphous Aggregation Channels on the Energy Landscapes of Tau Repeat Isoforms. *Proc. Natl. Acad. Sci. USA*. 117, 4125–4130. doi:10.1073/pnas.1921702117

Chen, Y. L., and Habeck, M. (2017). Data-driven Coarse Graining of Large Biomolecular Structures. *PloS one*. 12, e0183057. doi:10.1371/journal.pone.0183057

Cheon, M., Chang, I., and Hall, C. K. (2010). Extending the Prime Model for Protein Aggregation to All 20 Amino Acids. *Proteins* 78, 2950–2960. doi:10.1002/prot.22817

Cheon, M., Chang, I., and Hall, C. K. (2011). Spontaneous Formation of Twisted Aβ16-22 Fibrils in Large-Scale Molecular-Dynamics Simulations. *Biophysical J.* 101, 2493–2501. doi:10.1016/j.bpj.2011.08.042

Chiariello, M. G., Bolnykh, V., Ippoliti, E., Meloni, S., Olsen, J. M. H., Beck, T., et al. (2020). Molecular Basis of Clc Antiporter Inhibition by Fluoride. *J. Am. Chem. Soc.* 142, 7254–7258. doi:10.1021/jacs.9b13588

Chodera, J. D., and Noé, F. (2014). Markov State Models of Biomolecular Conformational Dynamics. *Curr. Opin. Struct. Biol.* 25, 135–144. doi:10.1016/j.sbi.2014.04.002

Chu, J.-W., and Voth, G. A. (2006). Coarse-grained Modeling of the Actin Filament Derived from Atomistic-Scale Simulations. *Biophysical J.* 90, 1572–1582. doi:10.1529/biophysj.105.073924

Clementi, C., Nymeyer, H., and Onuchic, J. N. (2000). Topological and Energetic Factors: what Determines the Structural Details of the Transition State Ensemble and "En-Route" Intermediates for Protein Folding? an Investigation for Small Globular Proteins. *J. Mol. Biol.* 298, 937–953. doi:10.1006/jmbi.2000.3693

Coluzza, I. (2011). A Coarse-Grained Approach to Protein Design: Learning from Design to Understand Folding. *PloS one* 6, e20853. doi:10.1371/journal.pone.0020853

Corradi, V., Mendez-Villuendas, E., Ingólfsson, H. I., Gu, R.-X., Siuda, I., Melo, M. N., et al. (2018). Lipid-Protein Interactions Are Unique Fingerprints for Membrane Proteins. *ACS Cent. Sci.* 4, 709–717. doi:10.1021/acscentsci.8b00143

D'Adamo, G., Menichetti, R., Pelissetto, A., and Pierleoni, C. (2015). Coarse-graining Polymer Solutions: A Critical Appraisal of Single-And Multi-Site Models. *Eur. Phys. J. Spec. Top.* 224, 2239–2267. doi:10.1140/epjst/e2015-02410-3

Dama, J. F., Jin, J., and Voth, G. A. (2017). The Theory of Ultra-coarse-graining. 3. Coarse-Grained Sites with Rapid Local Equilibrium of Internal States. *J. Chem. Theor. Comput.* 13, 1010–1022. doi:10.1021/acs.jctc.6b01081

Dama, J. F., Sinitskiy, A. V., McCullagh, M., Weare, J., Roux, B., Dinner, A. R., et al. (2013). The Theory of Ultra-coarse-graining. 1. General Principles. *J. Chem. Theor. Comput.* 9, 2466–2480. doi:10.1021/ct4000444

Dannenhoffer-Lafage, T., Wagner, J. W., Durumeric, A. E. P., and Voth, G. A. (2019). Compatible Observable Decompositions for Coarse-Grained Representations of Real Molecular Systems. *J. Chem. Phys.* 151, 134115. doi:10.1063/1.5116027

Dans, P. D., Zeida, A., Machado, M. R., and Pantano, S. (2010). A Coarse Grained Model for Atomic-Detailed Dna Simulations with Explicit Electrostatics. *J. Chem. Theor. Comput.* 6, 1711–1725. doi:10.1021/ct900653p

Darré, L., Machado, M. R., Brandner, A. F., González, H. C., Ferreira, S., and Pantano, S. (2015). Sirah: a Structurally Unbiased Coarse-Grained Force Field for Proteins with Aqueous Solvation and Long-Range Electrostatics. *J. Chem. Theor. Comput.* 11, 723–739. doi:10.1021/ct5007746

Davtyan, A., Dama, J. F., Sinitskiy, A. V., and Voth, G. A. (2014). The Theory of Ultra-coarse-graining. 2. Numerical Implementation. *J. Chem. Theor. Comput.* 10, 5265–5275. doi:10.1021/ct500834t

Davtyan, A., Schafer, N. P., Zheng, W., Clementi, C., Wolynes, P. G., and Papoian, G. A. (2012). Awsem-md: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. *J. Phys. Chem. B*. 116, 8494–8503. doi:10.1021/jp212541y

Delgado-Buscalioni, R., Kremer, K., and Praprotnik, M. (2009). Coupling Atomistic and Continuum Hydrodynamics through a Mesoscopic Model: Application to Liquid Water. *J. Chem. Phys.* 131, 244107. doi:10.1063/1.3272265

Delvenne, J. C., Yaliraki, S. N., and Barahona, M. (2010). Stability of Graph Communities across Time Scales. *Proc. Natl. Acad. Sci.* 107, 12755–12760. doi:10.1073/pnas.0903215107

DeLyser, M. R., and Noid, W. G. (2019). Analysis of Local Density Potentials. *J. Chem. Phys.* 151, 224106. doi:10.1063/1.5128665

Deng, Y., and Roux, B. (2008). Computation of Binding Free Energy with Molecular Dynamics and Grand Canonical Monte Carlo Simulations. *J. Chem. Phys.* 128, 03B611. doi:10.1063/1.2842080

Derreumaux, P. (1999). From Polypeptide Sequences to Structures Using Monte Carlo Simulations and an Optimized Potential. *J. Chem. Phys.* 111, 2301–2310. doi:10.1063/1.479501

Diggins, P., Liu, C., Deserno, M., and Potestio, R. (2018). Optimal Coarse-Grained Site Selection in Elastic Network Models of Biomolecules. *J. Chem. Theor. Comput.* 15, 648–664. doi:10.1021/acs.jctc.8b00654

Dignon, G. L., Zheng, W., Kim, Y. C., Best, R. B., and Mittal, J. (2018). Sequence Determinants of Protein Phase Behavior from a Coarse-Grained Model. *PLoS Comput. Biol.* 14, e1005941. doi:10.1371/journal.pcbi.1005941

Dignon, G. L., Zheng, W., Kim, Y. C., and Mittal, J. (2019). Temperature-Controlled Liquid-Liquid Phase Separation of Disordered Proteins. *ACS Cent. Sci.* 5, 821–830. doi:10.1021/acscentsci.9b00102

Dijkstra, M., van Roij, R., and Evans, R. (1999). Phase Diagram of Highly Asymmetric Binary Hard-Sphere Mixtures. *Phys. Rev. E.* 59, 5744–5771. doi:10.1103/physreve.59.5744

Doruker, P., Jernigan, R. L., and Bahar, I. (2002). Dynamics of Large Proteins through Hierarchical Levels of Coarse-Grained Structures. *J. Comput. Chem.* 23, 119–127. doi:10.1002/jcc.1160

Eom, K., Baek, S.-C., Ahn, J.-H., and Na, S. (2007). Coarse-graining of Protein Structures for the normal Mode Studies. *J. Comput. Chem.* 28, 1400–1410. doi:10.1002/jcc.20672

Errica, F., Giulini, M., Bacciu, D., Menichetti, R., Micheli, A., and Potestio, R. (2021). A Deep Graph Network-Enhanced Sampling Approach to Efficiently Explore the Space of Reduced Representations of Proteins. *Front. Mol. Biosciences*. 8, 136. doi:10.3389/fmolb.2021.637396

Fierro, F., Giorgetti, A., Carloni, P., Meyerhof, W., and Alfonso-Prieto, M. (2019). Dual Binding Mode of "Bitter Sugars" to Their Human Bitter Taste Receptor Target. *Scientific Rep.* 9, 1–16. doi:10.1038/s41598-019-44805-z

Fiorentini, R., Kremer, K., and Potestio, R. (2020). Ligand-protein Interactions in Lysozyme Investigated through a Dual-Resolution Model. *Proteins: Struct. Funct. Bioinformatics* 88, 1351-1360. https://doi.org/10.1002/prot.25954

Fiorentini, R., Kremer, K., Potestio, R., and Fogarty, A. C. (2017). Using Force-Based Adaptive Resolution Simulations to Calculate Solvation Free Energies of Amino Acid Sidechain Analogues. *J. Chem. Phys.* 146, 244113. doi:10.1063/1.4989486

Fogarty, A. C., Potestio, R., and Kremer, K. (2015). Adaptive Resolution Simulation of a Biomolecule and its Hydration Shell: Structural and Dynamical Properties. *J. Chem. Phys.* 142, 05B610_1. doi:10.1063/1.4921347

Fogarty, A. C., Potestio, R., and Kremer, K. (2016). A Multi-Resolution Model to Capture Both Global Fluctuations of an Enzyme and Molecular Recognition in the Ligand-Binding Site. *Proteins*. 84, 1902–1913. doi:10.1002/prot.25173

Foley, T. T., Shell, M. S., and Noid, W. G. (2015). The Impact of Resolution upon Entropy and Information in Coarse-Grained Models. *J. Chem. Phys.* 143, 12B601_1. doi:10.1063/1.4929836

Foley, T. T., Kidder, K. M., Shell, M. S., and Noid, W. G. (2020). Exploring the Landscape of Model Representations. *Proc. Natl. Acad. Sci. USA*. 117, 24061–24068. doi:10.1073/pnas.2000098117

Freddolino, P. L., Liu, F., Gruebele, M., and Schulten, K. (2008). Ten-microsecond Molecular Dynamics Simulation of a Fast-Folding Ww Domain. *Biophysical J.* 94, L75. doi:10.1529/biophysj.108.131565

Frishman, D., and Argos, P. (1995). Knowledge-based Protein Secondary Structure Assignment. *Proteins* 23, 566–579. doi:10.1002/prot.340230412

Garay, P. G., Barrera, E. E., and Pantano, S. (2019). Post-translational Modifications at the Coarse-Grained Level with the Sirah Force Field. *J. Chem. Inf. Model.* 60, 964–973. doi:10.1021/acs.jcim.9b00900

Giulini, M., Menichetti, R., Shell, M. S., and Potestio, R. (2020). An Information-Theory-Based Approach for Optimal Model Reduction of Biomolecules. *J. Chem. Theor. Comput.* 16, 6795–6813. doi:10.1021/acs.jctc.0c00676

Go, N. (1983). Theoretical Studies of Protein Folding. *Annu. Rev. Biophys. Bioeng.* 12, 183–210. doi:10.1146/annurev.bb.12.060183.001151

González, M. A. (2011). Force fields and Molecular Dynamics Simulations. *Journées de la Neutronique* 12, 169–200. doi:10.1051/sfn/201112009

Gowers, R. J., and Carbone, P. (2015). A Multiscale Approach to Model Hydrogen Bonding: The Case of Polyamide. *J. Chem. Phys.* 142, 224907. doi:10.1063/1.4922445

Gołaś, E., Maisuradze, G. G., Senet, P., Ołdziej, S., Czaplewski, C., Scheraga, H. A., et al. (2012). Simulation of the Opening and Closing of Hsp70 Chaperones by Coarse-Grained Molecular Dynamics. *J. Chem. Theor. Comput*. 8, 1750–1764. doi:10.1021/ct200680g

Grime, J. M., Dama, J. F., Ganser-Pornillos, B. K., Woodward, C. L., Jensen, G. J., Yeager, M., et al. (2016). Coarse-grained Simulation Reveals Key Features of Hiv-1 Capsid Self-Assembly. *Nat. Commun.* 7, 1–11. doi:10.1038/ncomms11568

Hagan, M. F., and Zandi, R. (2016). Recent Advances in Coarse-Grained Modeling of Virus Assembly. *Curr. Opin. Virol.* 18, 36–43. doi:10.1016/j.coviro.2016.02.012

Hagler, A. T., and Honig, B. (1978). On the Formation of Protein Tertiary Structure on a Computer. *Proc. Natl. Acad. Sci.* 75, 554–558. doi:10.1073/pnas.75.2.554

Haldar, S., Comitani, F., Saladino, G., Woods, C., van der Kamp, M. W., Mulholland, A. J., et al. (2018). A Multiscale Simulation Approach to Modeling Drug-Protein Binding Kinetics. *J. Chem. Theor. Comput.* 14, 6093–6101. doi:10.1021/acs.jctc.8b00687

Hanneke, D., Fogwell Hoogerheide, S., and Gabrielse, G. (2011). Cavity Control of a Single-Electron Quantum Cyclotron: Measuring the Electron Magnetic Moment. *Phys. Rev. A.* 83, 052122. doi:10.1103/physreva.83.052122

Hanson, B., Richardson, R., Oliver, R., Read, D. J., Harlen, O., and Harris, S. (2015). Modelling Biomacromolecular Assemblies with Continuum Mechanics. *Biochem. Soc. Trans.* 43, 186–192. doi:10.1042/bst20140294

Hanson, B. S., Iida, S., Read, D. J., Harlen, O. G., Kurisu, G., Nakamura, H., et al. (2021). Continuum Mechanical Parameterisation of Cytoplasmic Dynein from Atomistic Simulation. *Methods*. 185, 39–48. doi:10.1016/j.ymeth.2020.01.021

Harmandaris, V. A., Adhikari, N. P., van der Vegt, N. F. A., and Kremer, K. (2006). Hierarchical Modeling of Polystyrene: From Atomistic to Coarse-Grained Simulations. *Macromolecules*. 39, 6708–6719. doi:10.1021/ma0606399

Heidari, M., Cortes-Huerto, R., Potestio, R., and Kremer, K. (2019). Steering a Solute between Coexisting Solvation States: Revisiting Nonequilibrium Work Relations and the Calculation of Free Energy Differences. *J. Chem. Phys.* 151, 144105. doi:10.1063/1.5117780

Hills, R., and Brooks, C. (2009). Insights from Coarse-Grained Gō Models for Protein Folding and Dynamics. *Int J Mol Sci.* 10, 889–905. doi:10.3390/ijms10030889

Hinsen, K. (1998). Analysis of Domain Motions by Approximate normal Mode Calculations. *Proteins*. 33, 417–429. doi:10.1002/(sici)1097-0134(19981115)33:3<417::aid-prot10>3.0.co;2-8

Hu, J., Chen, T., Wang, M., Chan, H. S., and Zhang, Z. (2017). A Critical Comparison of Coarse-Grained Structure-Based Approaches and Atomic Models of Protein Folding. *Phys. Chem. Chem. Phys.* 19, 13629–13639. doi:10.1039/C7CP01532A

Hu, J., Korotkin, I. A., and Karabasov, S. A. (2019). Hybrid Multiscale Simulation Reveals Focusing of a Diffusing Peptide Molecule by Parallel Shear Flow in Water. *J. Mol. Liquids*. 280, 285–297. doi:10.1016/j.molliq.2019.01.152

Huang, J., and MacKerell, A. D., Jr (2013). Charmm36 All-Atom Additive Protein Force Field: Validation Based on Comparison to Nmr Data. *J. Comput. Chem.* 34, 2135–2145. doi:10.1002/jcc.23354

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., De Groot, B. L., et al. (2017). Charmm36m: an Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods*. 14, 71–73. doi:10.1038/nmeth.4067

Hyeon, C., Lorimer, G. H., and Thirumalai, D. (2006). Dynamics of Allosteric Transitions in Groel. *Proc. Natl. Acad. Sci.* 103, 18939–18944. doi:10.1073/pnas.0608759103

Im, W., Bernèche, S., and Roux, B. (2001). Generalized Solvent Boundary Potential for Computer Simulations. *J. Chem. Phys.* 114, 2924–2937. doi:10.1063/1.1336570

Izvekov, S., and Voth, G. A. (2005). A Multiscale Coarse-Graining Method for Biomolecular Systems. *J. Phys. Chem. B.* 109, 2469–2473. doi:10.1021/jp044629q

Jang, H., Na, S., and Eom, K. (2009). Multiscale Network Model for Large Protein Dynamics. *J. Chem. Phys.* 131, 12B623. doi:10.1063/1.3282447

Javanainen, M., Hammaren, H., Monticelli, L., Jeon, J.-H., Miettinen, M. S., Martinez-Seara, H., et al. (2013). Anomalous and normal Diffusion of Proteins and Lipids in Crowded Lipid Membranes. *Faraday Discuss.* 161, 397–417. doi:10.1039/c2fd20085f

Jin, S., Chen, M., Chen, X., Bueno, C., Lu, W., Schafer, N. P., et al. (2020). Protein Structure Prediction in Casp13 Using Awsem-Suite. *J. Chem. Theor. Comput.* 16, 3977–3988. doi:10.1021/acs.jctc.0c00188

Kabsch, W., and Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*. 22, 2577–2637. doi:10.1002/bip.360221211

Kanada, R., Kuwata, T., Kenzaki, H., and Takada, S. (2013). Structure-based Molecular Simulations Reveal the Enhancement of Biased Brownian

Motions in Single-Headed Kinesin. *Plos Comput. Biol.* 9, e1002907. doi:10.1371/journal.pcbi.1002907

Karanicolas, J., and Brooks, C. L., III (2002). The Origins of Asymmetry in the Folding Transition States of Protein L and Protein G. *Protein Sci.* 11, 2351–2361. doi:10.1110/ps.0205402

Kim, B. L., Schafer, N. P., and Wolynes, P. G. (2014). Predictive Energy Landscapes for Folding -helical Transmembrane Proteins. *Proc. Natl. Acad. Sci.* 111, 11031–11036. doi:10.1073/pnas.1410529111

Sanejouand, Y.-H. (2013). Elastic network models: theoretical and empirical foundations. *Biomol. Simulat.* 924, 601–616.

Kim, T.-R., Yang, J. S., Shin, S., and Lee, J. (2013). Statistical Torsion Angle Potential Energy Functions for Protein Structure Modeling: a Bicubic Interpolation Approach. *Proteins*. 81, 1156–1165. doi:10.1002/prot.24265

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., and Kolinski, A. (2016). Coarse-grained Protein Models and Their Applications. *Chem. Rev.* 116, 7898–7936. doi:10.1021/acs.chemrev.6b00163

Koehl, P., Poitevin, F., Navaza, R., and Delarue, M. (2017). The Renormalization Group and its Applications to Generating Coarse-Grained Models of Large Biological Molecular Systems. *J. Chem. Theor. Comput.* 13, 1424–1438. doi:10.1021/acs.jctc.6b01136

Kolinski, A., Godzik, A., and Skolnick, J. (1993). A General Method for the Prediction of the Three Dimensional Structure and Folding Pathway of Globular Proteins: Application to Designed Helical Proteins. *J. Chem. Phys.* 98, 7420–7433. doi:10.1063/1.464706

Korshunova, K., and Carloni, P. (2021). Ligand Affinities within the Open-Boundary Molecular Mechanics/coarse-Grained Framework (I): Alchemical Transformations within the Hamiltonian Adaptive Resolution Scheme. *J. Phys. Chem. B*. 125, 789–797. doi:10.1021/acs.jpcb.0c09805

Kreis, K., Fogarty, A. C., Kremer, K., and Potestio, R. (2015). Advantages and Challenges in Coupling an Ideal Gas to Atomistic Models in Adaptive Resolution Simulations. *Eur. Phys. J. Spec. Top.* 224, 2289–2304. doi:10.1140/epjst/e2015-02412-1

Kreis, K., Kremer, K., Potestio, R., and Tuckerman, M. E. (2017). From Classical to Quantum and Back: Hamiltonian Adaptive Resolution Path Integral, Ring Polymer, and Centroid Molecular Dynamics. *J. Chem. Phys.* 147, 244104. doi:10.1063/1.5000701

Kreis, K., Potestio, R., Kremer, K., and Fogarty, A. C. (2016). Adaptive Resolution Simulations with Self-Adjusting High-Resolution Regions. *J. Chem. Theor. Comput.* 12, 4067–4081. doi:10.1021/acs.jctc.6b00440

Kubelka, J., Hofrichter, J., and Eaton, W. A.. (2004) The Protein Folding 'speed Limit'. *Curr. Opin. Struct. Biol.* 14, 76–88. doi:10.1016/j.sbi.2004.01.013

Kurkcuoglu, O., Jernigan, R. L., and Doruker, P. (2004). Mixed Levels of Coarse-Graining of Large Proteins Using Elastic Network Model Succeeds in Extracting the Slowest Motions. *Polymer*. 45, 649–657. doi:10.1016/j.polymer.2003.10.071

Kutzner, C., Páll, S., Fechner, M., Esztermann, A., Groot, B. L., and Grubmüller, H. (2019). More Bang for Your Buck: Improved Use of Gpu Nodes for Gromacs 2018. *J. Comput. Chem.* 40, 2418–2431. doi:10.1002/jcc.26011

Kynast, P., Derreumaux, P., and Strodel, B. (2016). Evaluation of the Coarse-Grained Opep Force Field for Protein-Protein Docking. *BMC Biophys.* 9, 1–17. doi:10.1186/s13628-016-0029-y

Lamiable, A., Thévenet, P., Rey, J., Vavrusa, M., Derreumaux, P., and Tufféry, P. (2016). PEP-FOLD3: Fasterde Novostructure Prediction for Linear Peptides in Solution and in Complex. *Nucleic Acids Res.* 44, W449–W454. doi:10.1093/nar/gkw329

Lange, O. F., and Grubmüller, H. (2006). Generalized Correlation for Biomolecular Dynamics. *Proteins*. 62, 1053–1061. doi:10.1002/prot.20784

Lebold, K. M., and Noid, W. G. (2019a). Dual Approach for Effective Potentials that Accurately Model Structure and Energetics. *J. Chem. Phys.* 150, 234107. doi:10.1063/1.5094330

Lebold, K. M., and Noid, W. G. (2019b). Dual-potential Approach for Coarse-Grained Implicit Solvent Models with Accurate, Internally Consistent Energetics and Predictive Transferability. *J. Chem. Phys.* 151, 164113. doi:10.1063/1.5125246

Lee, E. H., Hsin, J., Sotomayor, M., Comellas, G., and Schulten, K. (2009). Discovery through the Computational Microscope. *Structure*. 17, 1295–1306. doi:10.1016/j.str.2009.09.001

Lee, M. S., Salsbury, F. R., Jr, and Olson, M. A. (2004). An Efficient Hybrid Explicit/implicit Solvent Method for Biomolecular Simulations. *J. Comput. Chem.* 25, 1967–1978. doi:10.1002/jcc.20119

Lee, T.-S., Cerutti, D. S., Mermelstein, D., Lin, C., LeGrand, S., Giese, T. J., et al. (2018). Gpu-accelerated Molecular Dynamics and Free Energy Methods in Amber18: Performance Enhancements and New Features. *J. Chem. Inf. Model.* 58, 2043–2050. doi:10.1021/acs.jcim.8b00462

Leguèbe, M., Nguyen, C., Capece, L., Hoang, Z., Giorgetti, A., and Carloni, P. (2012). Hybrid Molecular Mechanics/coarse-Grained Simulations for Structural Prediction of G-Protein Coupled Receptor/ligand Complexes. *PloS one*. 7, e47332. doi:10.1371/journal.pone.0047332

Levitt, M., and Warshel, A. (1975). Computer Simulation of Protein Folding. *Nature*. 253, 694–698. doi:10.1038/253694a0

Li, M., Zhang, J. Z. H., and Xia, F. (2016a). A New Algorithm for Construction of Coarse-Grained Sites of Large Biomolecules. *J. Comput. Chem.* 37, 795–804. doi:10.1002/jcc.24265

Li, M., Zhang, J. Z., and Xia, F. (2016b). Constructing Optimal Coarse-Grained Sites of Huge Biomolecules by Fluctuation Maximization. *J. Chem. Theor. Comput.* 12, 2091–2100. doi:10.1021/acs.jctc.6b00016

Lin, X., Kulkarni, P., Bocci, F., Schafer, N., Roy, S., Tsai, M.-Y., et al. (2019). Structural and Dynamical Order of a Disordered Protein: Molecular Insights into Conformational Switching of page4 at the Systems Level. *Biomolecules* 9, 77. doi:10.3390/biom9020077

Lindert, S., Bucher, D., Eastman, P., Pande, V., and McCammon, J. A. (2013). Accelerated Molecular Dynamics Simulations with the Amoeba Polarizable Force Field on Graphics Processing Units. *J. Chem. Theor. Comput.* 9, 4684–4691. doi:10.1021/ct400514p

Liwo, A., Baranowski, M., Czaplewski, C., Gołaś, E., He, Y., Jagieła, D., et al. (2014). A Unified Coarse-Grained Model of Biological Macromolecules Based on Mean-Field Multipole–Multipole Interactions. *J. Mol. Model.* 20, 2306. doi:10.1007/s00894-014-2306-5

Liwo, A. (2013). Coarse Graining: a Tool for Large-Scale Simulations or More?. *Physica Scripta*. 87, 058502. doi:10.1088/0031-8949/87/05/058502

Liwo, A., Czaplewski, C., Sieradzan, A. K., Lubecka, E. A., Lipska, A. G., Golon, Ł., et al. (2020). Scale-consistent Approach to the Derivation of Coarse-Grained Force fields for Simulating Structure, Dynamics, and Thermodynamics of Biopolymers. *Prog. Mol. Biol. translational Sci.* 170, 73–122. doi:10.1016/bs.pmbts.2019.12.004

Liwo, A., Khalili, M., and Scheraga, H. A. (2005). Ab Initio simulations of Protein-Folding Pathways by Molecular Dynamics with the United-Residue Model of Polypeptide Chains. *Proc. Natl. Acad. Sci.* 102, 2362–2367. doi:10.1073/pnas.0408885102

Liwo, A., Lee, J., Ripoll, D. R., Pillardy, J., and Scheraga, H. A. (1999). Protein Structure Prediction by Global Optimization of a Potential Energy Function. *Proc. Natl. Acad. Sci.* 96, 5482–5485. doi:10.1073/pnas.96.10.5482

Lodola, A., Mor, M., Rivara, S., Christov, C., Tarzia, G., Piomelli, D., et al. (2008). Identification of Productive Inhibitor Binding Orientation in Fatty Acid Amide Hydrolase (Faah) by QM/MM Mechanistic Modelling. *Chem. Commun.* 2008, 214–216. doi:10.1039/b714136j

Lonsdale, R., Rouse, S. L., Sansom, M. S., and Mulholland, A. J. (2014). A Multiscale Approach to Modelling Drug Metabolism by Membrane-Bound Cytochrome P450 Enzymes. *Plos Comput. Biol.* 10, e1003714. doi:10.1371/journal.pcbi.1003714

Lonsdale, R., Houghton, K. T., Żurek, J., Bathelt, C. M., Foloppe, N., de Groot, M. J., et al. (2013). Quantum Mechanics/molecular Mechanics Modeling of Regioselectivity of Drug Metabolism in Cytochrome P450 2c9. *J. Am. Chem. Soc.* 135, 8001–8015. doi:10.1021/ja402016p

Lu, Q., and Wang, J. (2008). Single Molecule Conformational Dynamics of Adenylate Kinase: Energy Landscape, Structural Correlations, and Transition State Ensembles. *J. Am. Chem. Soc.* 130, 4772–4783. doi:10.1021/ja0780481

Lu, W., Schafer, N. P., and Wolynes, P. G. (2018). Energy Landscape Underlying Spontaneous Insertion and Folding of an Alpha-Helical Transmembrane Protein into a Bilayer. *Nat. Commun.* 9, 1–10. doi:10.1038/s41467-018-07320-9

Lu, Y., Wei, G., and Derreumaux, P. (2012). Structural, Thermodynamical, and Dynamical Properties of Oligomers Formed by the Amyloid Nnqq Peptide: Insights from Coarse-Grained Simulations. *J. Chem. Phys.* 137, 025101. doi:10.1063/1.4732761

Lyubartsev, A., Mirzoev, A., Chen, L., and Laaksonen, A. (2010). Systematic Coarse-Graining of Molecular Models by the newton Inversion Method. *Faraday Discuss.* 144, 43–56. doi:10.1039/b901511f

Lyubartsev, A. P., and Laaksonen, A. (1995). Calculation of Effective Interaction Potentials from Radial Distribution Functions: A Reverse Monte Carlo Approach. *Phys. Rev. E.* 52, 3730–3737. doi:10.1103/physreve.52.3730

Machado, M. R., Barrera, E. E., Klein, F., Sóñora, M., Silva, S., and Pantano, S. (2019). The Sirah 2.0 Force Field: Altius, Fortius, Citius. *J. Chem. Theor. Comput.* 15, 2719–2733. doi:10.1021/acs.jctc.9b00006

Magalhães, R. P., Fernandes, H. S., and Sousa, S. F. (2020). Modelling Enzymatic Mechanisms with QM/MM Approaches: Current Status and Future Challenges. *Isr. J. Chem.* 60, 655-666. doi:10.1002/ijch.202000014

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14sb: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99sb. *J. Chem. Theor. Comput.* 11, 3696–3713. doi:10.1021/acs.jctc.5b00255

Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., and De Vries, A. H. (2007). The Martini Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B.* 111, 7812–7824. doi:10.1021/jp071097f

Masureel, M., Zou, Y., Picard, L.-P., van der Westhuizen, E., Mahoney, J. P., Rodrigues, J. P. G. L. M., et al. (2018). Structural Insights into Binding Specificity, Efficacy and Bias of a β2AR Partial Agonist. *Nat. Chem. Biol.* 14, 1059–1066. doi:10.1038/s41589-018-0145-x

Maupetit, J., Tuffery, P., and Derreumaux, P. (2007). A Coarse-Grained Protein Force Field for Folding and Structure Prediction. *Proteins*. 69, 394–408. doi:10.1002/prot.21505

Menichetti, R., Pelissetto, A., and Randisi, F. (2017). Thermodynamics of star Polymer Solutions: A Coarse-Grained Study. *J. Chem. Phys.* 146, 244908. doi:10.1063/1.4989476

Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., and Beckstein, O. (2011). Mdanalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* 32, 2319–2327. doi:10.1002/jcc.21787

Micheletti, C., Carloni, P., and Maritan, A. (2004). Accurate and Efficient Description of Protein Vibrational Dynamics: Comparing Molecular Dynamics and Gaussian Models. *Proteins*. 55, 635–645. doi:10.1002/prot.20049

Miyazawa, S., and Jernigan, R. L. (1996). Residue - Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *J. Mol. Biol.* 256, 623–644. doi:10.1006/jmbi.1996.0114

Mones, L., Jones, A., Götz, A. W., Laino, T., Walker, R. C., Leimkuhler, B., et al. (2015). The Adaptive Buffered Force QM/MM Method in the Cp2k and Amber Software Packages. *J. Comput. Chem.* 36, 633–648. doi:10.1002/jcc.23839

Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., and Marrink, S.-J. (2008). The Martini Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theor. Comput.* 4, 819–834. doi:10.1021/ct700324x

Morra, G., Potestio, R., Micheletti, C., and Colombo, G. (2012). Corresponding Functional Dynamics across the Hsp90 Chaperone Family: Insights from a Multiscale Analysis of Md Simulations. *Plos Comput. Biol.* 8, e1002433. doi:10.1371/journal.pcbi.1002433

Mullinax, J., and Noid, W. G. (2009a). Generalized Yvon-Born-green Theory for Molecular Systems. *Phys. Rev. Lett.* 103, 198104. doi:10.1103/physrevlett.103.198104

Mullinax, J. W., and Noid, W. G. (2009b). Extended Ensemble Approach for Deriving Transferable Coarse-Grained Potentials. *J. Chem. Phys.* 131, 104110. doi:10.1063/1.3220627

Najafi, S., and Potestio, R. (2015). Folding of Small Knotted Proteins: Insights from a Mean Field Coarse-Grained Model. *J. Chem. Phys.* 143, 12B606_1. doi:10.1063/1.4934541

Nasica-Labouze, J., and Mousseau, N. (2012). Kinetics of Amyloid Aggregation: a Study of the Gnnqqny Prion Sequence. *Plos Comput. Biol.* 8, e1002782. doi:10.1371/journal.pcbi.1002782

Neri, M., Anselmi, C., Cascella, M., Maritan, A., and Carloni, P. (2005). Coarse-grained Model of Proteins Incorporating Atomistic Detail of the Active Site. *Phys. Rev. Lett.* 95, 218102. doi:10.1103/physrevlett.95.218102

Neri, M., Anselmi, C., Carnevale, V., Vargiu, A. V., and Carloni, P. (2006). Molecular Dynamics Simulations of Outer-Membrane Protease T fromE. Colibased on a Hybrid Coarse-Grained/atomistic Potential. *J. Phys. Condens. Matter*. 18, S347–S355. doi:10.1088/0953-8984/18/14/s16

Nguyen, H. D., and Hall, C. K. (2004). Molecular Dynamics Simulations of Spontaneous Fibril Formation by Random-Coil Peptides. *Proc. Natl. Acad. Sci.* 101, 16180–16185. doi:10.1073/pnas.0407273101

Nguyen, H. D., Reddy, V. S., and Brooks III, C. L., Iii (2009). Invariant Polymorphism in Virus Capsid Assembly. *J. Am. Chem. Soc.* 131, 2606–2614. doi:10.1021/ja807730x

Nguyen, P. H., Ramamoorthy, A., Sahoo, B. R., Zheng, J., Faller, P., Straub, J. E., et al. (2021). Amyloid Oligomers: A Joint Experimental/computational Perspective on Alzheimer's Disease, Parkinson's Disease, Type Ii Diabetes, and Amyotrophic Lateral Sclerosis. *Chem. Rev.* 121, 2545-2647. doi:10.1021/acs.chemrev.0c01122

Nguyen, P. H., and Derreumaux, P. (2020). Structures of the Intrinsically Disordered Aβ, Tau and α-synuclein Proteins in Aqueous Solution from Computer Simulations. *Biophysical Chem.* 264, 106421. doi:10.1016/j.bpc.2020.106421

Noid, W. G. (2013). Perspective: Coarse-Grained Models for Biomolecular Systems. *J. Chem. Phys.* 139, 09B201_1. doi:10.1063/1.4818908

Noid, W. G., Chu, J.-W., Ayton, G. S., Krishna, V., Izvekov, S., Voth, G. A., et al. (2008a). The Multiscale Coarse-Graining Method. I. A Rigorous Bridge between Atomistic and Coarse-Grained Models. *J. Chem. Phys.* 128, 244114. doi:10.1063/1.2938860

Noid, W. G., Liu, P., Wang, Y., Chu, J.-W., Ayton, G. S., Izvekov, S., et al. (2008b). The Multiscale Coarse-Graining Method. Ii. Numerical Implementation for Coarse-Grained Molecular Models. *J. Chem. Phys.* 128, 244115. doi:10.1063/1.2938857

O'Meara, M. J., Leaver-Fay, A., Tyka, M. D., Stein, A., Houlihan, K., DiMaio, F., et al. (2015). Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *J. Chem. Theor. Comput.* 11, 609–622. doi:10.1021/ct500864r

Ohkuma, T., and Kremer, K. (2017). Comparison of Two Coarse-Grained Models of Cis -polyisoprene with and without Pressure Correction. *Polymer*. 130, 88–101. doi:10.1016/j.polymer.2017.09.062

Okazaki, K.-i., Koga, N., Takada, S., Onuchic, J. N., and Wolynes, P. G. (2006). Multiple-basin Energy Landscapes for Large-Amplitude Conformational Motions of Proteins: Structure-Based Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci.* 103, 11844–11849. doi:10.1073/pnas.0604375103

Oliver, R. C., Read, D. J., Harlen, O. G., and Harris, S. A. (2013). A Stochastic Finite Element Model for the Dynamics of Globular Macromolecules. *J. Comput. Phys.* 239, 147–165. doi:10.1016/j.jcp.2012.12.027

Palermo, G., Magistrato, A., Riedel, T., von Erlach, T., Davey, C. A., Dyson, P. J., et al. (2016). Fighting Cancer with Transition Metal Complexes: from Naked Dna to Protein and Chromatin Targeting Strategies. *ChemMedChem*. 11, 1199–1210. doi:10.1002/cmdc.201500478

Pan, A. C., Jacobson, D., Yatsenko, K., Sritharan, D., Weinreich, T. M., and Shaw, D. E. (2019). Atomic-level Characterization of Protein-Protein Association. *Proc. Natl. Acad. Sci. USA*. 116, 4244–4249. doi:10.1073/pnas.1815431116

Pan, A. C., Weinreich, T. M., Piana, S., and Shaw, D. E. (2016). Demonstrating an Order-Of-Magnitude Sampling Enhancement in Molecular Dynamics Simulations of Complex Protein Systems. *J. Chem. Theor. Comput.* 12, 1360–1367. doi:10.1021/acs.jctc.5b00913

Patel, J. S., and Ytreberg, F. M. (2018). Fast Calculation of Protein-Protein Binding Free Energies Using Umbrella Sampling with a Coarse-Grained Model. *J. Chem. Theor. Comput.* 14, 991–997. doi:10.1021/acs.jctc.7b00660

Perego, C., and Potestio, R. (2019). Searching the Optimal Folding Routes of a Complex Lasso Protein. *Biophysical J.* 117, 214–228. doi:10.1016/j.bpj.2019.05.025

Periole, X., Knepp, A. M., Sakmar, T. P., Marrink, S. J., and Huber, T. (2012). Structural Determinants of the Supramolecular Organization of G Protein-Coupled Receptors in Bilayers. *J. Am. Chem. Soc.* 134, 10959–10965. doi:10.1021/ja303286e

Petrone, P., and Pande, V. S. (2006). Can Conformational Change Be Described by Only a Few normal Modes?. *Biophysical J.* 90, 1583–1593. doi:10.1529/biophysj.105.070045

Petsev, N. D., Leal, L. G., and Shell, M. S. (2015). Hybrid Molecular-Continuum Simulations Using Smoothed Dissipative Particle Dynamics. *J. Chem. Phys.* 142, 044101. doi:10.1063/1.4905720

Phillips, J. C., Hardy, D. J., Maia, J. D., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., et al. (2020). Scalable Molecular Dynamics on Cpu and Gpu Architectures with Namd. *J. Chem. Phys.* 153, 044130. doi:10.1063/5.0014475

Polles, G., Indelicato, G., Potestio, R., Cermelli, P., Twarock, R., and Micheletti, C. (2013). Mechanical and Assembly Units of Viral Capsids Identified via Quasi-Rigid Domain Decomposition. *Plos Comput. Biol.* 9, e1003331. doi:10.1371/journal.pcbi.1003331

Poma, A. B., Cieplak, M., and Theodorakis, P. E. (2017). Combining the Martini and Structure-Based Coarse-Grained Approaches for the Molecular Dynamics Studies of Conformational Transitions in Proteins. *J. Chem. Theor. Comput.* 13, 1366–1374. doi:10.1021/acs.jctc.6b00986

Potestio, R., Fritsch, S., Espanol, P., Delgado-Buscalioni, R., Kremer, K., Everaers, R., et al. (2013). Hamiltonian Adaptive Resolution Simulation for Molecular Liquids. *Phys. Rev. Lett.* 110, 108301. doi:10.1103/physrevlett.110.108301

Potestio, R., Pontiggia, F., and Micheletti, C. (2009). Coarse-grained Description of Protein Internal Dynamics: an Optimal Strategy for Decomposing Proteins in Rigid Subunits. *Biophysical J.* 96, 4993–5002. doi:10.1016/j.bpj.2009.03.051

Praprotnik, M., Delle Site, L., and Kremer, K. (2005). Adaptive Resolution Molecular-Dynamics Simulation: Changing the Degrees of freedom on the Fly. *J. Chem. Phys.* 123, 224106. doi:10.1063/1.2132286

Ramis, R., Ortega-Castro, J., Casasnovas, R., Mariño, L., Vilanova, B., Adrover, M., et al. (2019). A Coarse-Grained Molecular Dynamics Approach to the Study of the Intrinsically Disordered Protein α-Synuclein. *J. Chem. Inf. Model.* 59, 1458–1471. doi:10.1021/acs.jcim.8b00921

Ranaghan, K. E., Hung, J. E., Bartlett, G. J., Mooibroek, T. J., Harvey, J. N., Woolfson, D. N., et al. (2014). A Catalytic Role for Methionine Revealed by a Combination of Computation and Experiments on Phosphite Dehydrogenase. *Chem. Sci.* 5, 2191–2199. doi:10.1039/c3sc53009d

Rapaport, D. C. (1978). Molecular Dynamics Simulation of Polymer Chains with Excluded Volume. *J. Phys. A: Math. Gen.* 11, L213–L217. doi:10.1088/0305-4470/11/8/008

Richardson, R. A., Hanson, B. S., Read, D. J., Harlen, O. G., and Harris, S. A. (2020). Exploring the Dynamics of Flagellar Dynein within the Axoneme with Fluctuating Finite Element Analysis. *Q. Rev. Biophys.* 53. doi:10.1017/s0033583520000062

Richardson, R. A., Papachristos, K., Read, D. J., Harlen, O. G., Harrison, M., Paci, E., et al. (2014). Understanding the Apparent Stator-Rotor Connections in the Rotary ATPase Family Using Coarse-Grained Computer Modeling. *Proteins.* 82, 3298–3311. doi:10.1002/prot.24680

Robustelli, P., Piana, S., and Shaw, D. E. (2018). Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. *Proc. Natl. Acad. Sci. USA.* 115, E4758–E4766. doi:10.1073/pnas.1800690115

Rojas, A., Liwo, A., Browne, D., and Scheraga, H. A. (2010). Mechanism of Fiber Assembly: Treatment of Aβ Peptide Aggregation with a Coarse-Grained United-Residue Force Field. *J. Mol. Biol.* 404, 537–552. doi:10.1016/j.jmb.2010.09.057

Rojas, A., Maisuradze, N., Kachlishvili, K., Scheraga, H. A., and Maisuradze, G. G. (2017). Elucidating Important Sites and the Mechanism for Amyloid Fibril Formation by Coarse-Grained Molecular Dynamics. *ACS Chem. Neurosci.* 8, 201–209. doi:10.1021/acschemneuro.6b00331

Rosenberger, D., Sanyal, T., Shell, M. S., and van der Vegt, N. F. A. (2019). Transferability of Local Density-Assisted Implicit Solvation Models for Homogeneous Fluid Mixtures. *J. Chem. Theor. Comput.* 15, 2881–2895. doi:10.1021/acs.jctc.8b01170

Rudzinski, J. F., and Noid, W. G. (2015). A Generalized-Yvon-Born-green Method for Coarse-Grained Modeling. *Eur. Phys. J. Spec. Top.* 224, 2193–2216. doi:10.1140/epjst/e2015-02408-9

Rudzinski, J. F., and Noid, W. G. (2011). Coarse-graining Entropy, Forces, and Structures. *J. Chem. Phys.* 135, 214101. doi:10.1063/1.3663709

Rudzinski, J. F., and Noid, W. G. (2014). Investigation of Coarse-Grained Mappings via an Iterative Generalized Yvon-Born-Green Method. *J. Phys. Chem. B.* 118, 8295–8312. doi:10.1021/jp501694z

Sanyal, T., and Shell, M. S. (2016). Coarse-grained Models Using Local-Density Potentials Optimized with the Relative Entropy: Application to Implicit Solvation. *J. Chem. Phys.* 145, 034109. doi:10.1063/1.4958629

Sanyal, T., and Shell, M. S. (2018). Transferable Coarse-Grained Models of Liquid-Liquid Equilibrium Using Local Density Potentials Optimized with the Relative Entropy. *J. Phys. Chem. B.* 122, 5678–5693. doi:10.1021/acs.jpcb.7b12446

Saunders, M. G., and Voth, G. A. (2012). Coarse-graining of Multiprotein Assemblies. *Curr. Opin. Struct. Biol.* 22, 144–150. doi:10.1016/j.sbi.2012.01.003

Schneider, J., Ribeiro, R., Alfonso-Prieto, M., Carloni, P., and Giorgetti, A. (2020). Hybrid MM/CG Webserver: Automatic Set up of Molecular Mechanics/coarse-Grained Simulations for Human G Protein-Coupled Receptor/ligand Complexes. *Front. Mol. Biosciences.* 7, 232. doi:10.3389/fmolb.2020.576689

Schneider, J., Korshunova, K., Musiani, F., Alfonso-Prieto, M., Giorgetti, A., and Carloni, P. (2018). Predicting Ligand Binding Poses for Low-Resolution Membrane Protein Models: Perspectives from Multiscale Simulations. *Biochem. biophysical Res. Commun.* 498, 366–374. doi:10.1016/j.bbrc.2018.01.160

Sept, D., and MacKintosh, F. C. (2010). Microtubule Elasticity: Connecting All-Atom Simulations with Continuum Mechanics. *Phys. Rev. Lett.* 104, 018101. doi:10.1103/physrevlett.104.018101

Shahidi, N., Chazirakis, A., Harmandaris, V., and Doxastakis, M. (2020). Coarse-graining of Polyisoprene Melts Using Inverse Monte Carlo and Local Density Potentials. *J. Chem. Phys.* 152, 124902. doi:10.1063/1.5143245

Shao, J., Tanner, S. W., Thompson, N., and Cheatham, T. E. (2007). Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theor. Comput.* 3, 2312–2334. doi:10.1021/ct700119m

Shaw, D. E., Dror, R. O., Salmon, J. K., Grossman, J., Mackenzie, K. M., Bank, J. A., et al. (2009). Millisecond-scale Molecular Dynamics Simulations on anton. In Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, 1–11.

Shaw, D. E., Grossman, J., Bank, J. A., Batson, B., Butts, J. A., Chao, J. C., et al. (2014). Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. In SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. . IEEE, 41–53.

Shaw, D. E., Deneroff, M. M., Dror, R. O., Kuskin, J. S., Larson, R. H., Salmon, J. K., et al. (2008). Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *Commun. ACM.* 51, 91–97. doi:10.1145/1364782.1364802

Shell, M. S. (2008). The Relative Entropy Is Fundamental to Multiscale and Inverse Thermodynamic Problems. *J. Chem. Phys.* 129, 144108. doi:10.1063/1.2992060

Shell, M. S. (2016). Coarse-graining with the Relative Entropy. *Adv. Chem. Phys.* 161, 395–441. doi:10.1002/9781119290971.ch5

Shen, Y., Maupetit, J., Derreumaux, P., and Tufféry, P. (2014). Improved Pep-fold Approach for Peptide and Miniprotein Structure Prediction. *J. Chem. Theor. Comput.* 10, 4745–4758. doi:10.1021/ct500592m

Sieradzan, A. K., Liwo, A., and Hansmann, U. H. E. (2012). Folding and Self-Assembly of a Small Protein Complex. *J. Chem. Theor. Comput.* 8, 3416–3422. doi:10.1021/ct300528r

Sieradzan, A. K., Makowski, M., Augustynowicz, A., and Liwo, A. (2017). A General Method for the Derivation of the Functional Forms of the Effective Energy Terms in Coarse-Grained Energy Functions of Polymers. I. Backbone Potentials of Coarse-Grained Polypeptide Chains. *J. Chem. Phys.* 146, 124106. doi:10.1063/1.4978680

Singh, N., and Li, W. (2019). Recent Advances in Coarse-Grained Models for Biomolecules and Their Applications. *Int J Mol Sci.* 20, 3774. doi:10.3390/ijms20153774

Sinitskiy, A. V., Saunders, M. G., and Voth, G. A. (2012). Optimal Number of Coarse-Grained Sites in Different Components of Large Biomolecular Complexes. *J. Phys. Chem. B.* 116, 8363–8374. doi:10.1021/jp2108895

Sittel, F., and Stock, G. (2018). Perspective: Identification of Collective Variables and Metastable States of Protein Dynamics. *J. Chem. Phys.* 149, 150901. doi:10.1063/1.5049637

Solernou, A., Hanson, B. S., Richardson, R. A., Welch, R., Read, D. J., Harlen, O. G., et al. (2018). Fluctuating Finite Element Analysis (Ffea): A Continuum Mechanics Software Tool for Mesoscale Simulation of Biomolecules. *PLoS Comput. Biol.* 14, e1005897. doi:10.1371/journal.pcbi.1005897

Soper, A. K. (1996). Empirical Potential Monte Carlo Simulation of Fluid Structure. *Chem. Phys.* 202, 295–306. doi:10.1016/0301-0104(95)00357-6

Souza, P. C., Thallmair, S., Conflitti, P., Ramírez-Palacios, C., Alessandri, R., Raniolo, S., et al. (2020). Protein–ligand Binding with the Coarse-Grained Martini Model. *Nat. Commun.* 11, 1–11. doi:10.1038/s41467-020-17437-5

Spagnolli, G., Rigoli, M., Orioli, S., Sevillano, A. M., Faccioli, P., Wille, H., et al. (2019). Full Atomistic Model of Prion Structure and Conversion. *PLoS Pathog.* 15. doi:10.1371/journal.ppat.1007864

Spagnolli, G., Rigoli, M., Novi Inverardi, G., Codeseira, Y. B., Biasini, E., and Requena, J. R. (2020). Modeling Prpsc Generation through Deformed Templating. *Front. Bioeng. Biotechnol.* 8, 1165. doi:10.3389/fbioe.2020.590501

Spiga, E., Alemani, D., Degiacomi, M. T., Cascella, M., and Dal Peraro, M. (2013). Electrostatic-consistent Coarse-Grained Potentials for Molecular Simulations of Proteins. *J. Chem. Theor. Comput.* 9, 3515–3526. doi:10.1021/ct400137q

Sterpone, F., Derreumaux, P., and Melchionna, S. (2015). Protein Simulations in Fluids: Coupling the Opep Coarse-Grained Force Field with Hydrodynamics. *J. Chem. Theor. Comput.* 11, 1843–1853. doi:10.1021/ct501015h

Sterpone, F., Derreumaux, P., and Melchionna, S. (2018). Molecular Mechanism of Protein Unfolding under Shear: A Lattice Boltzmann Molecular Dynamics Study. *J. Phys. Chem. B* 122, 1573–1579. doi:10.1021/acs.jpcb.7b10796

Sterpone, F., Nguyen, P. H., Kalimeri, M., and Derreumaux, P. (2013). Importance of the Ion-Pair Interactions in the Opep Coarse-Grained Force Field: Parametrization and Validation. *J. Chem. Theor. Comput.* 9, 4574–4584. doi:10.1021/ct4003493

Sterpone, F., Melchionna, S., Tuffery, P., Pasquali, S., Mousseau, N., Cragnolini, T., et al. (2014). The Opep Protein Model: from Single Molecules, Amyloid Formation, Crowding and Hydrodynamics to Dna/rna Systems. *Chem. Soc. Rev.* 43, 4871–4893. doi:10.1039/c4cs00048j

Stone, J. E., Hardy, D. J., Ufimtsev, I. S., and Schulten, K. (2010). Gpu-accelerated Molecular Modeling Coming of Age. *J. Mol. Graphics Model.* 29, 116–125. doi:10.1016/j.jmgm.2010.06.010

Sweet, J. C., Nowling, R. J., Cickovski, T., Sweet, C. R., Pande, V. S., and Izaguirre, J. A. (2013). Long Timestep Molecular Dynamics on the Graphical Processing Unit. *J. Chem. Theor. Comput.* 9, 3267–3281. doi:10.1021/ct400331r

Szklarczyk, O. M., Bieler, N. S., Hünenberger, P. H., and van Gunsteren, W. F. (2015). Flexible Boundaries for Multiresolution Solvation: an Algorithm for Spatial Multiscaling in Molecular Dynamics Simulations. *J. Chem. Theor. Comput.* 11, 5447–5463. doi:10.1021/acs.jctc.5b00406

Takada, S. (2019). Gō Model Revisited. *Biophysics*. 16, 248–255. doi:10.2142/biophysico.16.0_248

Taketomi, H., Ueda, Y., and Gō, N. (1975). Studies on Protein Folding, Unfolding and Fluctuations by Computer Simulation. I. The Effect of Specific Amino Acid Sequence Represented by Specific Inter-unit Interactions. *Int. J. Pept. Protein Res.* 7, 445–459.

Tama, F., and Brooks, C. L., III (2005). Diversity and Identity of Mechanical Properties of Icosahedral Viral Capsids Studied with Elastic Network normal Mode Analysis. *J. Mol. Biol.* 345, 299–314. doi:10.1016/j.jmb.2004.10.054

Tama, F., and Sanejouand, Y.-H. (2001). Conformational Change of Proteins Arising from normal Mode Calculations. *Protein Eng.* 14, 1–6. doi:10.1093/protein/14.1.1

Tama, F., Valle, M., Frank, J., and Brooks, C. L. (2003). Dynamic Reorganization of the Functionally Active Ribosome Explored by normal Mode Analysis and Cryo-Electron Microscopy. *Proc. Natl. Acad. Sci.* 100, 9319–9323. doi:10.1073/pnas.1632476100

Tanaka, S., and Scheraga, H. A. (1976). Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. *Macromolecules*. 9, 945–950. doi:10.1021/ma60054a013

Tarenzi, T., Calandrini, V., Potestio, R., and Carloni, P. (2019). Open-Boundary Molecular Mechanics/Coarse-Grained Framework for Simulations of Low-Resolution G-Protein-Coupled Receptor-Ligand Complexes. *J. Chem. Theor. Comput.* 15, 2101–2109. doi:10.1021/acs.jctc.9b00040

Tarenzi, T., Calandrini, V., Potestio, R., Giorgetti, A., and Carloni, P. (2017). Open Boundary Simulations of Proteins and Their Hydration Shells by Hamiltonian Adaptive Resolution Scheme. *J. Chem. Theor. Comput.* 13, 5647–5657. doi:10.1021/acs.jctc.7b00508

Tirion, M. M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* 77, 1905–1908. doi:10.1103/physrevlett.77.1905

Togashi, Y., and Flechsig, H. (2018). Coarse-grained Protein Dynamics Studies Using Elastic Network Models. *Int J Mol Sci.* 19, 3899. doi:10.3390/ijms19123899

Tribello, G. A., and Gasparotto, P. (2019). Using Dimensionality Reduction to Analyze Protein Trajectories. *Front. Mol. biosciences*. 6, 46. doi:10.3389/fmolb.2019.00046

Truong, H. H., Kim, B. L., Schafer, N. P., and Wolynes, P. G. (2015). Predictive Energy Landscapes for Folding Membrane Protein Assemblies. *J. Chem. Phys.* 143, 243101. doi:10.1063/1.4929598

Tschöp, W., Kremer, K., Batoulis, J., Bürger, T., and Hahn, O. (1998). Simulation of Polymer Melts. I. Coarse-Graining Procedure for Polycarbonates. *Acta Polym.* 49, 61–74. doi:10.1002/(sici)1521-4044(199802)49:2/3<61::aid-apol61>3.0.co;2-v

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. Inf. Model.* 56, 2180–2193. doi:10.1021/acs.jcim.6b00233

van der Heijden, T. W. G., Read, D. J., Harlen, O. G., van der Schoot, P., Harris, S. A., and Storm, C. (2020). Combined Force-Torque Spectroscopy of Proteins by Means of Multiscale Molecular Simulation. *Biophysical J.* 119, 2240–2250. doi:10.1016/j.bpj.2020.09.039

Villa, E., Balaeff, A., Mahadevan, L., and Schulten, K. (2004). Multiscale Method for Simulating Protein-Dna Complexes. *Multiscale Model. Simul.* 2, 527–553. doi:10.1137/040604789

Villa, E., Balaeff, A., and Schulten, K. (2005). Structural Dynamics of the Lac Repressor-DNA Complex Revealed by a Multiscale Simulation. *Proc. Natl. Acad. Sci.* 102, 6783–6788. doi:10.1073/pnas.0409387102

Voegler Smith, A., and Hall, C. K. (2001). ?-Helix Formation: Discontinuous Molecular Dynamics on an Intermediate-Resolution Protein Model. *Proteins*. 44, 344–360. doi:10.1002/prot.1100

Wagner, J. W., Dama, J. F., Durumeric, A. E. P., and Voth, G. A. (2016). On the Representability Problem and the Physical Meaning of Coarse-Grained Models. *J. Chem. Phys.* 145, 044108. doi:10.1063/1.4959168

Wagner, J. W., Dannenhoffer-Lafage, T., Jin, J., and Voth, G. A. (2017). Extending the Range and Physical Accuracy of Coarse-Grained Models: Order Parameter Dependent Interactions. *J. Chem. Phys.* 147, 044113. doi:10.1063/1.4995946

Wagoner, J. A., and Pande, V. S. (2011). A Smoothly Decoupled Particle Interface: New Methods for Coupling Explicit and Implicit Solvent. *J. Chem. Phys.* 134, 214103. doi:10.1063/1.3595262

Wagoner, J. A., and Pande, V. S. (2013). Finite Domain Simulations with Adaptive Boundaries: Accurate Potentials and Nonequilibrium Movesets. *J. Chem. Phys.* 139, 12B616_1. doi:10.1063/1.4848655

Wagoner, J. A., and Pande, V. S. (2018). Communication: Adaptive Boundaries in Multiscale Simulations. *J. Chem. Phys.* 148, 141104. doi:10.1063/1.5025826

Wang, J., Olsson, S., Wehmeyer, C., Pérez, A., Charron, N. E., De Fabritiis, G., et al. (2019). Machine Learning of Coarse-Grained Molecular Dynamics Force fields. *ACS Cent. Sci.* 5, 755–767. doi:10.1021/acscentsci.8b00913

Wang, W., and Gómez-Bombarelli, R. (2019). Coarse-graining Auto-Encoders for Molecular Dynamics. *npj Comput. Mater.* 5, 1–9. doi:10.1038/s41524-019-0261-5

Wang, Y., and Hall, C. K. (2018). Seeding and Cross-Seeding Fibrillation of N-Terminal Prion Protein Peptides PrP(120-144). *Protein Sci.* 27, 1304–1313. doi:10.1002/pro.3421

Wang, Y., Latshaw, D. C., and Hall, C. K. (2017). Aggregation of Aβ(17-36) in the Presence of Naturally Occurring Phenolic Inhibitors Using Coarse-Grained Simulations. *J. Mol. Biol.* 429, 3893–3908. doi:10.1016/j.jmb.2017.10.006

Wang, Y., and Zocchi, G. (2011). The Folded Protein as a Viscoelastic Solid. *Epl*. 96, 18003. doi:10.1209/0295-5075/96/18003

Warshel, A., and Karplus, M. (1972). Calculation of Ground and Excited State Potential Surfaces of Conjugated Molecules. I. Formulation and Parametrization. *J. Am. Chem. Soc.* 94, 5612–5625. doi:10.1021/ja00771a014

Warshel, A., and Levitt, M. (1976). Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol.* 103, 227–249. doi:10.1016/0022-2836(76)90311-9

Wei, G., Mousseau, N., and Derreumaux, P. (2004). Complex Folding Pathways in a Simple β-hairpin. *Proteins*. 56, 464–474. doi:10.1002/prot.20127

Welch, R., Harris, S. A., Harlen, O. G., and Read, D. J. (2020). Kobra: a Fluctuating Elastic Rod Model for Slender Biological Macromolecules. *Soft Matter*. 16, 7544–7555. doi:10.1039/d0sm00491j

Whitford, P. C., Noel, J. K., Gosavi, S., Schug, A., Sanbonmatsu, K. Y., and Onuchic, J. N. (2009). An All-Atom Structure-Based Potential for Proteins: Bridging

Minimal Models with All-Atom Empirical Forcefields. *Proteins*. 75, 430–441. doi:10.1002/prot.22253

Wolf, A., and Kirschner, K. N. (2013). Principal Component and Clustering Analysis on Molecular Dynamics Data of the Ribosomal L11·23S Subdomain. *J. Mol. Model.* 19, 539–549. doi:10.1007/s00894-012-1563-4

Wu, H., Wolynes, P. G., and Papoian, G. A. (2018). Awsem-idp: a Coarse-Grained Force Field for Intrinsically Disordered Proteins. *J. Phys. Chem. B.* 122, 11115–11125. doi:10.1021/acs.jpcb.8b05791

Wu, Z., Zhang, Y., Zhang, J. Z., Xia, K., and Xia, F. (2020). Determining Optimal Coarse-Grained Representation for Biomolecules Using Internal Cluster Validation Indexes. *J. Comput. Chem.* 41, 14–20. doi:10.1002/jcc.26070

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

Zavadlav, J., Marrink, S. J., and Praprotnik, M. (2019). Swinger: a Clustering Algorithm for Concurrent Coupling of Atomistic and Supramolecular Liquids. *Interf. Focus.* 9, 20180075. doi:10.1098/rsfs.2018.0075

Zavadlav, J., Sablić, J., Podgornik, R., and Praprotnik, M. (2018). Open-boundary Molecular Dynamics of a Dna Molecule in a Hybrid Explicit/implicit Salt Solution. *Biophysical J.* 114, 2352–2362. doi:10.1016/j.bpj.2018.02.042

Zhang, L., Han, J., Wang, H., Car, R., and Weinan, E. (2018). Deepcg: Constructing Coarse-Grained Models via Deep Neural Networks. *J. Chem. Phys.* 149, 034101. doi:10.1063/1.5027645

Zhang, Y., Cao, Z., and Xia, F. (2017). Construction of Ultra-coarse-grained Model of Protein with a Gō-like Potential. *Chem. Phys. Lett.* 681, 1–6. doi:10.1016/j.cplett.2017.05.039

Zhang, Y., Cao, Z., Zhang, J. Z., and Xia, F. (2020). Double-well Ultra-coarse-grained Model to Describe Protein Conformational Transitions. *J. Chem. Theor. Comput.* 16, 6678–6689. doi:10.1021/acs.jctc.0c00551

Zhang, Z., Lu, L., Noid, W. G., Krishna, V., Pfaendtner, J., and Voth, G. A. (2008). A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules. *Biophysical J.* 95, 5073–5083. doi:10.1529/biophysj.108.139626

Zhang, Z., Pfaendtner, J., Grafmüller, A., and Voth, G. A. (2009). Defining Coarse-Grained Representations of Large Biomolecules and Biomolecular Complexes from Elastic Network Models. *Biophysical J.* 97, 2327–2337. doi:10.1016/j.bpj.2009.08.007

Zhang, Z., and Voth, G. A. (2010). Coarse-grained Representations of Large Biomolecular Complexes from Low-Resolution Structural Data. *J. Chem. Theor. Comput.* 6, 2990–3002. doi:10.1021/ct100374a

Zheng, W., Schafer, N. P., Davtyan, A., Papoian, G. A., and Wolynes, P. G. (2012). Predictive Energy Landscapes for Protein-Protein Association. *Proc. Natl. Acad. Sci.* 109, 19244–19249. doi:10.1073/pnas.1216215109

Zheng, W., Tsai, M.-Y., Chen, M., and Wolynes, P. G. (2016). Exploring the Aggregation Free Energy Landscape of the Amyloid-β Protein (1-40). *Proc. Natl. Acad. Sci. USA*. 113, 11835–11840. doi:10.1073/pnas.1612362113

Reith, D., Pütz, M., and Müller-Plathe, F. (2003). Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.*. 24, 1624–1636. doi:10.1002/jcc.10307

Rosenberger, D., Hanke, M., and van der Vegt, N. F. A. (2016). Comparison of iterative inverse coarse-graining methods. *Eur. Phsy. J. Special Topics*. 225, 1323–1345. doi:10.1140/epjst/e2016-60120-1

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 © 2021 Giulini, Rigoli, Mattiotti, Menichetti, Tarenzi, Fiorentini and Potestio. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 11 June 2021
doi: 10.3389/fmolb.2021.689400

Thermodynamic Origin of Differential Excipient-Lysozyme Interactions

Jas Kalayan[1,2]*, *Robin A. Curtis*[1,3], *Jim Warwicker*[1,4] and *Richard H. Henchman*[1,2,5]

[1]Manchester Institute of Biotechnology, The University of Manchester, Manchester, United Kingdom, [2]Department of Chemistry, The University of Manchester, Manchester, United Kingdom, [3]Departments of Chemical Engineering and Analytical Science, The University of Manchester, Manchester, United Kingdom, [4]Division of Molecular and Cellular Function, School of Biological Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, United Kingdom, [5]Faculty of Medicine and Health, Sydney Medical School, The University of Sydney, Sydney, NSW, Australia

Understanding the intricate interplay of interactions between proteins, excipients, ions and water is important to achieve the effective purification and stable formulation of protein therapeutics. The free energy of lysozyme interacting with two kinds of polyanionic excipients, citrate and tripolyphosphate, together with sodium chloride and TRIS-buffer, are analysed in multiple-walker metadynamics simulations to understand why tripolyphosphate causes lysozyme to precipitate but citrate does not. The resulting multiscale decomposition of energy and entropy components for water, sodium chloride, excipients and lysozyme reveals that lysozyme is more stabilised by the interaction of tripolyphosphate with basic residues. This is accompanied by more sodium ions being released into solution from tripolyphosphate than for citrate, whilst the latter instead has more water molecules released into solution. Even though lysozyme aggregation is not directly probed in this study, these different mechanisms are suspected to drive the cross-linking between lysozyme molecules with vacant basic residues, ultimately leading to precipitation.

Keywords: statistical mechanics, entropy, free energy methods, multiscale, metadynamics method, protein-protein binding, protein-excipient binding, protein hydration

OPEN ACCESS

Edited by:
Fabio Trovato,
Freie Universität Berlin, Germany

Reviewed by:
Yandong Huang,
Jimei University, China
Riccardo Nifosì,
National Research Council (CNR), Italy
Giuseppe Pellicane,
University of Messina, Italy

****Correspondence:***
Jas Kalayan
jas.kalayan@manchester.ac.uk

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *31 March 2021*
Accepted: *25 May 2021*
Published: *11 June 2021*

Citation:
Kalayan J, Curtis RA, Warwicker J and Henchman RH (2021) Thermodynamic Origin of Differential Excipient-Lysozyme Interactions. Front. Mol. Biosci. 8:689400. doi: 10.3389/fmolb.2021.689400

1 INTRODUCTION

Protein therapeutics are increasingly being developed in the biopharmaceutical industry to combat a wide range of diseases (Dimitrov, 2012; Faber and Whitehead, 2019). Compared with traditional small drug molecules, the large and complex structures and marginal stability of biomolecules make necessary the development of sophisticated formulations which can stabilise the structure of such therapeutics to ensure their safe administration and efficacy (Wang, 2015). Without the correct formulation conditions, proteins are prone to aggregation, precipitation or phase separation (Moussa et al., 2016). The modulation of protein-protein interactions (PPIs) in formulations is commonly achieved by the addition of small molecules, termed excipients. However, a lack of understanding of how excipients operate is hampering further development because such systems comprise multiple, transient, weak interactions between proteins, excipients, other ions and water in solution (Falconer, 2019). This is an extension of the difficulties in understanding aqueous electrolytes, where behaviour even of simple salt solutions is not well explained, particularly at higher concentrations (Collins, 1997). It is therefore important to develop new strategies to identify the effects of excipients on protein stability to improve therapeutic formulations in the biopharmaceutical industry.

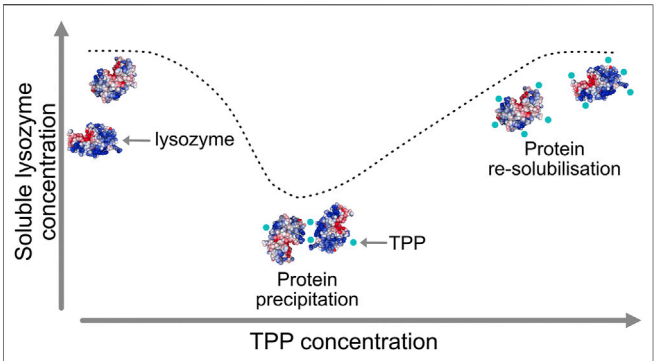


FIGURE 1 | Lysozyme precipitation at intermediate tripolyphosphate (TPP) concentrations and re-solubilisation at higher concentration.

A particularly intriguing phenomenon relating to the mechanism of excipients operation was recently revealed for the case of the protein lysozyme interacting with two excipients, namely tripolyphosphate (TPP) and citrate (CIT). Lysozyme is a small, stable protein that is widely studied and known to remain largely folded in both experiments and simulations and so does not require extensive sampling of protein tertiary structural changes. Experimental data, as illustrated schematically in **Figure 1**, has shown how TPP excipients cause lysozyme to precipitate out of solution and then resolubilise at higher TPP concentration, but similarly charged CIT excipients do not cause lysozyme precipitation at any concentration (Bye and Curtis, 2019).

The reason for these trends is not clearly understood, but it is hypothesised that TPP cross-links lysozyme due to more polarised, multi-directional P-O^- bonds which form stronger interactions with basic residues. Conversely, CIT may have sterically hindered hydrogen bonding with donors on the protein surface, which prevents cross-linking between multiple lysozyme molecules (Bye and Curtis, 2019). In this work we concentrate on the differences between TPP and CIT at the excipient concentration where TPP causes lysozyme precipitation but CIT does not.

Computational methods can contribute much to understanding protein-excipient binding, given the high level of detail that they provide. However, conventional methods for molecular binding are less suitable for such problems because they tend to be aimed at systems in which there are specific binding sites between a substrate and its target, owing to the long-standing influence of structure-based drug design. Rather, it is typically the case for excipients that they bind non-specifically to protein surfaces and in some cases not even directly to proteins at all (Zalar et al., 2020). This diminishes the usefulness of computational techniques such as docking studies and alchemical free-energy methods in favour of ensemble-based methods such as molecular dynamics (MD) or Monte Carlo (MC) simulations that are able to sample the wide range of relevant configurations (Schames et al., 2004; Spyrakis et al., 2015; Yu et al., 2018; Barata et al., 2016). Given the large number of possible configurations that need to be sampled, more efficient enhanced-sampling simulation methods are required. Most enhanced sampling methods for proteins have been conducted to explore protein folding, allostery or protein-ligand binding (Singh et al., 2017; Du et al., 2017; Barducci et al., 2013; Verona et al., 2017). Studies of multiple ligands interacting with a protein surface have shown how increased sampling produces binding affinities comparable to experiment (Troussicot et al., 2015) as well as how co-solutes can affect dissociation of proteins (Banerjee et al., 2019). The use of multiple simulations starting from different poses of two proteins has been shown to be effective in reproducing native protein-protein association structures and in the prediction of new bound configurations (Plattner et al., 2017). Such methods may yield the extent of excipient binding via brute-force probabilities between bound and unbound but are unable to explain the calculated stability for different excipients. Achieving this requires more detailed strategies such as determining how all of the molecules in a system contribute to the total Gibbs free energy. However, most such studies on proteins have focused on the contribution of water molecules or the protein itself (Tarek and Tobias, 2000; Gerogiokas et al., 2016; Chong and Ham, 2017).

Here we seek to understand how lysozyme is differentially stabilised by the excipients TPP and CIT to help explain its aggregation behaviour by applying a free energy method that calculates energy and entropy directly from a simulation called EE-MCC (Energy-Entropy Multiscale Cell Correlation). Energy is calculated directly from the system Hamiltonian by summing over per-atom energies. Entropy is calculated for all molecules in the system from forces and coordinates at multiple length scales using MCC, which has been applied to liquids (Higham et al., 2018; Ali et al., 2019), chemical reactions (Ali et al., 2020), and proteins (Chakravorty et al., 2020). The three length scales employed here from smallest to largest are 1) water and monatomic ions, 2) excipients and residues, and 3) the whole protein, which are classified here as united-atom, monomer, and polymer levels, respectively. We examine the Gibbs free energy for mixing a lysozyme dimer with TRIS buffer and counterions with five excipient molecules, either TPP or CIT, together with counterions, corresponding to the concentration at which TPP-induced aggregation is detected experimentally. Protein-excipient configurations are sampled using metadynamics and multiple-walker simulations to allow the exploration of more transient interactions than would be possible using a conventional molecular dynamics simulation. Although, the large phase space required to sample solute interactions remains a bottle-neck, we observe several important differences between CIT and TPP containing solutions that may help explain their effects on lysozyme aggregation.

2 METHODS

2.1 Multiscale Cell Correlation Entropy Theory

In MCC, entropy S is calculated in a multiscale fashion in terms of cells of correlated units. The total entropy is calculated as a sum of components S_{ij}^{kl} using the equation

$$S = \sum_{i}^{\text{molecule}} \sum_{j}^{\text{level}} \sum_{k}^{\text{type}} \sum_{l}^{\text{motion}} S_{ij}^{kl} \tag{1}$$

In this equation, S is calculated for each kind of molecule i, at the appropriate length scales j for each molecule, in terms of translational or rotational motion l over all units at that level, and in terms of vibration or topography k for each type of motion. Vibrational entropy relates to the average size of energy wells for that unit while topographical entropy relates to the probability distribution of the energy wells. Length scales are defined at the united-atom (UA), monomer (M) and polymer (P) levels for the protein, at the UA and M levels for excipients, and the UA level for water and monatomic ions.

2.1.1 Vibrational Entropy

The entropy of internal molecular vibrations of a unit in the x, y, z directions at each length scale are calculated from the frequencies ν_i of N_{vib} number of translational or rotational vibrations using

$$S^{\text{vib}} = k_{\text{B}} \sum_{i=1}^{N_{\text{vib}}} \left(\frac{h\nu_i / k_{\text{B}}T}{e^{h\nu_i/k_{\text{B}}T} - 1} - \ln\left(1 - e^{-h\nu_i/k_{\text{B}}T}\right) \right) \tag{2}$$

where h is Planck's constant, k_{B} is Boltzmann's constant and T is temperature. Frequencies ν_i are obtained from the eigenvalues λ_i of diagonalised mass-weighted force and inertia-weighted torque covariance matrices using

$$\nu_i = \frac{1}{2\pi} \sqrt{\frac{\lambda_i}{k_{\text{B}}T}} \tag{3}$$

The forces and torques are rotated into appropriate reference frames defined by the type of unit as shown in **Supplementary Table S1**. In the mean-field approximation, forces at the polymer level and torques at all levels are halved. For all but the highest length scale in each molecule, vibrational entropies associated with the six smallest frequencies in the force covariance matrix are removed to avoid double counting translation and rotation at the higher level.

2.1.2 Conformational Entropy

Translational topographical entropy at the united-atom level is otherwise known as conformational entropy. It is calculated from the probabilities p_i of sets of conformers for all dihedrals in the monomer using

$$S_{\text{UA}}^{\text{topo}} = -k_{\text{B}} \sum_{i=1}^{N_{\text{conf}}} p_i \ln p_i \tag{4}$$

where N_{conf} is the number of sets of conformers. Each conformer i is assigned by discretising the distribution of observed dihedral angles into 30° bins, conformers are centered on peaks defined at bins for which two adjacent bins are less populated and no peaks are closer than 60°, and each dihedral angle is assigned to the nearest peak.

2.1.3 Orientational Entropy

The entropy arising from water molecule orientations with respect to neighbouring united atoms is a generalisation from before (Higham et al., 2018) that captures anisotropy due to the presence of solutes. Orientational entropy is calculated using

$$S_{\text{UA}}^{\text{orient}} = k_{\text{B}} \sum_{c} p(c) \ln\left[\left(N_{\text{eff}}\pi\right)^{\frac{3}{2}} p(\text{HB}_{\text{av}}) \big/ \sigma \right] \tag{5}$$

where c is the coordination-shell type of a water molecule and σ is the symmetry number of water, equal to 2. The average bias in hydrogen bonds (HBs) $p(\text{HB}_{\text{av}})$ is a weighted average over the HB biases to all N_c neighbours in the coordination shell

$$p(\text{HB}_{\text{av}}) = \frac{\sum_n p(\text{HB}_i) N_i}{N_c} \tag{6}$$

where N_i is the number of neighbours of type i in the coordination shell and $p(\text{HB}_i)$ is the bias in hydrogen-bonding with neighbour i given by

$$p(\text{HB}_i) = \frac{p(\text{D}_i)}{p(\text{D}_i) + p(\text{A}_i)} \times \frac{p(\text{A}_i)}{p(\text{D}_i) + p(\text{A}_i)} \tag{7}$$

where $p(\text{A}_i)$ is the probability of accepting from neighboring species i over all other neighbours being accepted from and $p(\text{D}_i)$ is the probability of donating to neighbouring species i over all other neighbours being donated to. For bulk water, $p(\text{D}_i) = p(\text{A}_i) = 1$, but when solutes are nearby, the HBs may be biased in one direction, thus reducing the number of orientations of the water molecule. The effective number of neighbours N_{eff} that a water molecule can accept and donate HBs with is calculated from the HB probabilities relative to the unbiased value 0.25 as

$$N_{\text{eff}} = \sum_i \frac{p(\text{HB}_i) N_i}{0.25} \tag{8}$$

Neighbouring water molecules are identified by the solute they are closest to when in the solute coordination shell and labeled as bulk otherwise. Neighbouring solutes are identified as their particular united atoms that are proximal to the water molecule. Contacts between united atoms are defined using the relative angular distance (RAD) algorithm (Higham and Henchman, 2016). HBs are defined topologically (Henchman and Irudayam, 2010; Irudayam et al., 2010; Irudayam and Henchman, 2011) based on the acceptor with the most negative $q\text{D}q\text{A}/r^2$ for each donor, where qD and qA are the charges of the donor and acceptor, respectively, and r is the distance between them. We use a qualitative proxy for positional entropy of the mixing molecules by considering the number of solute-water contacts. We neglect the contribution of the orientational entropy of the excipients, assuming for these weakly binding molecules that it does not change significantly.

2.2 Molecular Energy

The potential (U) and kinetic (K) energies per atom i from simulation are summed to give the enthalpy H of a given unit

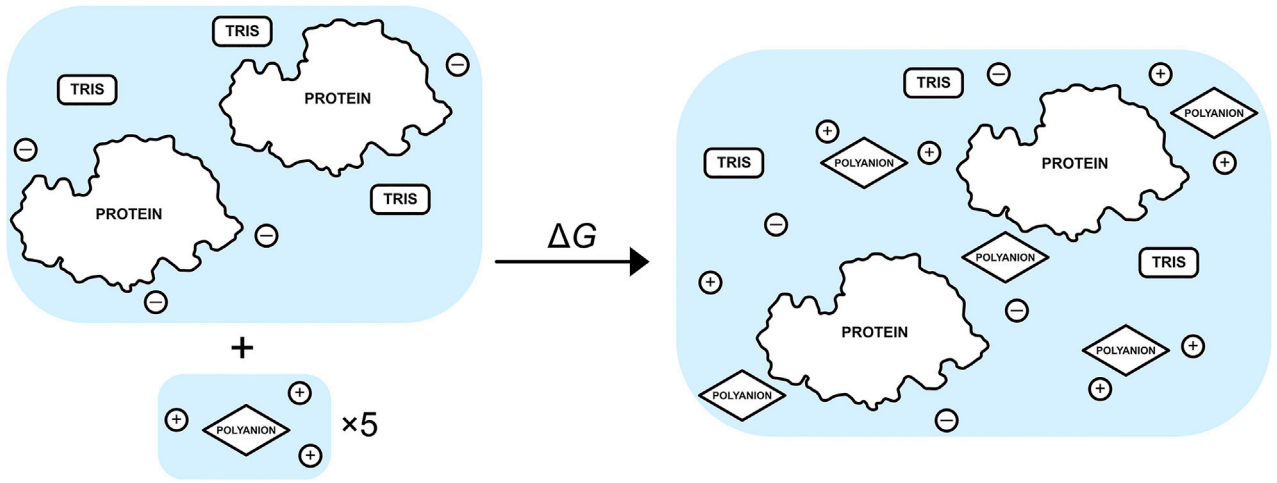


FIGURE 2 | Schematic of the binding process between a set of five excipients, each with neutralising Na^+ ions, and a buffered protein-dimer system with neutralising Cl^-.

$$H \approx \sum_i (K_i + U_i) \tag{9}$$

ignoring the negligible pressure-volume term.

2.3 Water Molecule Assessment Based on Coordination Shell Neighbours

The immediate environment of a water molecule is defined by what molecules are in its first coordination shell. This is used to assess the free energy of water molecules depending on what solutes they interact with. As there are many possible local configurations of a coordination shell, we simplify the definition of local water environments by defining five categories of water contacts with various combinations of excipients and counterion or lysozyme molecules as follows:

1) W_P: UAs from only one protein.
2) W_{PP}: UAs from two proteins.
3) W_{EP}: UAs of one protein and any excipient or counterion.
4) W_{EPP}: UAs of two proteins and any excipient or counterion.
5) W_E: UAs of any excipient or counterion.

2.4 Change in Free Energy for Excipient Binding

The free-energy change for binding five TPP or CIT excipient molecules to the buffered protein relative to an infinitely dilute excipient with their neutralising counterions is calculated using

$$\Delta G = \sum_X \left[G_{X,bnd} N_X + G_{W,bnd} N_{W_X,bnd} + G_{W,bulk} \left(N_{W_X,dil} - N_{W_X,bnd} \right) \right] - \left[G_{X,dil} N_{X,dil} + G_{W,dil} N_{W_X,dil} \right] \tag{10}$$

where X is each type of solute, N_X is the number of solute molecule X, and N_{W_X} is the number of water molecules in the hydration shell of solute X. The subscripts "bnd", "dil" and "bulk" refer to the bound or unbound dilute solutions and bulk water, respectively. The unbound protein dimer comprises chloride ions and five TRIS molecules (2-amino-2-hydroxymethyl-propane-1,3-diol) as shown in **Figure 2**. The five unbound excipients comprise a single dilute excipient molecule with three neutralising sodium counterions. The number of water molecules that are released from solvation shells go into bulk, and so they are assigned the free energy of pure water, which is obtained from a simulation of pure water.

2.5 Well-Tempered Metadynamics

Metadynamics allows for enhanced sampling compared to standard molecular dynamics (MD) simulations by forcing simulations to sample unexplored regions of free energy landscapes (Laio and Parrinello, 2002). Sampling is directed through the use of collective variables (CVs), whose choice depends on the system studied and the problem under consideration. Each CV is selected to most efficiently sample regions of interest, such as sampling dihedrals on a flexible peptide to find new and metastable conformations. To bias the system to previously unexplored regions, a biasing potential is used to add Gaussian functions to the potential, which depends on what regions of the potential energy landscape have been historically explored. This potential has the form

$$V\left(\overrightarrow{s}, t\right) = \sum_{k\tau<t} W(k\tau) \exp\left(- \sum_{i=1}^{d} \frac{\left(s_i - s_i^{(0)}(k\tau) \right)^2}{2\sigma_i^2} \right) \tag{11}$$

where $V(\overrightarrow{s,t})$ is the history-dependent bias potential summed over all selected instantaneous collective variables $\overrightarrow{s}$ at time t. For each collective variable i, Gaussian functions (kernals) with width σ_i and height $W(k\tau)$ are deposited every interval of t steps. Over a long enough time, the bias potential converges to minus the free energy $G(\overrightarrow{s})$ (plus a constant C) as a function of all the CVs

$$V\left(\vec{s}, t \to \infty\right) = -G\left(\vec{s}\right) + C \tag{12}$$

The CVs chosen are to efficiently fill metastable states and overcome barriers to neighbouring unexplored metastable states. To improve the convergence of metadynamics simulations, the height of the Gaussian is decreased with longer simulation time using well-tempered metadynamics (WT-MTD) (Barducci et al., 2008)

$$W(k\tau) = W_0 \exp\left(-\frac{V\left(s_i^{(0)}(k\tau), k\tau\right)}{k_B \Delta T} \right) \tag{13}$$

where W_0 is the initial Gaussian height and ΔT is a temperature value selected to regulate the extent of exploration of free-energy, selected based on the bias factor γ as follows

$$\gamma = \frac{T + \Delta T}{T} \tag{14}$$

This is the ratio between the temperature of the CVs and the system temperature. WT-MTD helps with free-energy barrier crossing by simulating CVs at higher temperatures and reduces noise by gradually reducing Gaussian heights.

2.5.1 Reweighting Simulations From the Bias Potential

Because the bias potential in WT-MTD simulations is time-dependent, the weighting w for each simulation frame is calculated from the saved history-dependent bias potential applied at a given time frame t (Tiwary and Parrinello, 2015)

$$w(t) = \exp\left(\frac{V(\mathrm{u}) - V(\vec{s}, t)}{k_B T} \right) \tag{15}$$

where $V(\mathrm{u})$ is the unbiased potential, equal to zero. Each statistical contribution to thermodynamics calculated here using MCC theory is weighted according to the above equation at a given simulation frame over all analysed frames using

$$A_w = \frac{A_n w_n + \sum_t^{n-1} A_t w_t}{w_n + \sum_t^{n-1} w_t} \tag{16}$$

Statistical values (A) used to calculate the free energy of excipients and water molecules at frame t are weighted (A_w) using a rolling average, where the weighted value of current frame n is added to the weighted values of all previous frames. The main influence of the biased potential between solutes and solvent is on the topographical entropy of water molecules, which is accounted for here. Protein values are not weighted because protein conformations are not sampled using a bias and the weightings would not greatly affect protein entropy, and in practice because the program used to calculate protein entropy does not account for biases.

2.6 Selection of Collective Variables and Metadynamics Variables

CVs are selected based on how efficiently the interactions between solutes can be sampled without having to sample more expensive molecular conformations. We therefore consider the number of contacts made between molecule types in the system. Three possible contact CVs are sampled: 1) protein-protein contacts of any side-chain oxygen or nitrogen between each protein, 2) polyanion-protein contacts of any oxygen on the polyanion with any oxygen or nitrogen on residue side chains and 3) any contacts between water oxygens and any O or N atom on a protein or polyanion. We therefore efficiently sample protein-protein interactions, protein-polyanion interactions and water-solute interactions. For each CV, the rational switching function, $s(r)$ is used to set boundaries between a contact and no contact such that

$$s\left(r_{ij}\right) = \frac{1 - \left(\frac{r_{ij} - d_0}{r_0}\right)^6}{1 - \left(\frac{r_{ij} - d_0}{r_0}\right)^{12}} \tag{17}$$

where a contact $s(r_{ij})$ between atoms i and j and distance r_{ij} is 1 if less than distance r_0 and zero if beyond this distance. The switching function allows for the transition between 1 to 0 to be a continuous value for CV derivatives. A 10 Å neighbour list cutoff is used and updated every 2000 steps, with a contact being defined as two atoms between 0.5 and 6.5 Å distance. Contacts are described as O or N atoms within 6.5 Å with the switching function starting at 5 Å to gradually set the contact to zero at 6.5 Å. These contact definitions are used during metadynamics simulations. For post-processing of simulations, contacts are defined by coordination shells using the RAD algorithm as mentioned in **Section 2.4**.

The Gaussian width σ for each CV is selected based on the standard deviation observed from unbiased simulations. Contacts between solutes are set with $\sigma = 4$ and for contacts with water molecules $\sigma = 30$. The Gaussian height, $W(k\tau)$, is set to 1.5 kJ mol^{-1} and Guassians are deposited every 500 steps. Simulations are run at $T = 298$ K and the bias factor γ is set to 20. WT-MTD calculations for each system are performed across 25 simulations with differing starting poses of two lysozyme proteins and described in more detail next.

2.7 Multiple Walkers

Multiple walkers are multiple simulations running independently, but sharing information about already visited configurations along several CVs. Because we want to study non-specific interactions, which is characterised by many possibilities of weak or indirect interactions with the protein, there is no precise starting configuration to use. We thus consider multiple starting structures for sampling of interactions between proteins and polyanions. We use multiple walkers of 25 possible starting structures of various orientations between two proteins as shown in **Figure 3**. Starting poses are generated in VMD by sampling 90° orientations about the protein principal axes, giving 21 starting dimers with different protein surfaces facing each other. An additional four poses are included of the same protein surfaces facing each other but rotated about the dimer principal axes. While more starting poses and simulations would be required to achieve full protein-excipient sampling, this number of structures was chosen to resolve the different

FIGURE 3 | The 25 starting poses for two lysozyme proteins for each multiple-walker simulation. First and last protein residues in the sequence are highlighted in orange and cyan, respectively. The *x*, *y* and *z* axes (red, green, blue) lie at the origin of each simulation box.

effects of the two excipients in a computationally efficient manner. A similar method to sample between multiple starting configurations is bias-exchange metadynamics, where many collective variables are sampled and exchanged between multiple simulations (Piana and Laio, 2007). Here we use multiple walkers instead of bias-exchange for a few reasons. Multiple walkers allow for simulations to start and run at different times, unlike with the bias-exchange method, where information is swapped every step between simulations. For multiple walkers, previously visited configurations over all steps and simulations are read in and biased toward unvisited configurations. Given that we use a high performance computing (HPC) cluster, our simulations can run independently without the need to wait for all 25 simulations to be simultanously running, therefore efficiently using the CPU resources. However, a disadvantage of multiple walkers is that only a few collective variables can be sampled. Guided by the previous experimental work (Bye and Curtis, 2019), we assume that sampling contacts between protein and excipients is sufficient to determine possible mechanisms for solute interactions. Lastly, because we use LAMMPS for per-atom energy, we are restricted to using multiple walkers rather than replica exchange, which is not currently supported for use between PLUMED and LAMMPS, as the latter uses temperature swaps, while PLUMED uses coordinate swaps between biased simulations.

2.8 System Setup

Structures of systems containing two lysozyme proteins surrounded by 14,400 water molecules to represent 100 mM protein concentration are created for three different conditions of excipient: tris(hydroxymethyl)aminomethane (TRIS) buffer only, citrate (CIT) with buffer and tripolyphosphate (TPP) with buffer. The two polyanion-containing systems have five polyanions to represent a concentration of approximately 20 mM and three TRIS molecules for approximately 10 mM buffer concentration for all systems. These conditions are comparable to experiment at the concentration where TPP precipitates lysozyme, while CIT does not (Bye and Curtis, 2019). Each system is created with 25 replicates of different starting poses of lysozymes with respect to each other. CIT or TPP are randomly assigned within a 30 Å radius sphere centered within a box of approximately 80 $Å^3$ for each system respectively.

TABLE 1 | Total ΔG, ΔH and $T\Delta S$ and molecular components to form the lysozyme-polyanion systems.

	ΔG / kJ mol^{-1}		ΔH / kJ mol^{-1}		$T\Delta S$ / kJ mol^{-1}	
Species X	**TPP**	**CIT**	**TPP**	**CIT**	**TPP**	**CIT**
Lysozyme	−297	−226	−287	−267	10	−42
Polyanion	131	−76	136	−29	5	47
TRIS	−106	−85	−97	−80	9	6
Na^+	4	−4	4	−4	1	0
Cl^-	−3	−1	−3	−1	0	0
$W_{lysozyme}$	4	−14	28	17	23	32
$W_{polyanion}$	−43	31	−84	30	−41	−1
W_{TRIS}	34	40	38	43	4	3
W Na^+	2	−18	−28	−27	−29	−9
W Cl^-	−2	−1	0	1	2	2
Total	−276	−355	−293	−315	−17	39

Each molecule in the system is neutralised with sodium or chloride ions depending on charge. Protein structure coordinates are taken from the Protein Data Bank with PDB-ID 2VB1 (Berman et al., 2000). The protein is ionised according to pKa of side chains at pH 9 using the PDB2PQR web server (Dolinsky et al., 2004). TPP and TRIS have a population of two charges at pH 9, so two TRIS molecules are set with 1+ charge and the other neutral, while for TPP one molecule is 4− charge and the other four have 5− charge. All five CIT molecules are 3− charge. Small-molecule geometries are built using Avogadro (Hanwell et al., 2012) and initial protein poses are created using VMD (Humphrey et al., 1996). Then all excipient, counterion and solvent molecules are populated around proteins, with at least an 8 Å layer of water molecules to the edge of the box using Packmol (Martínez et al., 2009). Additional systems of each excipient and counterion species surrounded by 900 water molecules are created to represent dilute reference systems with which to calculate changes in energy and entropy. Lysozyme is parametrised using ff14SB (Maier et al., 2015), water with TIP3P (Jorgensen et al., 1983) and and counter-ions with Joung and Cheetham parameters for TIP3P water (Joung and Cheatham, 2008). TRIS and CIT parameters are calculated using the AMBER GAFF (Wang et al., 2004) and TPP parameters with GAFF2 force field, partial charges are parametrised with the AM1-BCC charge method using antechamber (Wang et al., 2006), with additional parameters for the TPP O-P-O-P dihedral and P-O-P angle from Meagher et al. (Meagher et al., 2003).

2.9 Simulation Protocol

Each system is treated as a periodic box and initially minimized for 5,000 steepest descent steps with sander in AMBER18 (Pearlman et al., 1995). Topology and initial coordinates are converted to LAMMPS formatted files with InterMol (Shirts et al., 2017). A short 0.2 ns equilibration is conducted at NVT (number, volume, temperature) conditions to slowly heat the systems to 298 K and then equilibrated for 12 ns at constant NPT (number, pressure, temperature) conditions at 1 atm pressure with 1 fs time-steps in the LAMMPS simulation package (Plimpton, 1995). Restraints on protein Cα atoms are placed with a 500 kJ mol^{-1} spring constant. After standard deviations for each CV are calculated, restraints on backbone atoms are removed and 10 ns of production simulations are performed. In total, 300 ns of equilibration and 250 ns production simulation time are conducted for each of the three systems over all 25 walkers. To maintain constant pressure and temperature, each system is controlled using a Nose-Hoover thermostat and barostat respectively. Temperature is relaxed every 0.2 ps and pressure is relaxed every 0.5 ps with an isotropic stress tensor across x, y, z box dimensions. For non-bonded iteractions, these are truncated to 8 Å, beyond which the particle-particle particle-mesh (PPPM) (Hockney and Eastwood, 1988) is used. All bonds to hydrogen atoms are constrained using the SHAKE algorithm (Ryckaert et al., 1977). Multiple walker well-tempered metadynamics sampling is performed using the PLUMED-2.0 plugin within LAMMPS (Tribello et al., 2014; Bonomi et al., 2019). Per-atom coordinates, forces, potential and kinetic energies are saved every 10 ps and a total of 20,000 frames are analysed for each system. Potential and kinetic energies are outputted per atom using the pe/atom and ke/atom flags respectively within LAMMPS. For potential energy, two and three-body energy terms are divided evenly over involved atoms and the long-range PPPM contributions are calculated using the method from Heyes (Heyes, 1994) within LAMMPS. For water and excipient free energy calculations, output files and topologies are read using the MDAnalysis python library (Michaud-Agrawal et al., 2011) and analysed with an in-house developed python program POSEIDON Beta V2 available at https://github.com/jkalayan/PoseidonBeta. For proteins, LAMMPS output files are first converted to CHARMM. psf and. dcd formats and stripped of solvent and excipients using CPPTRAJ (Roe and Cheatham, 2013) and protein entropy is calculated using code available at https://github.com/arghya90/CodeEntropy.

3 RESULTS

3.1 Free Energy Change for the Formation of Each Lysozyme-Polyanion System

The total change in Gibbs energy, enthalpy and entropy for mixing each lysozyme-polyanion system from the individually separated polyanions and the protein-TRIS systems are shown in **Table 1**.

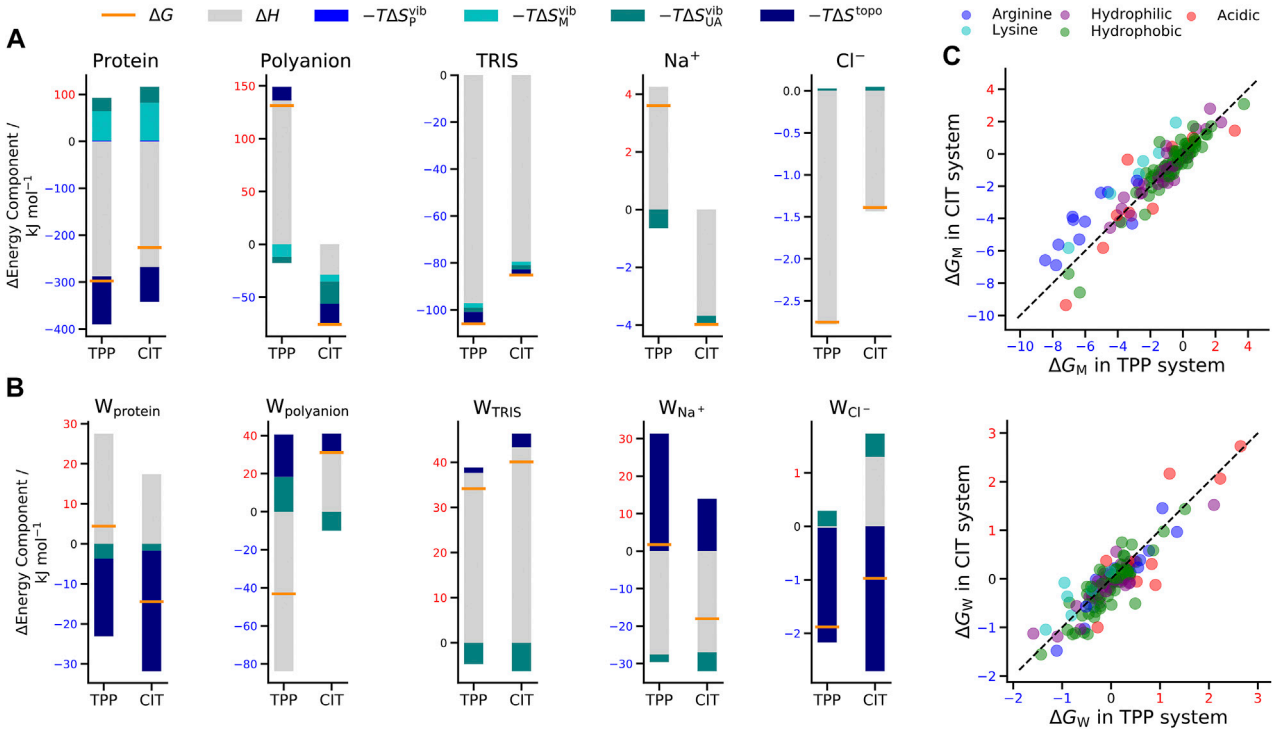


FIGURE 4 | Change in free energy (orange line), enthalpy (grey bar) and entropy (bars in shades of blue) of **(A)** solutes and **(B)** water molecules around solutes in the TPP and CIT systems. **(C)** Change in free energy for CIT vs. TPP of each residue classified by type (top) and of water around each residue (bottom). The dashed line represents $y = x$.

Also included is a decomposition over all five kinds of solute in the system, namely protein lysozyme, polyanion TPP or CIT, buffer TRIS, counterions Na^+ or Cl^- and all the water in the first hydration shell of each of species, denoted by W_X. Estimation of standard errors in each thermodynamic property are extrapolated from a repeated simulation of dilute citrate as detailed and presented in **Supplementary Tables S2 and S3**. The dominant error is that of the protein ΔG at 12.1 kJ mol^{-1} for both excipient systems, with slightly larger errors in ΔH and smaller errors in $T\Delta S$. Standard errors are also estimated based on the number of starting poses, which are analysed in five groups of five poses and are presented in **Supplementary Table S4**. These also show a similar largest contribution from the protein, with values of 15 kJ mol^{-1} for TPP and 21 kJ mol^{-1} for CIT systems. Further entropy decompositions are illustrated in **Figures 4A,B** according to the vibrational entropy at polymer, monomer and united-atom levels and the topographical entropy, which is conformational for proteins and polyanions and orientational for all molecules.

When interpreting the results, we refer to ΔG or ΔH values that are negative as stabilising the system, and conversely positive changes in components as destabilising the system. The opposite holds for $T\Delta S$. The total free energy change for polyanion mixing is large and negative, indicating that the CIT system is stabilised compared to TPP. The main contribution to the greater stability overall is lower CIT free energy compared to the destabilisation of TPP. This is offset by the reverse trend for hydrating water molecules around polyanions, with are stabilised near TPP but destabilised near CIT. Substantial energy-entropy compensation occurs in the stabilising of water molecules hydrating TPP, with water energy decreasing more than entropy. TRIS molecules are consistently stabilised for either polyanion, largely due to energy, while their surrounding waters are both destabilised. Cl^- and Na^+ions are weakly affected upon mixing, although there is a greater change for the water molecules. The water solvating Na^+ is stabilised in the CIT system because of enthalpy but destabilised in TPP due to enthalpy-entropy compensation because of a loss in orientational entropy.

Also of note is the number of contacts between each kind of species. **Table 2** gives the number of contacts for the unbound systems, namely the separated excipient with Na^+ counterions and the two lysozymes buffered with TRIS and Cl^- counterions, while **Table 3** gives the contacts for the mixed polyanion-protein systems. A comparison of the tables shows the most noticeable change is a reduction in the number of contacts between Na^+ and polyanions upon binding for both polyanions, from 51.1 to 30.5 for TPP and from 16.3 to 7.7 for CIT. However, **Table 1** indicates that this release of Na^+ into solution is only destabilising in the TPP system.

For proteins, there is a lower energy for both polyanions due to favorable electrostatic interactions between oppositely charges molecules, in particular, involving charged patches on lysozyme as shown in **Figure 5**. However, protein free energy is seen to be more stable with TPP than CIT. This is due to both energy and entropy (**Table 1**), with contributions from all components of protein entropy (**Figures 4A,B**). Interestingly, both protein

TABLE 2 | Number of contacts between species in the separated excipient and lysozyme systems.

Species X	Lysozyme	TPP	CIT	TRIS	Na^+ (TPP)	Na^+ (CIT)	Cl^-
Lysozyme	3.8	-	-	5.7	-	-	0.9
TPP	-	-	-	-	51.1	-	-
CIT	-	-	-	-	-	16.3	-
TRIS	5.7	-	-	0.0	-	-	0.1
Na^+	-	51.1	16.3	-	0.0	-	-
Cl^-	0.9	-	-	0.1	-	-	-
Water	918.5	76.0	91.3	54.1	79.9	69.8	137.8

TABLE 3 | Number of contacts between species in the polyanion-lysozyme systems.

	TPP system					
	Lysozyme	Polyanion	TRIS	Na^+	Cl^-	N_X
Lysozyme	15.1	17.5	7.3	3.2	0.8	2
Polyanion	17.5	0.1	8.1	30.5	0.0	5
TRIS	7.3	8.1	0.4	0.2	0.0	3
Na^+	3.2	30.5	0.2	-	0.1	24
Cl^-	0.8	0.0	0.0	0.1	0.0	18
Water	866.1	76.8	34.9	100.5	133.2	-
	CIT system					
Lysozyme	16.4	16.7	10.2	2.4	0.8	2
Polyanion	16.7	0.7	9.0	7.7	0.0	5
TRIS	10.2	9.0	0.6	0.1	0.0	3
Na^+	2.4	7.7	0.1	0.0	0.1	15
Cl^-	0.8	0.0	0.0	0.1	-	18
Water	864.1	77.1	31.1	77.3	133.0	-

TABLE 4 | Number of water-molecule environments for each protein-excipient system.

Water contact type	TRIS	TPP	CIT
W_P	906.9	843.1	837.2
W_{PP}	5.2	15.0	15.4
W_{EP}	6.4	7.8	11.0
W_{EPP}	0.0	0.3	0.5
Total	918.5	866.1	864.1

conformational and surrounding water orientational entropy are larger in the presence of polyanions, which suggest weaker interactions of the protein with its environment. Correspondingly, an increase in residue RMSD is observed in the CIT and TPP systems in **Figure 5C**. Water around proteins is destabilised in both energy and entropy in the presence of polyanions, bringing about slightly more stabilisation for CIT than TPP. Further decomposition of the change in free energies for each kind of residue according to charge, hydrophilicity and hydrophobicity and for their surrounding waters are plotted in **Figure 4C**. The free energy of basic residues, particularly Arg, is reduced for both polyanions but more so in the TPP system than CIT system. The free energy of acidic residues is generally more destabilised with TPP than with CIT.

3.2 Interactions Between Solutes

Given the free-energy destabilisation for TPP but stabilisation for CIT, we assess further how interactions with other solute molecules may cause these differences in free energy. In dilute solution TPP has more UA contacts with Na^+, namely 10.2 contacts per TPP with an average TPP charge of −4.8, suggesting strongly bound Na^+ as calculated from the number of contacts between molecules in **Tables 2**, **3**. Dilute CIT has 3.3 contacts with Na^+, matching its −3 cha$^+$rge and suggesting Na^+ is not as tightly bound. In protein solutions, TPP and CIT have approximately 60 and 50% reductions in contacts with Na^+ respectively, but only TPP is destabilised.

From **Figure 5A**, both polyanions have similar numbers of contacts in similar regions of positively charged patches on the protein surface. Patches are generated on the protein using Poisson-Boltzmann electrostatics as described in previous work (Kalayan et al., 2020). CIT has overall weaker interactions over the protein surface, while TPP interacts with fewer regions of the protein but with higher occupancy. From **Figure 5B**, TPP still has a high percentage of Na^+ contacts compared with all other solutes, even though TPP forms more contacts with the protein that CIT. Similarly, Na^+ interacts more with TPP than the protein or CIT. As TPP has more strongly bound Na^+, we assume that the loss of some bound counterions destabilises TPP. Interactions between TPP and basic residues stabilise the protein, but do not appear to stabilise TPP. For CIT, the loss in half of the interacting Na^+ is compensated by the interactions with basic residues which stabilise CIT.

As to why TPP remains bound to the lysozyme surface even though it is destabilised, we observe stabilisation of residues that TPP interacts with most frequently from **Figures 5C**, **4C**, where highest percentage occupied residues are also reduced the most in free energy. Therefore protein stabilisation may prevent TPP from being released into solution. Another possible reason for TPP remaining in the bound state may be attributed to surrounding water molecules as described in more detail next.

3.3 Specific Residue Interactions of Water Molecules

In **Figure 6**, water molecules are analysed based on the solutes in their first coordination shell. Total numbers of each water type are given in **Table 4**. By studying interactions of water with solutes, we can also infer which solutes are in close proximity to one another from the reduction in water coordination.

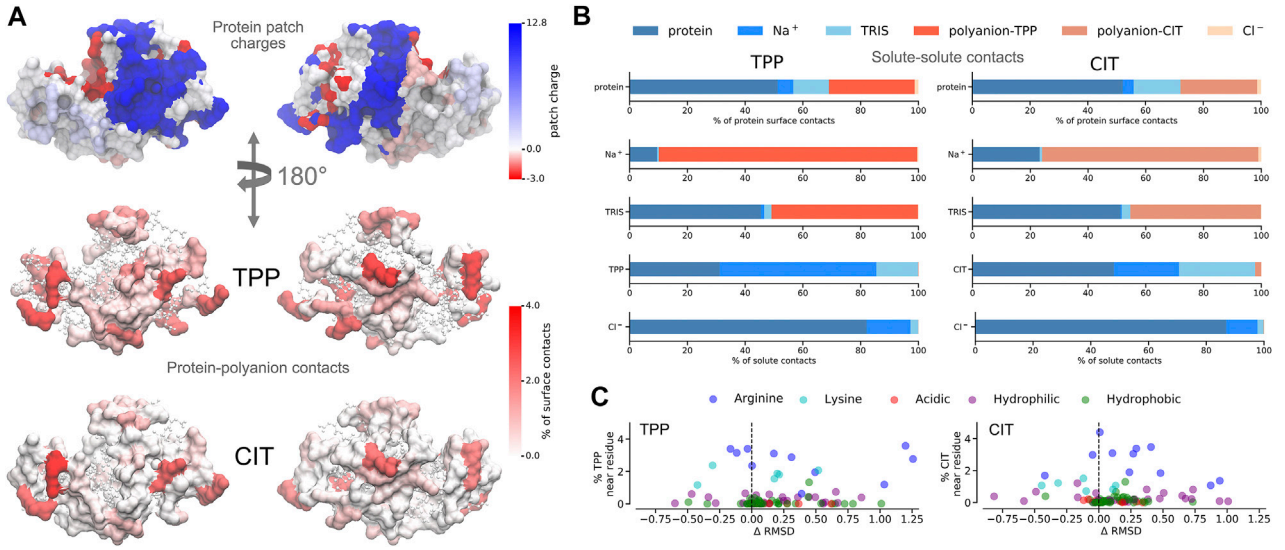


FIGURE 5 | A) Charged patches on lysozyme at pH 9 for two opposing orientations (top). Average percentage of residue contacts (darker regions have more contacts) of polyanions mapped onto the protein surface for the same orientations (middle and bottom). Regions with no contacts are represented as CPK structures. **(B)** Percentages of solute-solute interactions in TPP (left) and CIT (right) systems. **C** Change of per-residue RMSD compared to the protein in buffer only plotted against percentages of polyanion interactions with each residue.

Representation of each water type is shown in part A. In part B, distributions of water energy and entropy for each water type are shown, where water molecules interacting with just protein residues (W_P and W_{PP}) have similar distributions for all three systems. However, when a polyanion is in the system, W_{EP} and W_{EPP} water molecules are most energetically stabilised in the TPP system. Some water molecules are stabilised in the CIT system, but to a greater extent many water molecules remain less stable compared to bulk water.

Part C in **Figure 6** shows the water environments based on their interactions with the nearest two residues, ordered by acidic, basic, uncharged polar and non-polar from left to right. More destabilised W_{EP} and W_{EPP} are present for CIT compared to TPP because excipients are near more hydrophobic residues than in the TPP system. W_P and WW_{PP} water molecules generally are less stable than in bulk. The distribution of neighbouring residues interacting with WW_P water molecules is less specific than residues involved in W_{PP} interactions. This also indicates that PPIs between lysozyme molecules without excipients involved occur mostly between hydrophilic residues. In the presence of polyanions, protein-protein interactions occur mainly between arginine residues. However, in the presence of CIT, PPIs also occur between hydrophobic residues, which explains the distribution of less stable W_{EPP} water molecules in the CIT system in part B.

4 DISCUSSION

Protein precipitation is often characterised by the propensity for ions to "salt-out" a protein. Furthermore, the ion-specificity to cause precipitation can be ranked in the Hofmeister series (Hofmeister, 1888). The mechanism by which protein precipitation occurs is generally considered to be caused by ions forming stronger interactions than proteins with surrounding solvent. It is assumed that stronger ion-solvent interactions leads to a more structured water HB network, explaining the designation of such ions as kosmotropes (Collins, 1997). It is the loss of water at the protein surface that then drives PPIs between hydrophobic regions. Suggested mechanisms for protein precipitation assume that ions do not interact preferentially with the protein surface, and instead remain fully hydrated in solution due to strong interactions with water. In the case of multivalent anions or cations, however, these ions bind onto the protein surface as shown by the change in net protein charge in experimental measurements of their zeta potentials (Matsarskaia et al., 2016; Roosen-Runge et al., 2013; Matsarskaia et al., 2018). At the concentration when the net charge of a globular protein is neutral, the propensity to precipitate is high because repulsive interactions are screened by bound ions. This phenomenon is observed for polyvalent cations (Y^{3+}, La^{3+}, Al^{3+}, Fe^{3+}) when interacting with net-negatively charged globular proteins (Roosen-Runge et al., 2013). Yet similar studies with polyvalent anions shows selectivity between anions to salt out proteins, even though ion binding to the protein and net-neutrality both occur universally (Bye and Curtis, 2019).

To understand why ion specific protein precipitation occurs, the change in both the stability and contacts between solutes molecules is studied for conditions where TPP causes lysozyme to precipitate but CIT does not. Stability is assessed by the change in Gibbs free energy for mixing lysozyme with each excipient and the contributions of all molecules involved. This analysis shows several differences between TPP and CIT-containing systems. First we observe TPP destabilisation upon interaction with

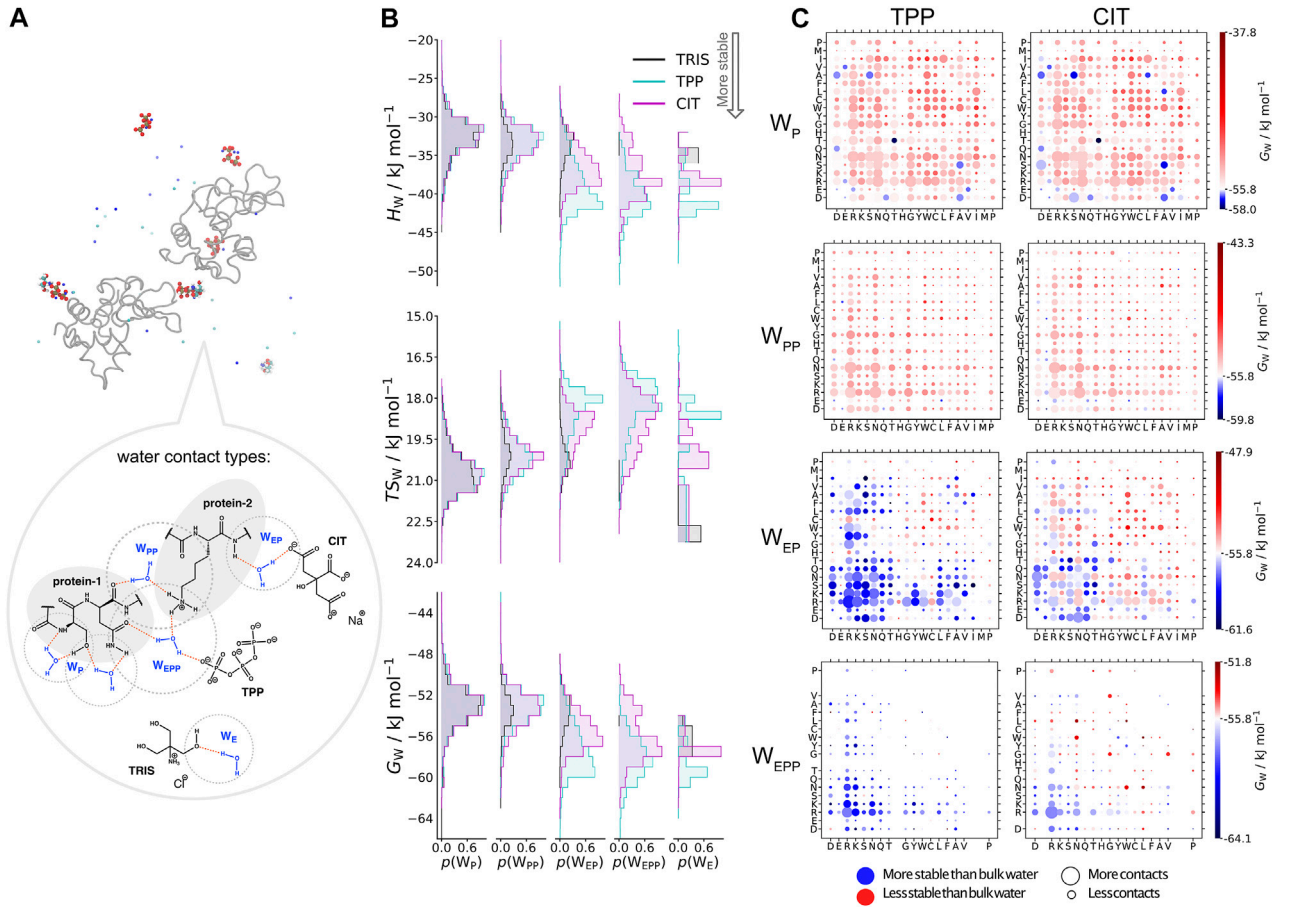


FIGURE 6 | A) Water types (W_P, W_{PP}, W_{EP}, W_{EPP}, W_E) based on what solute types are in their coordination shell (grey dashed circles). **(B)** Distribution of water types energy, entropy and free energy (top to bottom) for each system: TRIS only buffer (black), buffer + TPP (cyan) and buffer + CIT (magenta). **(C)** Water free energy for each water type in TPP and CIT systems is based on which residue pairs are in a water coordination shell.

lysozyme, caused by the release of half of the strongly bound Na^+ ions surrounding TPP. Although unfavourable for TPP, protein binding occurs due to the overall stabilisation of the protein, which is stronger than TPP destabilisation. Conversely, CIT is not destabilised upon protein binding, even with the loss of Na^+ ions, which are not strongly bound to CIT. However, cross-linking with another protein does not occur with CIT because this is not favourable for the surrounding water molecules.

The stability and contacts of water surrounding each solute reveals the stabilisation of water around TPP, but overall destabilisation around CIT. Upon binding to the protein, TPP does not lose any water-molecule contacts. Instead, water-molecule energy is stabilised but entropy is lost, therefore forming a more structured HB network around TPP and agreeing with the description of TPP as a kosmotropic ion. Thus instead of water being stabilised upon release into solution as is usually the case for PPIs, water is stabilised by forming strong interactions with protein-bound TPP when replacing released Na^+ counterions. Water entropy around CIT does not change greatly ($T\Delta S = -1$ kJ mol^{-1}) when CIT interacts with the protein but the contacts of CIT with water moleculesreduce by ~ 15%. The lack of strongly bound, ordered water molecules suggests that CIT behaves as a chaotropic salt, thereby not salting out the protein. Overall, the stabilisation of water around lysozyme upon CIT binding maintains lysozyme solubility.

Although water is slightly destabilised around lysozyme in the presence of TPP, it does not seem to be destabilised enough to cause precipitation via hydrophobic interactions. Instead the combination of overall solvent and protein stabilisation upon TPP binding is expected to drive binding via cross-linking with other proteins at moderate TPP concentration. In the conditions simulated here, five polyanion molecules per lysozyme pair closely represents a 20 mM concentration. There are a total of 34 basic residues on the lysozyme pair at pH 9 with a net charge of 16+, therefore giving many vacancies for TPP cross-linking. At 100 mM TPP concentration, experiment shows that lysozyme is fully resolublised into solution (Bye and Curtis, 2019) and there are approximately 29 TPP molecules per lysozyme pair. Each basic residue can be bound individually at 100 mM TPP, and therefore no cross-linking between proteins would be required. To more conclusively explain the experimental data, much longer

simulations with more starting poses would be required that can explicitly account for direct lysozyme-lysozyme binding and cross-linking of lysozyme molecules by excipients at varying concentrations of each excipient. Adequately addressing such questions may require coarse-grain simulations to achieve the necessary sampling. At the same time, greater model accuracy may also be required by accounting for the possibility of variable protonation states of amino acids and excipients arising from their changing environments using constant pH simulations. This makes clear that much work remains in order to fully understand the effects of excipients on protein behaviour.

5 CONCLUSION

Lysozyme-excipient systems have been studied to understand how excipients differentially interact with lysozyme using a statistical-mechanics based free-energy method called Energy-Entropy Multiscale Cell Correlation that calculates the energy and entropy of each solute and hydration shell water molecules. Simulations are conducted using multiple walker metadynamics simulations followed by reweighting to give a Boltzmann distribution. We observe different contacts and thermodynamic properties when the excipient TPP interacts with lysozyme compared to the excipient CIT. A possible mechanism by which TPP precipitates lysozyme is suggested to be due to the stabilisation in free energy of both lysozyme and solvent surrounding TPP upon TPP interacting with basic residues and the release of bound Na^+ ions.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

JK carried out the research; all authors designed the research; JK and RH devised the theory; JK and RH wrote the manuscript with support from RC and JW; RH supervised the project.

FUNDING

JK was supported by Centre of Doctoral Training in Emergent Macromolecular Therapy funded by the EPSRC under grant codes EP/L015218/1 and EP/N025105/1.

ACKNOWLEDGMENTS

The authors thank IT Services at the University of Manchester for access to the Computational Shared Facility.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.689400/full#supplementary-material

REFERENCES

Ali, H. S., Higham, J., de Visser, S. P., and Henchman, R. H. (2020). Comparison of Free-Energy Methods to Calculate the Barriers for the Nucleophilic Substitution of Alkyl Halides by Hydroxide. *J. Phys. Chem. B.* 124(31): 6835–6842. doi:10.1021/acs.jpcb.0c02264

Ali, H. S., Higham, J., and Henchman, R. H. (2019). Entropy of Simulated Liquids Using Multiscale Cell Correlation. *Entropy* 21, 750. doi:10.3390/e21080750

Banerjee, P., Mondal, S., and Bagchi, B. (2019). Effect of Ethanol on Insulin Dimer Dissociation. *J. Chem. Phys.* 150, 084902. doi:10.1063/1.5079501

Barata, T., Zhang, C., Dalby, P., Brocchini, S., and Zloh, M. (2016). Identification of Protein-Excipient Interaction Hotspots Using Computational Approaches. *Ijms* 17, 853. doi:10.3390/ijms17060853

Barducci, A., Bonomi, M., Prakash, M. K., and Parrinello, M. (2013). Free-energy Landscape of Protein Oligomerization from Atomistic Simulations. *Proc. Natl. Acad. Sci.* 110, E4708–E4713. doi:10.1073/pnas.1320077110

Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* 100, 020603. doi:10.1103/PhysRevLett.100.020603

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

Bonomi, M., Bussi, G., Camilloni, C., Tribello, G. A., Banáš, P., Barducci, A., et al. (2019). Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* 16, 670–673. doi:10.1038/s41592-019-0506-8

Bye, J. W., and Curtis, R. A. (2019). Controlling Phase Separation of Lysozyme with Polyvalent Anions. *J. Phys. Chem. B.* 123, 593–605. doi:10.1021/acs.jpcb.8b10868

Chakravorty, A., Higham, J., and Henchman, R. H. (2020). Entropy of Proteins Using Multiscale Cell Correlation. *J. Chem. Inf. Model.* 60, 5540–5551. doi:10.1021/acs.jcim.0c00611

Chong, S.-H., and Ham, S. (2017). Dynamics of Hydration Water Plays a Key Role in Determining the Binding Thermodynamics of Protein Complexes. *Sci. Rep.* 7, 8744. doi:10.1038/s41598-017-09466-w

Collins, K. D. (1997). Charge Density-dependent Strength of Hydration and Biological Structure. *Biophysical J.* 72, 65–76. doi:10.1016/S0006-3495(97)78647-8

Dimitrov, D. S. (2012). "Therapeutic Proteins," in *Methods in Molecular Biology*. Editors V. Voynov and J. A. Caravella (Humana Press), Vol. 899, 1–26. doi:10.1007/978-1-61779-921-1_1

Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., and Baker, N. A. (2004). PDB2PQR: an Automated Pipeline for the Setup of Poisson-Boltzmann Electrostatics Calculations. *Nucleic Acids Res.* 32, W665–W667. doi:10.1093/nar/gkh381

Du, H., Liu, Z., Jennings, R., and Qian, X. (2017). The Effects of Salt Ions on the Dynamics and Thermodynamics of Lysozyme Unfolding. *Separation Sci. Technol.* 52, 320–331. doi:10.1080/01496395.2016.1229336

Faber, M. S., and Whitehead, T. A. (2019). Data-driven Engineering of Protein Therapeutics. *Curr. Opin. Biotechnol.* 60, 104–110. doi:10.1016/j.copbio.2019.01.015

Falconer, R. J. (2019). Advances in Liquid Formulations of Parenteral Therapeutic Proteins. *Biotechnol. Adv.* 37, 107412. doi:10.1016/j.biotechadv.2019.06.011

Gerogiokas, G., Southey, M. W. Y., Mazanetz, M. P., Heifetz, A., Bodkin, M., Law, R. J., et al. (2016). Assessment of Hydration Thermodynamics at Protein Interfaces with Grid Cell Theory. *J. Phys. Chem. B.* 120, 10442–10452. doi:10.1021/acs.jpcb.6b07993

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

Henchman, R. H., and Irudayam, S. J. (2010). Topological Hydrogen-Bond Definition to Characterize the Structure and Dynamics of Liquid Water. *J. Phys. Chem. B.* 114, 16792–16810. doi:10.1021/jp105381s

Heyes, D. M. (1994). Pressure Tensor of Partial-Charge and point-dipole Lattices with Bulk and Surface Geometries. *Phys. Rev. B.* 49, 755–764. doi:10.1103/PhysRevB.49.755

Higham, J., Chou, S.-Y., Gräter, F., and Henchman, R. H. (2018). Entropy of Flexible Liquids from Hierarchical Force-Torque Covariance and Coordination. *Mol. Phys.* 116, 1965–1976. doi:10.1080/00268976.2018.1459002

Higham, J., and Henchman, R. H. (2016). Locally Adaptive Method to Define Coordination Shell. *J. Chem. Phys.* 145, 084108. doi:10.1063/1.4961439

Hockney, R. W., and Eastwood, J. W. (1988). *Computer Simulation Using Particles*. New York, NY: Taylor and Francis Group.

Hofmeister, F. (1888). Zur Lehre von der Wirkung der Salze. *Archiv F. Experiment. Pathol. U. Pharmakol* 24, 247–260. doi:10.1007/BF01918191

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual Molecular Dynamics. *J. Mol. Graphics* 14 (33–38), 27–28. doi:10.1016/0263-7855(96)00018-5

Irudayam, S. J., and Henchman, R. H. (2011). Prediction and Interpretation of the Hydration Entropies of Monovalent Cations and Anions. *Mol. Phys.* 109, 37–48. doi:10.1080/00268976.2010.532162

Irudayam, S. J., Plumb, R. D., and Henchman, R. H. (2010). Entropic Trends in Aqueous Solutions of the Common Functional Groups. *Faraday Discuss.* 145, 467–485. doi:10.1039/b907383c

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

Joung, I. S., and Cheatham, T. E. (2008). Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B.* 112, 9020–9041. doi:10.1021/jp8001614

Kalayan, J., Henchman, R. H., and Warwicker, J. (2020). Model for Counterion Binding and Charge Reversal on Protein Surfaces. *Mol. Pharmaceutics* 17, 595–603. doi:10.1021/acs.molpharmaceut.9b01047

Laio, A., and Parrinello, M. (2002). Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci.* 99, 12562–12566. doi:10.1073/pnas.202427399

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theor. Comput.* 11, 3696–3713. doi:10.1021/acs.jctc.5b00255

Martínez, L., Andrade, R., Birgin, E. G., and Martínez, J. M. (2009). PACKMOL: A Package for Building Initial Configurations for Molecular Dynamics Simulations. *J. Comput. Chem.* 30, 2157–2164. doi:10.1002/jcc.21224

Matsarskaia, O., Braun, M. K., Roosen-Runge, F., Wolf, M., Zhang, F., Roth, R., et al. (2016). Cation-Induced Hydration Effects Cause Lower Critical Solution Temperature Behavior in Protein Solutions. *J. Phys. Chem. B.* 120, 7731–7736. doi:10.1021/acs.jpcb.6b04506

Matsarskaia, O., Roosen-Runge, F., Lotze, G., Möller, J., Mariani, A., Zhang, F., et al. (2018). Tuning Phase Transitions of Aqueous Protein Solutions by Multivalent Cations. *Phys. Chem. Chem. Phys.* 20, 27214–27225. doi:10.1039/C8CP05884A

Meagher, K. L., Redman, L. T., and Carlson, H. A. (2003). Development of Polyphosphate Parameters for Use with the AMBER Force Field. *J. Comput. Chem.* 24, 1016–1025. doi:10.1002/jcc.10262

Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., and Beckstein, O. (2011). MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* 32, 2319–2327. doi:10.1002/jcc.21787

Moussa, E. M., Panchal, J. P., Moorthy, B. S., Blum, J. S., Joubert, M. K., Narhi, L. O., et al. (2016). Immunogenicity of Therapeutic Protein Aggregates. *J. Pharm. Sci.* 105, 417–430. doi:10.1016/j.xphs.2015.11.002

Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., DeBolt, S., et al. (1995). AMBER, a Package of Computer Programs for Applying Molecular Mechanics, normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Computer Phys. Commun.* 91, 1–41. doi:10.1016/0010-4655(95)00041-D

Piana, S., and Laio, A. (2007). A Bias-Exchange Approach to Protein Folding. *J. Phys. Chem. B.* 111, 4553–4559. doi:10.1021/jp067873l

Plattner, N., Doerr, S., De Fabritiis, G., and Noé, F. (2017). Complete Protein-Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling. *Nat. Chem.* 9, 1005–1011. doi:10.1038/nchem.2785

Plimpton, S. (1995). Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* 117, 1–19. doi:10.1006/jcph.1995.1039

Roe, D. R., and Cheatham, T. E. (2013). PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theor. Comput.* 9, 3084–3095. doi:10.1021/ct400341p

Roosen-Runge, F., Heck, B. S., Zhang, F., Kohlbacher, O., and Schreiber, F. (2013). Interplay of pH and Binding of Multivalent Metal Ions: Charge Inversion and Reentrant Condensation in Protein Solutions. *J. Phys. Chem. B.* 117, 5777–5787. doi:10.1021/jp401874t

Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977). Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* 23, 327–341. doi:10.1016/0021-9991(77)90098-5

Schames, J. R., Henchman, R. H., Siegel, J. S., Sotriffer, C. A., Ni, H., and McCammon, J. A. (2004). Discovery of a Novel Binding Trench in HIV Integrase. *J. Med. Chem.* 47, 1879–1881. doi:10.1021/jm0341913

Shirts, M. R., Klein, C., Swails, J. M., Yin, J., Gilson, M. K., Mobley, D. L., et al. (2017). Lessons Learned from Comparing Molecular Dynamics Engines on the SAMPL5 Dataset. *J. Comput. Aided Mol. Des.* 31, 147–161. doi:10.1007/s10822-016-9977-1

Singh, R., Bansal, R., Rathore, A. S., and Goel, G. (2017). Equilibrium Ensembles for Insulin Folding from Bias-Exchange Metadynamics. *Biophysical J.* 112, 1571–1585. doi:10.1016/j.bpj.2017.03.015

Spyrakis, F., Benedetti, P., Decherchi, S., Rocchia, W., Cavalli, A., Alcaro, S., et al. (2015). A Pipeline to Enhance Ligand Virtual Screening: Integrating Molecular Dynamics and Fingerprints for Ligand and Proteins. *J. Chem. Inf. Model.* 55, 2256–2274. doi:10.1021/acs.jcim.5b00169

Tarek, M., and Tobias, D. J. (2000). The Dynamics of Protein Hydration Water: A Quantitative Comparison of Molecular Dynamics Simulations and Neutron-Scattering Experiments. *Biophysical J.* 79, 3244–3257. doi:10.1016/S0006-3495(00)76557-X

Tiwary, P., and Parrinello, M. (2015). A Time-independent Free Energy Estimator for Metadynamics. *J. Phys. Chem. B* 119, 736–742. doi:10.1021/jp504920s

Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C., and Bussi, G. (2014). PLUMED 2: New Feathers for an Old Bird. *Computer Phys. Commun.* 185, 604–613. doi:10.1016/j.cpc.2013.09.018

Troussicot, L., Guillière, F., Limongelli, V., Walker, O., and Lancelin, J.-M. (2015). Funnel-Metadynamics and Solution NMR to Estimate Protein-Ligand Affinities. *J. Am. Chem. Soc.* 137, 1273–1281. doi:10.1021/ja511336z

Verona, M. D., Verdolino, V., Palazzesi, F., and Corradini, R. (2017). Focus on PNA Flexibility and RNA Binding Using Molecular Dynamics and Metadynamics. *Sci. Rep.* 7, 42799. doi:10.1038/srep42799

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. Graphics 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

Wang, W. (2015). Advanced Protein Formulations. *Protein Sci.* 24, 1031–1039. doi:10.1002/pro.2684

Yu, J., Ciancetta, A., Dudas, S., Duca, S., Lottermoser, J., and Jacobson, K. A. (2018). Structure-Guided Modification of Heterocyclic Antagonists of the P2Y14 Receptor. *J. Med. Chem.* 61, 4860–4882. doi:10.1021/acs.jmedchem.8b00168

Zalar, M., Svilenov, H. L., and Golovanov, A. P. (2020). Binding of Excipients Is a Poor Predictor for Aggregation Kinetics of Biopharmaceutical Proteins. *Eur. J. Pharmaceutics Biopharmaceutics* 151, 127–136. doi:10.1016/j.ejpb.2020.04.002

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.

Copyright © 2021 Kalayan, Curtis, Warwicker and Henchman. 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 Molecular Biosciences

REVIEW
published: 17 June 2021
doi: 10.3389/fmolb.2021.693710

A Primer on the oxDNA Model of DNA: When to Use it, How to Simulate it and How to Interpret the Results

A. Sengar[1], T. E. Ouldridge[1], O. Henrich[2], L. Rovigatti[3,4] and P. Šulc[5,6]*

[1]Centre for Synthetic Biology, Department of Bioengineering, Imperial College London, London, United Kingdom, [2]Department of Physics, SUPA, University of Strathclyde, Glasgow, United Kingdom, [3]Department of Physics, Sapienza University of Rome, Rome, Italy, [4]CNR Institute of Complex Systems, Sapienza University of Rome, Rome, Italy, [5]Center for Molecular Design and Biomimetics, The Biodesign Institute, Arizona State University, Tempe, AZ, United States, [6]School of Molecular Sciences, Arizona State University, Tempe, AZ, United States

The oxDNA model of Deoxyribonucleic acid has been applied widely to systems in biology, biophysics and nanotechnology. It is currently available via two independent open source packages. Here we present a set of clearly documented exemplar simulations that simultaneously provide both an introduction to simulating the model, and a review of the model's fundamental properties. We outline how simulation results can be interpreted in terms of—and feed into our understanding of—less detailed models that operate at larger length scales, and provide guidance on whether simulating a system with oxDNA is worthwhile.

Keywords: coarse-grained modelling, deoxyribonucleic acid, nanotechnology, biophysics, molecular simulation (molecular modeling)

OPEN ACCESS

Edited by:
Fabio Trovato,
Freie Universität Berlin, Germany

Reviewed by:
Naoto Hori,
University of Nottingham, United Kingdom
Monte Pettitt,
University of Texas Medical Branch at Galveston, United States

****Correspondence:***
T. E. Ouldridge
t.ouldridge@imperial.ac.uk

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *11 April 2021*
Accepted: *27 May 2021*
Published: *17 June 2021*

Citation:
Sengar A, Ouldridge TE, Henrich O, Rovigatti L and Šulc P (2021) A Primer on the oxDNA Model of DNA: When to Use it, How to Simulate it and How to Interpret the Results. Front. Mol. Biosci. 8:693710. doi: 10.3389/fmolb.2021.693710

1 INTRODUCTION

Deoxyribonucleic acid (DNA) is a macromolecule that acts as a storage medium for genetic information for all living organisms (Alberts et al., 2002). In nature, the molecule is most often found as a double helix of two strands. The structure of each strand comprises of a backbone of covalently linked sugar and phosphate groups. Each sugar is further attached to a base moiety: adenine (A), guanine (G), cytosine (C) or thymine (T). Certain intra- and intermolecular interactions between these bases drive the formation of the aforementioned double helical structure.

Crucially, the base pairing that holds these duplexes together is highly specific; to a first approximation, A will only bind to T and C will only bind to G, and vice versa. Matching—or complementary—sequences therefore bind to each other much more strongly than to non-complementary sequences. The different base identities, along with the rules of complementarity, allow information to be encoded into the single strands and copied from generation to generation (Watson and Crick, 1953).

The DNA double helix has a diameter of about 2 nm, and a helical pitch of about 3.4–3.6 nm. Double strands are relatively stiff, with large bending disfavoured on lengthscales below around 40–50 nm (Seeman, 2003). By contrast, single strands are very flexible (Murphy et al., 2004; Chen et al., 2012) forming loops and kinks with only a handful of bases or fewer.

These thermodynamic, mechanical and structural properties influence DNA's biological role, but also make it an ideal material for nanoscale engineering. The simplicity of interactions between strands, and the predictability of the structural and mechanical properties of the product, have enabled the rational design of a host of synthetic structures (Fu and Seeman, 1993; Goodman et al., 2005; Rothemund, 2006; Douglas et al., 2009; Ke et al., 2012; Zhang et al., 2015; Tikhomirov et al., 2017; Wagenbauer et al., 2017), computing architectures (Adleman, 1994; Rothemund et al., 2004;

Qian et al., 2011; Cherry and Qian, 2018; Woods et al., 2019) and dynamic systems (Yurke et al., 2000; Shin and Pierce, 2004; Muscat et al., 2011; Zhang and Seelig, 2011; Wickham et al., 2012; Srinivas et al., 2017; Tomov et al., 2017).

DNA's importance to biology, nanotechnology and simply as a canonical model biopolymer for biophysicists means that modelling its behaviour is a key challenge. Unsurprisingly, therefore, models spanning an enormous range of complexity have been proposed to analyse and rationalize the behaviour of DNA. In this pedagogical review, we will first discuss this range of models and their interplay, before focusing on a particular coarse-grained model, oxDNA.

The oxDNA model, first published in 2010 (Ouldridge et al., 2010a) (and with a slightly updated potential in 2011 (Ouldridge et al., 2011)), has now been extensively applied to problems in nanotechnology (Ouldridge et al., 2013a; Doye et al., 2013; Srinivas et al., 2013; Machinek et al., 2014; Snodin et al., 2016, 2019; Henning-Knechtel et al., 2017; Hong et al., 2018), soft matter (De Michele et al., 2012; Rovigatti et al., 2014; Procyk et al., 2020; Stoev et al., 2020), biophysics (Matek et al., 2012; Romano et al., 2013; Matek et al., 2015; Mosayebi et al., 2015; Harrison et al., 2019; Nomidis et al., 2019) and biology (Lee et al., 2015; Wang et al., 2015; Craggs et al., 2019). Numerous tools exist to generate and visualize systems with oxDNA (Henrich et al., 2018; Suma et al., 2019), alongside two independent, publicly-available code bases for actually running simulations with at least three qualitatively distinct algorithms for simulating the model (Ouldridge et al., 2011; Snodin et al., 2015). One of these code bases has recently been incorporated into a webserver (Poppleton et al., 2020).

Despite this uptake, however, there is insufficient clarity on how the basic properties of the oxDNA model make it well- or poorly-suited to studying certain systems. Moreover, many interesting phenomena require non-trivial simulation techniques if they are to be probed with oxDNA. Although those techniques have been widely applied, and software implementing them with oxDNA is available, documentation supporting their use is limited. Equally, there is very little help with the intuition required to use these techniques successfully. Finally, a major aspect to interpreting the results from oxDNA is rationalizing its predictions in terms of less detailed models. Unfortunately, however, there are many subtleties in doing so.

In this pedagogical review we implement a series of exemplar simulations that allow us to address these shortcomings. These simulations will establish a well-documented set of examples for a series of approaches that can be adapted by users, and this review will provide some of the intuition for how to use these approaches successfully. Simultaneously, we will use these examples to illustrate key aspects of the oxDNA model that determine its usefulness, and will explore how to interpret the results in terms of DNA models at different scales.

2 DNA MODELS ACROSS LENGTH SCALES

At the smallest and most fundamental scale, quantum chemistry calculations can be used to estimate the nucleotide properties from first principles (Hobza and Šponer, 1999; Pérez et al., 2004; Šponer et al., 2004; Šponer et al., 2008). However, these calculations are computationally extremely expensive and are unable to capture the collective behaviour of whole strands in solution. Nonetheless, insight from this field has been incorporated into classical atomistic force fields AMBER (Cornell et al., 1996) and CHARMM (Brooks et al., 1983) that use empirical force fields to model interactions between atoms. These force fields are iteratively parameterised using both comparison to experimental data and information from lower-level quantum mechanical descriptions. In recent years, advances in computational resources have allowed these models to simulate large systems—such as DNA origami—for long enough timescales to analyse their equilibrium properties. Given long simulations, these atomistic models are able to sample the conformation of large structures (Nguyen et al., 2014; Rocklin et al., 2017) and the breaking and formation of base pairs (Brown et al., 2015). However, at the time of writing, a systematic study of DNA duplex formation thermodynamics, as represented by atomistic models, has not been performed. As such, it is unknown how well these atomistic models represent DNA thermodynamics—historically, the force fields have required adjustment as new systems and longer time scales are studied (Pérez et al., 2007; Yoo and Aksimentiev, 2012). This fact, alongside the heavy computational load in simulating large systems or significant structural changes, mean that atomistic approaches are currently limited to a fraction of the systems of interest in DNA-based biophysics, biology, soft matter and nanotechnology.

In an effort to access longer timescales, a number of "coarse-grained" or "mesoscale" models have been introduced (Savelyev and Papoian, 2009; Ouldridge et al., 2011; Hinckley et al., 2013; Korolev et al., 2014; Maciejczyk et al., 2014; Maffeo et al., 2014; Machado and Pantano, 2015; Uusitalo et al., 2015; Dans et al., 2016; Ivani et al., 2016; Chakraborty et al., 2018; Maffeo and Aksimentiev, 2020). These models represent DNA with a much-reduced set of degrees of freedom relative to atomistic approaches. In particular, solvent (and solvated ions) are usually treated implicitly, and groups of atoms in the DNA are replaced by a single site with effective interactions. As a result, these models can access longer length and time scales than atomistic descriptions.

The procedure for coarse-graining ranges from "bottom-up" approaches that seek to formally map the statistical behaviour of a more detailed model into a coarse-grained description (Savelyev and Papoian, 2009; Maciejczyk et al., 2014; Maffeo et al., 2014), to "top-down" approaches such as oxDNA that are more *ad hoc*, instead seeking to reproduce as many experimentally relevant properties as possible (Ouldridge et al., 2009; Hinckley et al., 2013; Machado and Pantano, 2015; Uusitalo et al., 2015; Chakraborty et al., 2018; Maffeo and Aksimentiev, 2020). Bottom-up approaches have been most successfully used to study fluctuations within the duplex state, where the atomistic models on which they are built are best parameterised. Top-down approaches, by contrast, have found their application in the analysis of processes that involve DNA outside of its canonical B-form, including duplex hybridization (Ouldridge et al., 2013b), strand displacement (Srinivas et al., 2013; Irmisch et al., 2020),

stress-induced structural transitions (Romano et al., 2013; Wang and Pettitt, 2014; Sutthibutpong et al., 2016) and the properties of nanostructures with branched helices and single-stranded sections (Rovigatti et al., 2014; Engel et al., 2020).

Although highly-simplified, all of the coarse-grained models cited above attempt to represent the discrete, three-dimensional structure of DNA explicitly. An important role in our understanding of DNA is played by even simpler models. In thermodynamic terms, two classes of model have received particular attention. Firstly, the Peyrard-Bishop-Dauxois model and its variants have been used to probe the statistical properties of the duplex denaturation transition in the thermodynamic limit (Dauxois et al., 1993; Cocco and Monasson, 1999; Nisoli and Bishop, 2011). These models represent DNA through two or three continuous degrees of freedom per base pair.

A second approach dispenses with continuous degrees of freedom altogether, taking an Ising-like approach in which base pairs are either present or absent. Originally introduced by Poland and Scheraga to probe the duplex denaturation phase transition (Poland and Scheraga, 1966), the approach was adapted and carefully parameterized (SantaLucia, 1998; SantaLucia and Hicks, 2004; Huguet et al., 2010; Bae et al., 2020) to describe binding equilibria for strands of moderate length (oligonucleotides). It is difficult to overstate just how influential the nearest neighbour model has been, particularly in the development of nucleic acid nanotechnology, as it allows rational design of an ensemble of strands to produce the desired thermodynamics. The NUPACK software suite automates this process of system analysis and thermodynamics-based design by implementing the nearest-neighbour model (SantaLucia and Hicks, 2004). A number of attempts have been made to augment this thermodynamic model with realistic kinetics (Flamm et al., 2000; Xayaphoummine et al., 2005; Srinivas et al., 2013; Schaeffer et al., 2015).

At its simplest, the nearest-neighbour model allows a two-state approximation to the binding of A and B, in which the strands are either fully bound or fully dissociated. In this limit, the concentration of the product $[AB]$ can be estimated using the equation

$$\frac{[AB]}{[A][B]} = \exp\left(-\Delta H_{AB} - T\Delta S_{AB}/kT\right). \quad (1)$$

Here ΔH_{AB} and ΔS_{AB} are computed by summing contributions from each nearest-neighbor set of two base pairs, together with terms for helix initiation and various structural features, all of which are assumed to be temperature independent.

Another class of models ignores thermodynamics entirely, instead providing a continuum-level description of DNA mechanics. Most notably, DNA is frequently modelled as a semi-flexible polymer (or worm-like chain, WLC) characterised by a bending modulus (Kratky and Porod, 1949). This model can be augmented with an extensional modulus (Odijk, 1995) and a representation of twist with associated twist modulus (Yamakawa, 1977). It is also possible to consider coupling between the modes of deformation (Gore et al., 2006; Nomidis et al., 2019). As with the nearest-neighbour model of DNA, the influence of these approaches is enormous, particularly within the biophysics community. The elastic rod is the starting point for understanding the geometry of DNA, and the null model against which results are compared and interpreted.

There is actually quite a large gap in complexity between mesoscopic models such as oxDNA and the continuum WLC models or the nearest neighbour model of thermodynamics. The time required to analyse the same system with these methods differs by many orders of magnitude. It is intriguing that, to our knowledge, there are few approaches that come close to bridging this gap. Fundamentally, it is not easy to combine the mechanics of semiflexible DNA as captured by the WLC, the geometry and topology of DNA structures, and the thermodynamics of DNA duplex formation as described in the nearest-neighbour model, in a representation that is simultaneously quantitatively useful and substantially simpler than the existing mesoscale models. Approaches such as Benham's description of melting in circularly negatively-supercoiled DNA (Fye and Benham, 1999) achieve this marriage in specific contexts. The variety of possible behaviour, however, and the sensitive interplay of topology, structure, mechanics and thermodynamics in many systems of interest, make the development of such models extremely hard and currently necessitate the application of coarse-grained models such as oxDNA.

In the rest of this pedagogical review, we first provide a high-level description of the basics of the oxDNA model and simulation techniques. We then present prototypical simulations to demonstrate key properties of oxDNA, and discuss how results from these simulations can be interpreted in terms of simpler DNA models at different length scales. While doing so, we discuss specific challenges in obtaining meaningful data from oxDNA simulations, and discuss where oxDNA provides added value. Initialisation files, processing scripts and supporting instructions are provided for all simulations presented here at (Sengar, 2021). This review should then serve as an introductory tutorial to applying oxDNA.

One drawback of this format is that the examples are presented as a fait accompli; just re-running the code will provide a limited experience of the real process of simulating oxDNA. We strongly encourage readers using this document as a tutorial to attempt to construct as much as possible of the simulations for themselves, and then to compare to the results obtained here. Alternatively, users may try to construct variants to simulate similar systems. Additional guidance on the nuts and bolts of running simulations can be found at Ref (oxDNA wiki, 2015; LAMMPS Documentation, 2021), where instructions on visualizing the output can also be found. In general, we have found that checking one or two snapshots of a simulation can avoid many wasted hours simulating and studying faulty systems.

3 THE OXDNA MODEL

The oxDNA model was originally developed to study the self-assembly, structure and mechanical properties of DNA

nanostructures, and the action of DNA nanodevices—although it has since been applied more broadly. To describe such systems, a model needs to capture the structural, mechanical and thermodynamic properties of single-stranded DNA, double-stranded DNA, and the transition between the two states. It must also be feasible to simulate large enough systems for long enough to sample the key phenomena. As discussed in **Section 2**, mesoscopic models in which multiple atoms are represented by a single interaction site are the appropriate resolution for these goals.

We will now outline the key features of the oxDNA model, the specific mesoscale model that is the focus of this review. While doing so, we note that there are effectively three versions of the oxDNA potential that are publicly available. The original model, oxDNA1.0 (Ouldridge et al., 2011), lacks sequence-specific interaction strengths, electrostatic effects and major/minor grooving. oxDNA1.5 adds sequence-dependent interaction strengths to oxDNA1.0 (Šulc et al., 2012), and oxDNA2.0 (Snodin et al., 2015) also includes a more accurate structural model, alongside an explicit term in the potential for screened electrostatic interactions between negatively charged sites on the nucleic acid backbone. In addition to these three versions of the DNA model, an RNA parameterisation "oxRNA" has also been introduced (Šulc et al., 2014).

It is worth noting that the three versions of oxDNA are very similar; most of the changes involve small adjustments of the geometry and strength of interactions. Structurally, the most significant change is the addition of a screened electrostatic interaction in oxDNA 2.0, which is typically small unless low salt concentrations are used. Moreover, subsequent versions of the model have been explicitly designed to preserve aspects of earlier versions that performed well. So oxDNA 1.5 and oxDNA 1.0 are very similar, except that oxDNA predicts sequence-dependent thermodynamic effects that are absent in oxDNA 1.0. oxDNA 2.0 is designed to preserve the thermodynamic and mechanical properties of oxDNA 1.5 at high salt as far as possible, but improves the structural description of duplexes (and structures built from duplexes) and allows for accurate thermodynamics at lower salt concentrations.

As a result, therefore, the discussion provided here for simulation of one version of the model largely applies to all. Moreover, it is worth noting that, even given the improved accuracy of oxDNA 2.0 in certain contexts, simulations of earlier versions of the model are still potentially valuable. oxDNA 2.0 comes into its own when it is essential to incorporate longer-range electrostatics at low salt concentrations, or when the detailed geometry of the helices are particularly important. A good example would be when simulating densely packed helices connected by crossover junctions in DNA origami (see **Section 8**). In other contexts, the reduced complexity of oxDNA 1.5 and hence its improved computational efficiency (along with slightly greater focus on basic thermodynamics and mechanics) may be beneficial. Further, it is often helpful to use oxDNA 1.0 in these contexts, as the comparison of versions 1.0 and 1.5 can help to distinguish sequence-dependent and generic effects.

In all three parameterisations, oxDNA represents each nucleotide as a rigid body with several interaction sites, namely the backbone, base repulsion, stacking and hydrogen-bonding sites, as shown in **Figure 1**. In oxDNA1.0 and oxDNA1.5, these sites are co-linear; the more realisic geometry of oxDNA2.0 offsets the backbone to allow for major and minor grooving.

Interactions between nucleotides depend on the orientation of the nucleotides as a whole, rather than just the position of the interaction sites. In particular, there is a vector that is perpendicular to the notional plane of the base, and a vector that indicates the direction of the hydrogen bonding interface. These vectors are used to modulate the orientational dependence of the interactions, which allows the model to represent the coplanar base stacking, the linearity of hydrogen bonding and the edge-to-edge character of the Watson–Crick base pairing. Furthermore, this representation allows the encoding of more detailed structural features of DNA, for example, the right-handed character of the double helix and the anti-parallel nature of the strands in the helix.

The potential energy of the system is calculated as:

$$V_0 = \sum_{\langle ij \rangle} \left(V_{b.b.} + V_{stack} + V'_{exc} \right) + \sum_{i,j \notin \langle ij \rangle} \left(V_{HB} + V_{cr.st.} + V_{exc} + V_{coax} \right), \tag{2}$$

with an additional screened electrostatic repulsion term for oxDNA 2.0. In **Eq. 2**, the first sum is taken over all pairs of nucleotides that are nearest neighbors on the same strand and the second sum comprises all remaining pairs. The terms represent backbone connectivity ($V_{b.b.}$), excluded volume (V_{exc} and V'_{exc}), hydrogen bonding between complementary bases (V_{HB}), stacking between adjacent bases on a strand (V_{stack}), cross-stacking ($V_{cr.st.}$) across the duplex axis and coaxial stacking (V_{coax}) across a nicked backbone. The excluded volume and backbone interactions are a function of the distance between repulsion sites. The backbone potential is a spring potential mimicking the covalent bonds along the strand. All other interactions depend on the relative orientations of the nucleotides and the distance between the hydrogen-bonding and stacking interaction sites.

A crucial feature of the oxDNA model is that the double helical structure is driven by the interplay between the hydrogen-bonding, stacking and backbone connectivity bonds. The stacking interaction tends to encourage the nucleotides to form co-planar stacks; the fact that this stacking distance is shorter than the backbone bond length results in a tendency to form helical stacked structures. In the single-stranded state, these stacks can easily break, allowing the single strands to be flexible. The geometry of base pairing with a complementary strand locks the nucleotides into a much more stable double helical structure.

The model was deliberately constructed with all interactions pairwise (i.e., only involving two nucleotides, which are taken as rigid bodies). This pairwise character allows us to make effective use of cluster-move Monte Carlo (MC) algorithms, which provide efficient equilibrium sampling (see **Section 8**).

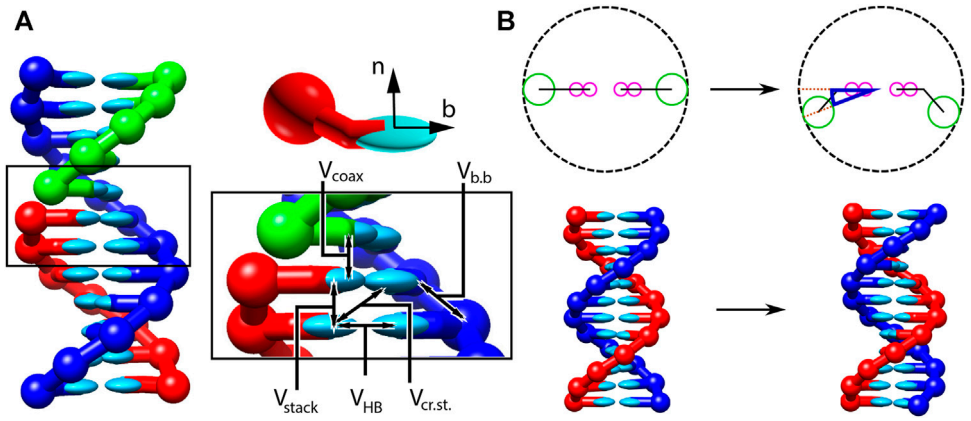


FIGURE 1 | Structure and interactions of the oxDNA model (adapted from (Snodin et al., 2015; Doye et al., 2020). **(A)** Three strands forming a nicked duplex as represented by oxDNA2.0, with the central section of the complex illustrating key interactions from **Eq. 2** highlighted. Individual nucleotides have an orientation described by a vector normal to the plane of the base (labelled n), and a vector indicating the direction of the hydrogen bonding interface (labelled b). **(B)** Comparison of structure in oxDNA1.0 and oxDNA1.5 vs oxDNA2.0. In the earlier version of the model, all interaction sites are co-linear; in oxDNA2.0, offsetting the backbone site allows for major and minor grooving.

It is convenient to use reduced units to describe lengths, energies and times in the system. A summary of the conversion of these "oxDNA units" to SI units is provided in **Supplementary Appendix A**.

4 SIMULATING THE MODEL: MOLECULAR DYNAMICS (MD) VS VIRTUAL-MOVE MONTE CARLO (VMMC)

The oxDNA model is far too complicated to approach analytically. Publicly released code to simulate oxDNA is available as a standalone package (oxDNA wiki, 2015), or as a module (LAMMPS Documentation, 2021) for the popular LAMMPS simulation software. We note that in the process of preparing this manuscript, a small error was identified in the potential as implemented in LAMMPS (see **Supplementary Appendix B**). This error is only really noticeable when simulating unpaired DNA bases. Nonetheless, we recommend that potential users of the LAMMPS implementation wait for the stable LAMMPS release in Summer 2021. The fix will be verified through an erratum attached to the original publication Henrich et al. (2018).

There are two broad types of simulation technique that can be applied to probe the model: molecular dynamics (MD) and Monte Carlo (MC). Molecular dynamics (Frenkel and Smit, 2002) algorithms evolve their constituent molecules according to Newton's laws of motion, and so are a natural choice for simulating particle systems. For coarse-grained models such as oxDNA, in which the solvent is implicit, it is necessary to include a thermostat to both set the temperature and ensure diffusive rather than ballistic dynamics. The default MD algorithm for the standalone version of oxDNA is an Andersen-like algorithm (Russo et al., 2009), in which particle velocities and angular velocities are resampled from a Boltzmann distribution with a frequency that sets the effective diffusion coefficient. In the LAMMPS implementation, the model utilises a Langevin thermostat for rigid bodies (Davidchack et al., 2015), which applies small friction- and noise-based updates to the momentum and angular momentum at each step. The relative size of these contributions sets the temperature.

A challenge of MD simulations is that when strong, short-ranged interactions are present—as in oxDNA—they place a limit on the maximum integration time step that can be used while preserving numerical stability. Interestingly both the Andersen-like and Langevin thermostats act to stabilise the simulations, allowing larger time steps to be used than if the equations of motion were integrated without noise or drag to generate energy-conserving, ballistic motion.

Both MD algorithms generate dynamical trajectories that can be used to probe system kinetics (more on this in **Section 7**). However, it is also common to use MD to take equilibrium averages over the configurations of a particular system. In the limit of small time steps, both Andersen-like and Langevin algorithms will converge on a steady state in which they sample configurations $\mathbf{x}$ from the Boltzmann distribution $p_{\mathrm{eq}}(\mathbf{x}) \propto \exp(-\beta V_0(\mathbf{x})/k_B T)$, where $V_0(\mathbf{x})$ is the potential energy of the model. How small the step size needs to be depends on a number of details, such as the strength of coupling to the thermostat. For parameters that have become an unofficial default for oxDNA, we illustrate the accuracy of the algorithm as a function of step size in **Supplementary Appendix B**.

Monte Carlo (MC) (Metropolis and Ulam, 1949) simulations are an alternative approach for sampling from the same Boltzmann distribution, but evading the drawbacks caused by the presence of a timestep altogether. In the standard MC approach (Frenkel and Smit, 2002), configurational moves $\mathbf{x} \rightarrow \mathbf{y}$ are proposed randomly, with a symmetric probability distribution that satisfies $p_{\mathrm{gen}}(\mathbf{x} \rightarrow \mathbf{y}|\mathbf{x}) = p_{\mathrm{gen}}(\mathbf{y} \rightarrow \mathbf{x}|\mathbf{y})$. If these proposed moves are accepted with a probability

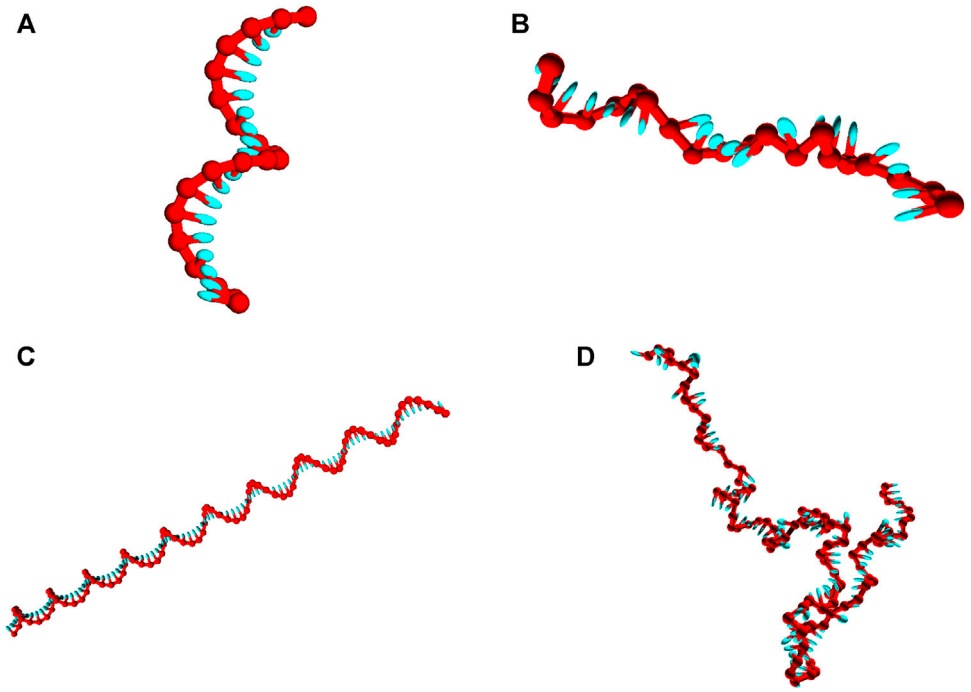


FIGURE 2 | Snapshots of poly (dT) molecules used in the equilibration time tests. Non-representative initial states of poly (dT) molecules **(left)**, and representative configurations obtained post-equilibration **(right)**. **(A)** and **(B)**: poly (dT) with 20 nucleotides; **(C)** and **(D)** poly (dT) with 100 nucleotides.

$p_{\mathrm{acc}}(\mathbf{x} \to \mathbf{y}) = \max(\exp(-(V_0(\mathbf{y}) - V_0(\mathbf{x}))/k_B T, 1))$, then the Boltzmann distribution $p_{\mathrm{eq}}(\mathbf{x}) \propto \exp(-\beta V_0(\mathbf{x}))$ is the stationary distribution of the simulation and a long simulation will sample from that distribution, assuming ergodicity.

In principle, the moves $\mathbf{x} \to \mathbf{y}$ can be arbitrarily large without leading to errors, since it is not necessary to integrate the derivative of the potential, only calculate its values at the endpoints. However, standard MC techniques incorporate sequential updates of individual particles as the moves $\mathbf{x} \to \mathbf{y}$. For a model of a strongly-attractive system such as oxDNA, these moves must be extremely small or the acceptance factor will always be small. The result is painfully slow equilibration, particularly if large scale movements of strands is required to observe it.

Virtual-move Monte Carlo (VMMC) (Whitelam and Geissler, 2007; Whitelam et al., 2009) is an alternative that circumvents the drawbacks of MC algorithms. VMMC first proposes a single particle move, then generates a co-moving cluster of particles based on which interactions are best preserved by moving the particles in unison while ensuring that the correct, detailed-balanced stationary distribution is retained. The cluster building process is based on assessing the change in pairwise interactions, and so VMMC is especially suited to oxDNA, which has exclusively pairwise interactions. We have implemented the variant from the appendix of Ref (Whitelam et al., 2009). in the standalone code.

For those with limited experience of simulating oxDNA, it is not obvious whether VMMC or MD is the optimal approach to sampling a given system. We illustrate the relative efficiencies of the two algorithms when simulating ssDNA of length 20, 100 and 1,000 bases, in terms of the computational time required to reach states representative of equilibrium from the same unrepresentative starting condition.

We simulate the poly (dT) molecules with (a) 20 bases, (b) 100 bases and (c) 1,000 bases, using oxDNA1.0 (Ouldridge et al., 2011). Simulations are performed at $T = 27°\mathrm{C}$, in a periodic box of 20, 100 and 1,000 simulation units for 20, 100 and 1,000 bases, respectively. For the MD simulations, we simulate for 60,000 simulation units of time (a nominal 182 ns), with a time step of $dt = 0.003$ (see **Supplementary Appendix B**); each simulation therefore has 2×10^7 steps in total. For VMMC, we attempted 60,000 VMMC steps per particle. The proposed moves are: rotation about a random axis, through an angle up to 0.22 radians; and translation through a distance of up to 0.22 units. These choices produce a nice balance of cluster sizes, ranging from individual nucleotides to entire strands.

Strands are initialized in a fully stacked, helical conformation as illustrated in **Figures 2A,C**, and relax to more-representative, partially-stacked conformations (**Figures 2B,D**) as the simulation is run. The relaxation of the strands is associated with an increase in the potential V_0, and so we illustrate equilibration by plotting that potential averaged over 20 independent simulations as a function of simulation progress in **Figure 3**. The average value of the potential in equilibrium, $\langle V_0 \rangle_{\mathrm{eq}}$, can be approximated by the average over the data collected in the second half of the simulations. We then estimate the equilibration time scale as the time required for $\langle V_0(t) \rangle - \langle V_0 \rangle_{\mathrm{eq}}$ to reach $1/e$ of its initial value for the first time.

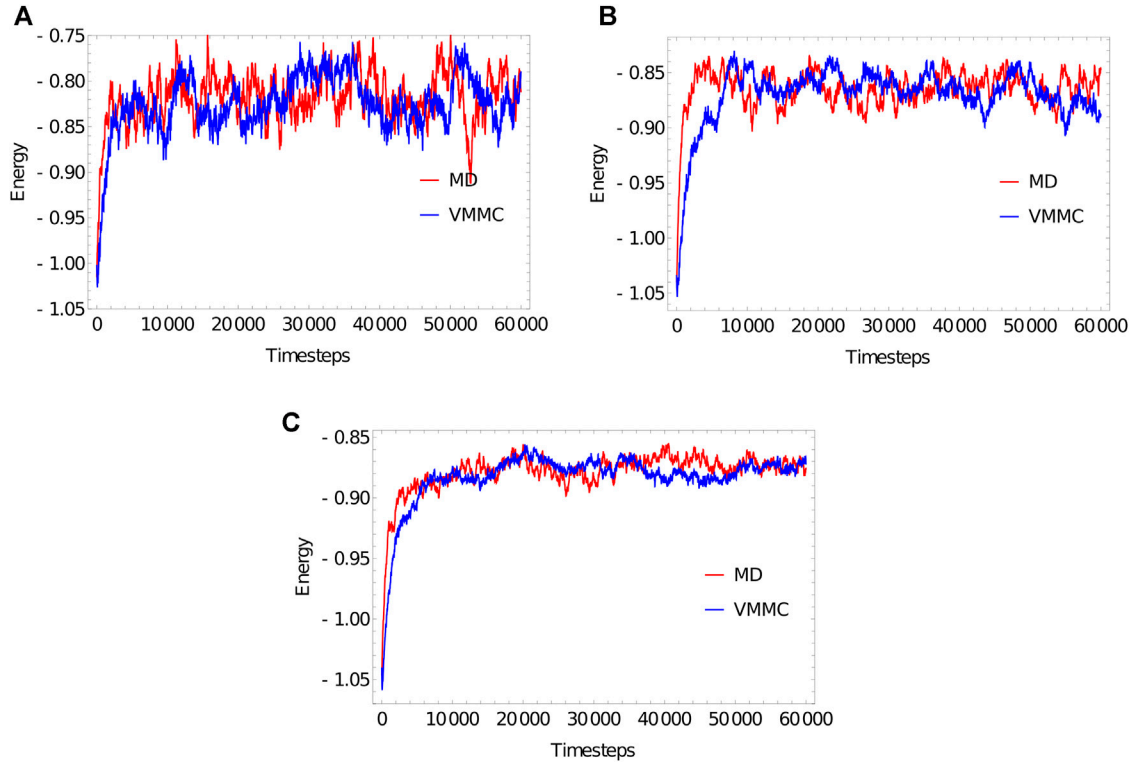


FIGURE 3 | Equilibration plots for the ploy (dT) molecules with **(A)** 20 bases, **(B)** 100 bases, **(C)** 1,000 bases, obtained as averages over 20 independent simulations. For both MD and VMMC, the potential V_0 is plotted as a function of simulation progress, in units of reduced time (3.03 ps) for MD and attempted steps per particle for VMMC.

TABLE 1 | Computational time for equilibration of poly (dT) molecules of various lengths. Simulations were performed using a single core Intel(R) Core(TM) i5-4300U CPU @ 1.90GHz.

Strand length	Total runtime (MD) in seconds	Total runtime (VMMC) in seconds	Equilibration time as a fraction of total runtime (MD)	Equilibration time as a fraction of total runtime (VMMC)
20	409.67	14.804	0.0118	0.0244
100	1804.18	147.57	0.0128	0.0287
1,000	32493.0	4187.03	0.0134	0.0299

The "simulation progress" axes in **Figure 3** are not directly comparable for MD and VMMC; one measures simulation time, the other attempted VMMC steps per particle. The most relevant quantity is the actual computational time required to equilibrate the system on a given architecture; in simple contexts, this time is also indicative of the speed with which the algorithm samples the equilibrium ensemble. **Table 1** shows the total runtime of the simulations and the equilibration time as a fraction of that runtime. In computational time, the VMMC algorithm is able to equilibrate the poly (dT) molecules more quickly (compared to MD algorithm) in all the three cases. VMMC is around 15 times as fast for the 20-nucleotide strand, dropping to around 4 times as fast for the 1000-nucleotide strand. The large moves available to VMMC, and the lack of a requirement to differentiate potentials, provide this benefit. Note, however, that the ratio of the equilibration times for MD to VMMC algorithms decreases as the system size increases; as can be seen in **Table 1**, the equilibration time for VMMC simulations increases super-linearly with system size. This super-linear increase arises because more steps must be taken to equilibrate larger systems *and* because larger clusters tend to be built for larger systems, resulting in each step taking longer on average.

The relative efficiency of VMMC and MD approaches will depend to some degree on the choice of damping parameters and seed moves; we have not carefully optimised our choices for either technique, but have used values that generally work well. The relative efficiency will also depend on the particular system: VMMC lends itself to systems in which large movements are important. Our metric for efficiency (speed with which systems reach a state that is representative of equilibrium) is also fairly

crude. Nonetheless, the general rule of thumb that VMMC is more efficient for smaller systems—particularly those with significantly fewer than 1,000 nucleotides, such as the sampling of duplex formation in oligonucleotides - is a helpful one. It is also particularly easy to enhance VMMC using umbrella sampling, as explained in **Section 6.1.1**.

For sufficiently large systems, such as DNA origami, MD should equilibrate faster, and therefore provide improved sampling. Another major advantage of the MD approach is much more facile parallelisation when simulating large systems. The standalone code allows for parallel simulation on GPUs (graphical processing units), and the LAMMPS module for parallel simulations across multiple CPUs (central processing units) using MPI. These approaches are demonstrated in **Section 8**.

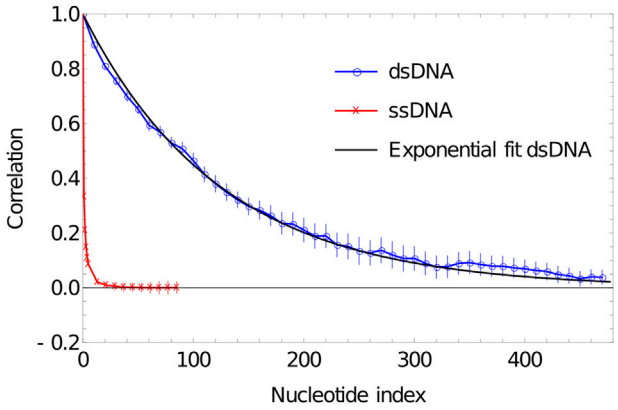


FIGURE 4 | Plot of correlation of inter-segment vectors vs distance (number of base pairs along the DNA) for 500 dsDNA (blue curve) and 100 ssDNA (red curve). dsDNA fitted with an exponential decay with a decay constant of 0.0076134.

5 MECHANICAL PROPERTIES OF DNA

The mechanical properties of DNA are central to its role across nanotechnological, biophysical and biological contexts. DNA's flexibility and response to applied stress determine the conformation and accessibility of the genome inside cells (Lewis et al., 1996; Nikolov et al., 1996; Widom, 2001; Richmond and Davey, 2003). Moreover, not only is the stiffness of dsDNA important in maintaining the conformation of DNA nanostructures, but the relative flexibility of ssDNA crucially allows for joints and flexible hinges. These properties are widely-studied in bulk and single-molecule experiments *in vitro* (Crothers et al., 1992; Smith et al., 1996; Strick et al., 1996; Wang et al., 1997; Rivetti et al., 1998; Mills et al., 1999; Podtelezhnikov et al., 2000; Dessinges et al., 2002; Bryant et al., 2003; Seol et al., 2004; Fujimoto et al., 2006; Gore et al., 2006; Lionnet et al., 2006; Seol et al., 2007; Du et al., 2008; Forth et al., 2008; Mosconi et al., 2009; Demurtas et al., 2009; Brutzer et al., 2010; Gross et al., 2011; Salerno et al., 2012; Tempestini et al., 2013; Fields et al., 2013; Le and Kim, 2014; Kim et al., 2015).

It is therefore essential that a coarse-grained model provides a reasonable representation of these properties. In this section, we both discuss the mechanical properties of oxDNA, and show how to construct simulations that can probe these properties.

5.1 Stiffness of Duplex and Single-Stranded DNA

The most common metric used to quantify the stiffness of DNA is the persistence length, defined in the textbook of Cantor and Schimmel as (Cantor and Schimmel, 1980)

$$L_{\mathrm{ps}} = \frac{\langle \mathbf{L} \cdot \mathbf{l}_0 \rangle}{\langle l_0 \rangle}. \tag{3}$$

Here, $\mathbf{L}$ is the end-to-end vector of the polymer and $\mathbf{l}_0$ represents the vector between the first two monomer units. dsDNA is most commonly thought of as a semi-flexible polymer or wormlike chain (Kratky and Porod, 1949). In this picture, the discrete series of inter-base pair vectors are approximated as a continuous, differentiable polymer axis with a quadratic free energy of curvature. For an infinitely long, semi-flexible polymer, correlations in the alignment of the polymer axis decay exponentially with separation, with a decay rate given determined by L_{ps}. When translated back to the language of inter-base-pair vectors, we obtain

$$\frac{\langle \mathbf{l}_n \cdot \mathbf{l}_0 \rangle}{\langle l_0 \rangle^2} = \exp\left(-n\langle l_0 \rangle / L_{ps}\right), \tag{4}$$

where $\mathbf{l}_n$ is the vector between base pair $n-1$ and base pair n.

It is relatively straightforward to both assess whether the wormlike chain model is a good model for oxDNA, and to extract L_{ps}. We simply simulate a duplex system for long enough to sample a representative set of configurations, calculate the correlation between inter-base-pair vectors as a function of separation, and fit the results to the exponential decay of **Eq. 4**.

In **Figure 4**, we plot the results of such a procedure. To obtain these data, we simulate a DNA duplex of length 500 base pairs at 27°C using oxDNA1.5. We perform 20 VMMC simulations with 2×10^7 attempted steps per particle for VMMC. The first 10^5 moves are treated as an initialization period and no data is collected. Additionally, the base pairs at the ends of the duplex are more flexible than those well within the bulk; to obtain properties representative of bulk DNA, we therefore do not include the five base pairs at either end in our analysis. Correlations are calculated from the default configurational outputs of the model, using the code provided in (Sengar, 2021). In order to obtain a good sample, it is helpful to output these configurations with a high frequency; we use a small value for the parameter to output energy configurations after a single VMMC move per particle.

As is evident from **Figure 5**, the correlation of the duplex axis indeed follows an exponential fall-off, to within sampling error. Fitting **Eq. 4** to $\ln\left(\frac{\langle \mathbf{l}_n \cdot \mathbf{l}_0 \rangle}{\langle l_0 \rangle^2}\right)$ gives $L_{\mathrm{ps}} \approx 131\langle l_0 \rangle = 131 \times$

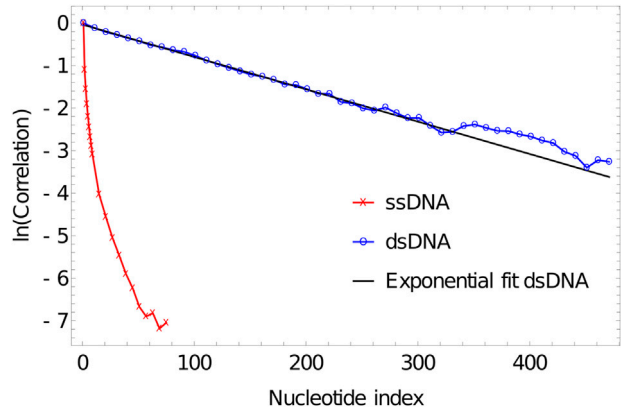


FIGURE 5 | Log plot of correlation of inter-segment vectors vs distance (number of base pairs along the DNA) for 500 dsDNA (blue curve) and 100ssDNA (red curve). dsDNA fitted with an exponential decay with a decay constant of 0.0076134.

0.4118 ≈ 53.95 simulation units= 45.91 nm, consistent with experimental estimates of 40–50 nm (120–150 base pairs) at high [Na+] concentrations (Savelyev, 2012; Herrero-Galán et al., 2013).

Indeed, more generally, the mechanical properties of double-stranded oxDNA are well-described by a semiflexible polymer model, and its torsional and extensional moduli have been analysed elsewhere (Ouldridge et al., 2011; Matek et al., 2015). Significant deviations from this behaviour - such as sharp kinks facilitated by broken base pairs—are generally only observed when large stresses are applied to the molecule (Romano et al., 2013; Matek et al., 2015), in agreement with experiment.

ssDNA behaves very differently in oxDNA. In **Figure 4**, we plot the correlations of backbone-site-to-backbone-site vector for a 100-base poly (dT) ssDNA, obtained from running simulation in oxDNA1.5 at 27°C. We perform 20 VMMC simulations with 2×10^7 attempted steps per particle. The first 10^5 moves are treated as an initialization period and no data is collected. For these simulations, we have set the stacking strength between the nucleotides to zero (the consequences of non-zero stacking strength will be addressed in Section 5.2). From **Figure 4** and **Figure 5**, it is apparent that the correlation drops very rapidly, meaning that unstacked ssDNA is very flexible in oxDNA, as it should be; adjacent backbone-to-backbone vectors can bend through a large angle. But importantly, it is worth noting that the drop in correlation between vectors with separation along the polymer cannot be well described by an exponential as in **Eq. 4**. The convexity of $\ln\left(\frac{\langle \mathbf{l}_u.\mathbf{l}_0\rangle}{\langle l_0\rangle^2}\right)$ is indicative of more distant backbone-to-backbone vectors being aligned more strongly than would be expected from the alignment of two adjacent backbone-to-backbone vectors.

The reason for this behaviour is that it is the excluded volume of nucleotides that gives unstacked ssDNA its "stiffness" in oxDNA. The excluded volume of nucleotides discourages ssDNA from folding back on itself, but importantly it leads to very different polymer properties than assumed in common polymer models such as the freely-jointed chain and the wormlike chain. For these classic polymer models, the statistical properties are entirely determined by interactions between parts of the polymer that are adjacent along the backbone, whereas the curvature of $\ln\left(\frac{\langle \mathbf{l}_u.\mathbf{l}_0\rangle}{\langle l_0\rangle^2}\right)$ in **Figure 5** is indicative of interactions between more distant points along the polymer contour playing a role.

As a result, using a wormlike chain with a given L_{ps} (or a freely jointed chain with a given Kuhn length) to understand ssDNA in oxDNA is misleading. The overall tendency of the polymer to swell to fill a large volume - due to its excluded volume - would suggest a far greater degree of local stiffness than actually present. This effect is retained even when stacking between adjacent nucleotides is included.

Importantly, these complexities also apply to physical ssDNA, as well as the oxDNA model. Single DNA have linear dimensions on the order of 1 nm, and experimental attempts to measure the mechanical properties of oxDNA (usually reported as persistence lengths) are of a similar order of magnitude (Smith et al., 1996; Rivetti et al., 1998; Mills et al., 1999; Murphy et al., 2004). Describing ssDNA in this way is not self-consistent; any polymer with this cross-section and flexibility would be strongly affected by excluded volume, so these models cannot be accurate. The result has been that experiments on large scale properties of relaxed ssDNA (Rivetti et al., 1998; Murphy et al., 2004), which are sensitive to excluded volume effects, tend to produce larger estimates for quantities like L_{ps} than experiments on shorter sections of ssDNA, or ssDNA under high tension (Smith et al., 1996; Rivetti et al., 1998).

Low salt concentrations, which lead to weaker screening of electrostatic interactions between non-adjacent nucleotides make the above effect stronger (Smith et al., 1996; Dessinges et al., 2002). Base-pairing interactions in non-homopolymeric ssDNA have a confounding effect; the formation of secondary structure tends to condense the strand, making it appear more flexible when its statistics are modelled with a wormlike chain or a freely-jointed chain (Smith et al., 1996). Overall, as for oxDNA, simple descriptions of the mechanical properties of physical ssDNA should be treated with caution.

5.2 Response of ssDNA to Tension

A common mechanism for probing the mechanical properties of DNA is to apply force, whether torsional (Strick et al., 1996; Bryant et al., 2003; Forth et al., 2008; Mosconi et al., 2009; Brutzer et al., 2010; Salerno et al., 2012; Tempestini et al., 2013), extensional (Smith et al., 1996; Strick et al., 1996; Wang et al., 1997; Dessinges et al., 2002; Seol et al., 2004; Gore et al., 2006; Seol et al., 2007; Huguet et al., 2010; Gross et al., 2011) or shearing (Forth et al., 2008; Hatch et al., 2008; Mosconi et al., 2009; van Mameren et al., 2009; Brutzer et al., 2010; Salerno et al., 2012; Tempestini et al., 2013; Wang and Ha, 2013).

Experiments have focused on both the elastic properties associated with small deformations (Smith et al., 1996; Strick et al., 1996; Wang et al., 1997; Bryant et al., 2003), and large scale structural transitions (Smith et al., 1996; Strick et al., 1996; Dessinges et al., 2002; Bryant et al., 2003; Seol et al., 2004; Gore et al., 2006; Seol et al., 2007; Hatch et al., 2008; Forth et al., 2008; Mosconi et al., 2009; van Mameren et al., 2009;

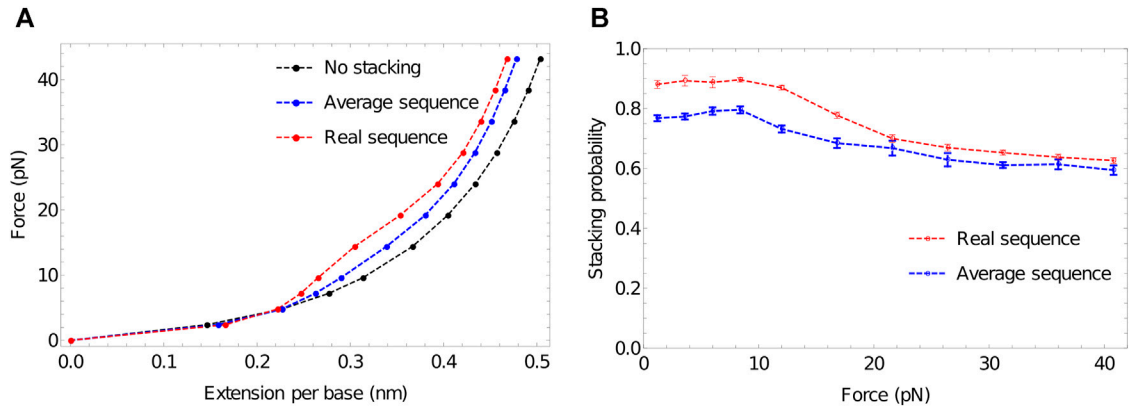


FIGURE 6 | The response of ssDNA to tension, and the role of stacking therein. **(A)** Force-extension plots for 100-nucleotide poly (dA) using three models: no stacking (black); average stacking strength (oxDNA1.0, blue) and sequence-dependent stacking (oxDNA1.5, red). Stronger stacking leads to an increased force at larger extensions, and extremely strong stacking results in a plateau-like feature as stacking is disrupted. **(B)** Stacking probability for average stacking strength (oxDNA1.0, blue) and sequence-dependent stacking (oxDNA1.5, red) as a function of applied force. Adjacent nucleotides are defined as stacked if the stacking energy between the pair is more than -0.1 units.

Brutzer et al., 2010; Gross et al., 2011; Salerno et al., 2012; Wang and Ha, 2013; Tempestini et al., 2013).

Applying external tension is relatively straightforward in molecular simulation; there are more subtleties associated with applying boundary conditions for external torsion (Matek et al., 2012; Matek et al., 2015), but it is also possible. In the case of oxDNA, small external stresses have been used to help parameterise and characterise the model (Ouldridge et al., 2011; Matek et al., 2015; Skoruppa et al., 2017; Nomidis et al., 2019); larger stresses have been applied to provide insight into experiments on structural transitions (Matek et al., 2012; Romano et al., 2013; Wang and Pettitt, 2014; Matek et al., 2015; Mosayebi et al., 2015; Wang et al., 2015; Engel et al., 2018; Desai et al., 2020).

Systems with internally-induced stress, where the drive to form base pairs in one part of an assembly applies stress to another part, have also been studied (Harrison et al., 2015; Sutthibutpong et al., 2016; Wang and Pettitt, 2016; Wang et al., 2017; Tee and Wang, 2018; Caraglio et al., 2019; Harrison et al., 2019; Engel et al., 2020; Fosado et al., 2021; Park et al., 2021).

As an example, in this section we demonstrate the force-extension properties of ssDNA as represented by oxDNA. Optical tweezer experiments with ssDNA have a long history (Smith et al., 1996). These original experiments with naturally-occurring DNA exhibited formation and stabilization of secondary structure in high salt conditions and low-moderate force, although this was not explicitly modelled at the time. The presence of this secondary structure makes simulation of DNA heteropolymers hard; it is challenging to equilibrate a long strand with many competing base-pairing configurations (we have had some success using methods based on parallel tempering (Romano et al., 2013). Instead, therefore, we simulate 100-nucleotide-long homopolymeric poly (dA) using oxDNA1.0 and oxDNA1.5.

The helicity in oxDNA is driven by stacking interactions between adjacent nucleotides. As is evident from **Figure 2**, this stacking has a residual effect on the structure of ssDNA strands, which are partially stacked in equilibrium. We will use force-extension simulations to probe the consequences of single-stranded stacking in oxDNA.

For these simulations, which are similar to original results in (Šulc et al., 2012), we use both a version of the parameters with no stacking interacting, a sequence-averaged stacking interaction (oxDNA1.0), and a sequence-specific stacking interaction (oxDNA1.5) for which poly (dA) has the strongest interaction of all sequences. All simulations are performed at $T = 27°C$, in a periodic box of length 100 simulation units with each simulation running for 4×8 steps with $dt = 0.005$ which equals to a total run time of 2×10^6 units. Four sets of simulations are performed for 12 difference values of force.

Figure 6A shows that extensive stacking has only a moderate effect on the force-extension properties of the ssDNA at low force. In the sequence-dependent model, poly (dA) is close to 90% stacked at 27°C—see **Figure 6B**. However, the increased stiffness due to the tendency to form stacked single helices (akin to the initial state in **Figure 3**) is counteracting by the shorter end-to-end distance of the backbone when it is forced to wind around the helix.

At larger forces, however, we clearly see a signal of stronger stacking. Larger force is required to extend the strands with stronger stacking, and a plateau-like feature is evident in the system with the strongest stacking. A similar plateau was observed by Seol *et al.* for RNA stretching (poly(A) and poly(C)) (Seol et al., 2004; Seol et al., 2007) but was absent for poly(U). Those authors hypothesised that the plateau arises as the shorter end-to-end distance in helical stacked confirmations becomes prohibitive; additional force is then required to disrupt the stacking interaction to allow further extension, causing an increase in the gradient of the force-extension curve. After the bases have unstacked, the gradient becomes less steep again.

Broadly speaking, this explanation is borne out by oxDNA. Notably, however, Seol *et al.* (Seol et al., 2004; Seol et al., 2007)

concluded that a relatively low stacking probability should give a pronounced plateau. By contrast, in **Figure 6B**—obtained by probing configuration output files to assess the degree of stacking (Sengar, 2021) we see that strands with an initial stacking probability of 78% show only a hint of the plateau. This discrepancy arises because, in the minimal model of Seol *et al.* (Seol et al., 2004; Seol et al., 2007), even a single pair of stacked nucleotides has a much shorter end-to-end distance along the ssDNA backbone than an unstacked pair. The explicit representation of 3D structure in oxDNA, however, captures the fact that the shortening of the end-to-end distance along the DNA backbone is only significant when several bases in a row are stacked into a helix, so that the backbone really has to wrap back round upon itself. As a result, extension can occur while disrupting only a fraction of the stacking interactions, and ssDNA in oxDNA remains significantly stacked even at high force (**Figure 6B**).

While oxDNA's representation of the polynucleotide backbone is simplistic, these geometrical arguments also apply to physical DNA - suggesting that even weak plateau-like behaviour in ssDNA force-extension curves is evidence of strong stacking, and the absence of a plateau is not proof of an absence of stacking. More generally, this system is indicative of the value that oxDNA can provide. The system involves an interplay between basic structure, mechanics and thermodynamics of ssDNA. When applied, oxDNA reveals subtleties that are not directly apparent from a more minimal model. Indeed, it is quite common to construct very simple models to interpret biophysical experiments on the mechanical properties of DNA (Hatch et al., 2008; Qu and Zocchi, 2011; Salerno et al., 2012; Vafabakhsh and Ha, 2012; Fields et al., 2013; Tempestini et al., 2013; Wang and Ha, 2013; Meng et al., 2014); simulations with oxDNA often reveal physically reasonable relaxation mechanisms that aren't factored into these simpler models (Harrison et al., 2015; Matek et al., 2015; Mosayebi et al., 2015; Skoruppa et al., 2017; Harrison et al., 2019). At this stage, it is also worth highlighting a general virtue of coarse-grained models that is apparent in these simulations. It is very simple just to switch off interactions—such as the stacking here—to isolate the effect those interactions have on the system. Doing so can be incredibly helpful in interpreting the physical cause of experimental signals.

6 THERMODYNAMIC SIMULATIONS WITH OXDNA

6.1 Duplex Formation Thermodynamics

As well as representing the structure and mechanical properties of ssDNA and dsDNA, oxDNA is also designed to capture the thermodynamics of the hybridization transition from ssDNA to dsDNA. Needless to say, accurately capturing the thermodynamics of this transition is essential for any model hoping to describe biological and nanotechnological processes involving the forming and disruption of base pairs.

To assess the thermodynamics of a simple duplex, it is typical to simulate an isolated pair of strands in a periodic cell that is large enough to prohibit self-interactions (unit cell size of $\gtrapprox 2n$ oxDNA length units, where n is the duplex length, is generally sufficient). Given a sufficiently long VMMC or MD simulation, the fraction of time spent in the bound state can be estimated and used to infer quantities such as melting temperatures, as outlined below.

However, particularly for longer strands, simulating this process can be prohibitively slow. For two short strands in solution, the vast majority of configurations have well-separated strands and no base-pairing interactions. Enthalpically favourable base-pairing provides a compensatory advantage to configurations with many well-formed base pairs (fully-formed duplexes). To obtain a good estimate of the fraction of strands bound in equilibrium, it is necessary to pass between these two sub-ensembles (completely unbound and fully bound) many times; as a rule of thumb, we have found that around 10 interconversions will start to provide meaningful statistics.

Unfortunately, interconversion requires the system to transition through states with only one or two base pairs that benefit neither from the large ensemble of configurations accessible to dissociated strands, nor the favourable interactions of fully-bound strands. These configurations with $Q = 1, 2$ base pairs are rare in the equilibrium ensemble, and have a relatively high free energy

$$F(Q) = -kT\ln\left(p_{\mathrm{eq}}(Q)\right) + C, \tag{5}$$

where C is a Q-independent constant, and $p_{\mathrm{eq}}(Q)$ is the probability of observing Q base pairs in equilibrium. The high free energy of these intermediate states makes dissociation and association rare event processes that are challenging to sample directly.

6.1.1 Umbrella Sampling

To overcome this difficulty, simulations can be augmented with umbrella sampling (Torrie and Valleau, 1977). For a system with coordinates x (in our case, nucleotide positions and orientations), umbrella sampling involves identifying a collective order parameter $\lambda(x)$ for the transition of interest, and then applying a bias $W(\lambda(x))$ to force the system to occupy otherwise undesirable values of $\lambda(x)$ that lie along the transition path more frequently. Unbiased statistical averages can be extracted from these biased samples using

$$\langle A(x)\rangle_{\mathrm{eq}} = \left\langle \frac{A(x)}{W(\lambda(x))} \right\rangle_{\mathrm{biased}}, \tag{6}$$

where $A(x)$ is a quantity of interest. Essentially, the contribution of each configuration sampled to the average is reduced by a factor of the bias applied.

A common approach with umbrella sampling is to perform a series of separate simulations with very strong biases tightly centred on distinct values of $\lambda(x)$. Simulations centred on adjacent values of $\lambda(x)$ can then be knitted together using procedures such as the Weighted Histogram Analysis Method (WHAM), allowing the calculate of the free energy difference between the start and end point (Kumar et al., 1992).

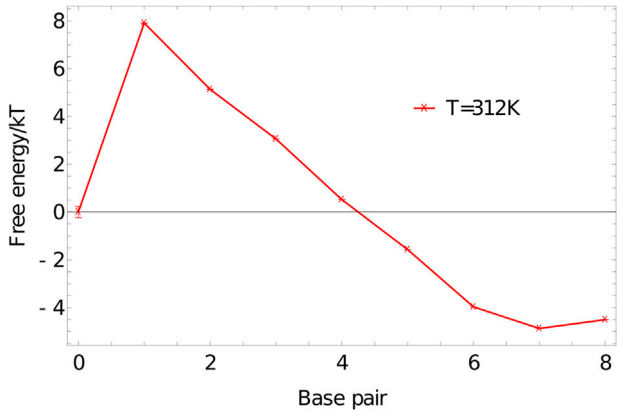


FIGURE 7 | Free-energy profile of an 8-base-pair duplex (3′-ACTGACGT-5′ and 3′-ACGTCAGT-5′) at 312K in a simulation volume of side length 15 units.

Generally, however, we have found that this sophisticated approach is not necessary for oxDNA, and a particularly straightforward umbrella sampling method is built into the standalone oxDNA code. When using VMMC, it is possible to specify discrete order parameters $\lambda(x)$ based on the number of base pairs between user-defined groups of nucleotides. For duplex formation, it is fairly straightforward to iteratively identify a biasing potential that facilitates both the sampling of all states and the rapid transition between fully bound and completely detached configurations.

This biasing potential doesn't need to be fine tuned so that all values of $\lambda(x)$ are equally probable in the biased sample—just good enough to facilitate multiple transitions backwards and forwards. Typical examples for 5-base and 8-base duplexes are given in (Sengar, 2021). For more complex systems, more sophisticated $\lambda(x)$ and the use of multiple sampling windows are sometimes necessary—we refer the reader to (Ouldridge et al., 2010a; Ouldridge et al., 2013a; Machinek et al., 2014; Harrison et al., 2019). Even in these cases, however, the principles are similar to those outlined here.

We perform umbrella sampling simulations on an 8-nucleotide duplex at 312K in a simulation volume of side length 15 units, using the oxDNA1.5 version of the model. Five independent simulations are performed for 7.7×10^8 VMMC steps per particle. The quantity $p_{\mathrm{eq}}(Q)$ obtained from simulations is used to calculate a free energy $F(Q)$ according to **Eq. 5** and plotted in **Figure 7**. The shape of this graph is typical for duplex formation, showing the expected large jump in free energy from 0 to 1 base pairs. From 1 to 6 base pairs there is a steady drop in the free energy as configurations are stabilised by additional base-pairing interactions that are favoured once the strands are in close proximity. The final base pairs are less favourable, as base pairs at the end of a duplex are prone to fraying (SantaLucia and Hicks, 2004; Ouldridge et al., 2011).

The shape of $F(Q)$ gives a good guide to constructing first estimates of umbrella biases $W(\lambda(x))$ for duplex formation in general. Ignoring the dissociated state, $W(\lambda(x))$ should increase roughly exponentially with the number of base pairs broken, since it must counteract $\exp(-F(Q)/kT)$. The slope of $F(Q)$, and hence the required rate of exponential growth in $W(\lambda(x))$, is determined by the temperature; as a crude rule of thumb, a bias of a factor of 10–15 is required per base pair broken at 300K; this required bias falls to a factor of 3-4 by 330 K.

The initial jump in free energy from 0 to 1 in 7 is largely determined by the simulation volume; for simulation cells similar in size to this one, a factor of 3000–10000 is a reasonable first guess for the required weight of the 1-base-pair state relative to the 0-base-pair state.

In addition to biasing by the number of base pairs formed, it is sometimes helpful to also use a distance-based contribution to the order parameter. Built in to the standalone oxDNA code is the ability to define additional dimensions of $\lambda(x)$ that depend on the minimum separation between sets of nucleotides, rather than the number of base pairs. We have found that a simple division of the 0-base-pair state into configurations in which the strands are close (less than 4 units apart) and far apart (4 or more units apart) can reduce the amount of time spent sampling the independent diffusion of strands around the simulation volume. For simulation volumes similar to this one, the close state should be weighted by around 5–10 relative to the distant state, which dominates the unbound ensemble.

6.1.2 Melting Temperature Curves

Although free-energy profiles for a single pair of strands are informative, they aren't directly comparable to the majority of experiments. Indeed, the thermodynamics of the oxDNA model was parameterised to reproduce the nearest neighbour model (SantaLucia, 1998), which in turn was fitted to—and predicts—experimental melting curves in bulk conditions. The Santalucia parameterisation of the nearest neighbour model (SantaLucia and Hicks, 2004) assumes that DNA duplex formation is essentially a two-state transition between a well-formed duplex and separated single strands. The free-energy profiles produced by oxDNA, such as **Figure 7** are consistent with this picture; the ensemble is dominated by configuration with either zero base pairs, or a large number. In this limit, the melting behaviour can be well-characterised by the fraction of strands that are expected to have base pairing with another strand at a temperature T in a bulk system, $f_\infty(T)$.

To calculate $f_\infty(T)$ in this two-state description, it is first necessary to obtain data at a range of temperatures. In principle, these data can be obtained through separate simulations. However, we have found that a technique called single histogram reweighting (Ferrenberg and Swendsen, 1988) is sufficient to infer $f_\infty(T)$ accurately over a large enough range of temperatures to describe the melting transition. The basic idea is to treat a simulation at a temperature T as a biased sample of the ensemble at another temperature T'; this bias can be corrected in the same way as the bias applied during umbrella sampling:

$$\langle A(x)\rangle_{T'} = \langle A(x)\exp(V_0(x,T)/kT - V_0(x,T')/kT')\rangle_T. \quad (7)$$

Here $V_0(x,T)$ is the value of the potential in the original simulation at temperature T ($V_0(x,T')$ is slightly different due to

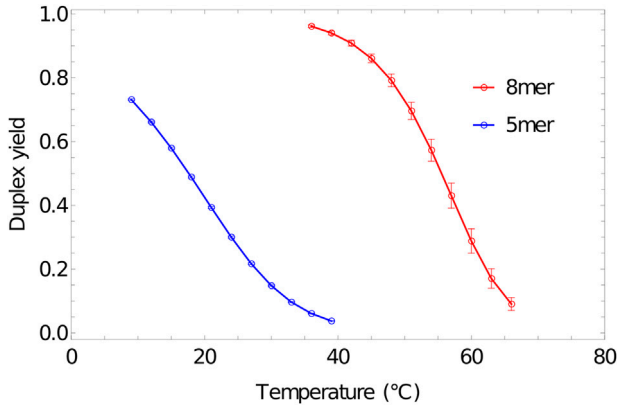


FIGURE 8 | Melting transition of oligonucleotides. Fractional yield of 5mer (3′-AGTCT-5′/3′-AGACT-5′) and 8mer (3′-ACTGACGT-5′/3′-ACGTCA GT-5′) duplexes in bulk ($f_\infty(T)$)) as predicted by oxDNA1.5 at a total concentration of 7.96×10^{-4} M for each strand.

a T-dependent term in the potential (Ouldridge et al., 2011)). Extrapolation to nearby temperatures using single histogram reweighting is built into the oxDNA standalone code. It is important to note that if umbrella sampling, and particularly temperature reweighting, are applied, then it is especially important to simulate for a good equilibration time before results are collected. Normally, any initial unrepresentative states will be swiftly overwhelmed within an average taken over the whole course of the simulation. The unbiasing factors in **Eq. 6** and **Eq. 7**, however, can cause unrepresentative initial states to be assigned enormous weights in the ensemble average that are effectively insurmountable, rendering the simulation results meaningless.

Given well-sampled data of the formation of a single duplex in a simulation volume, it is tempting to assume that the fractional yield of states with more than one base pair in a single duplex simulation, $f_1(T)$, is equal to the bulk yield of duplexes $f_\infty(T)$ in a system with the same total concentration of strands. Unfortunately this is not the case; simulations of only a single target duplex neglect concentration fluctuations within unit cells that have large effects on the yield of products (Ouldridge et al., 2010b; Ouldridge, 2012). Quantitative comparison to experimental data is therefore impossible unless extrapolations to bulk conditions can be performed. Assuming ideal behaviour of solutes, Extrapolation is possible. For dimerisation between non-self-complementary strands (Ouldridge et al., 2010b)

$$f_\infty(T) = \left(1 + \frac{1}{2\Phi(T)}\right) - \sqrt{\left(1 + \frac{1}{2\Phi(T)}\right)^2 - 1}, \tag{8}$$

where $\Phi(T) = f_1(T)/(1 - f_1(T))$. A similar result holds for self-complementary duplexes (Ouldridge et al., 2010b), and algorithms exist to extrapolate to bulk for more complex assemblies (Ouldridge, 2012).

Melting curves obtained for 5-base and 8-base duplexes, using umbrella sampling, temperature reweighting and extrapolation to bulk, are reported in **Figure 8**. The melting temperatures T_m for these duplexes – defined, in the two state model, as the temperature at which $f_\infty(T)$ is 0.5 – are close to the values predicted by the nearest neighbour model at the same conditions (17.8°C and 56.1°C) (SantaLucia and Hicks, 2004). This agreement is, of course, due to the model being fitted to these data. However, it is worth noting that although duplexes are often described as having a single "melting temperature", the temperature at which $f_\infty(T)$ is 0.5 depends on the concentration of the individual strands, $[C]$ with (Ouldridge et al., 2011)

$$\frac{\mathrm{d}T_m}{\mathrm{d}[C]} \sim \frac{\Delta T}{[C]}. \tag{9}$$

Here, ΔT is the width of the transition over which $f_\infty(T)$ goes from largely bound to largely unbound. To match nearest neighbour predictions for melting temperatures over a range of concentrations, therefore, it is necessary that transition widths are also comparable; achieving a good match was a major part of oxDNA's parameterisation.

6.2 Thermodynamics of More Complex Structures

Although accurately simulating basic duplex formation was necessary for the parameterisation of oxDNA, little new information is to be gained from performing these simulations again. The model is trained to reproduce the thermodynamics of the nearest neighbour model, so simulating the thermodynamics of duplex formation is an expensive way to get at an approximation to said nearest neighbour model.Where oxDNA can add value is if duplex formation occurs as part of some more complex system - possibly one in which internally or externally-applied stresses, or topological constraints, are relevant (Romano et al., 2012; Šulc et al., 2012; Ouldridge et al., 2013a; Mosayebi et al., 2015; Kočar et al., 2016; Fonseca et al., 2018; Tee and Wang, 2018; Harrison et al., 2019). As an example, we simulate the formation of a small pseudoknotted structure (**Figure 9**) leveraging the intuition and techniques discussed in **Section 6**. Here, the two sequences 3′-AGCTTCCATG-5′ and 3′-AAGCTCATGG-5′ cannot form a single continuous duplex, but can form two 5-bp duplexes section if both strands bend back on themselves. The stability of this structure cannot be inferred from the nearest neighbour model, but it can easily be simulated with oxDNA. Applying umbrella sampling, we simulate the system at a temperature of 308K in a periodic cell of side-length 20 using oxDNA1.5.

The resulting free-energy profile, **Figure 10**, shows that, at the temperatures of interest, forming two arms is less favourable than forming only one. The advantage obtained by bringing the strand into close proximity via the binding of the first duplex is not enough to overcome the internal stress generated by the structure. This internal stress is evidenced by the much shallower slope of the free energy profile for forming base pairs 6–10 than 1-5.

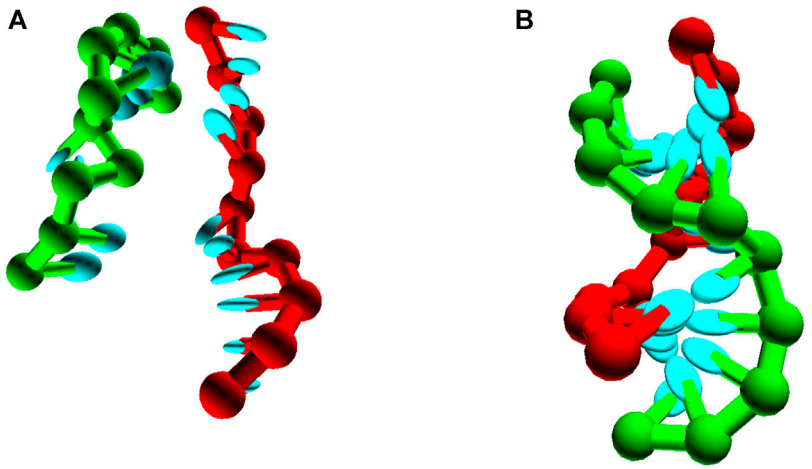


FIGURE 9 | Pseudoknot unbound **(A)** and bound **(B)** states, for seqeunces 3′-AGCTTCCATG-5’/3′-AAGCTCATGG-5’.

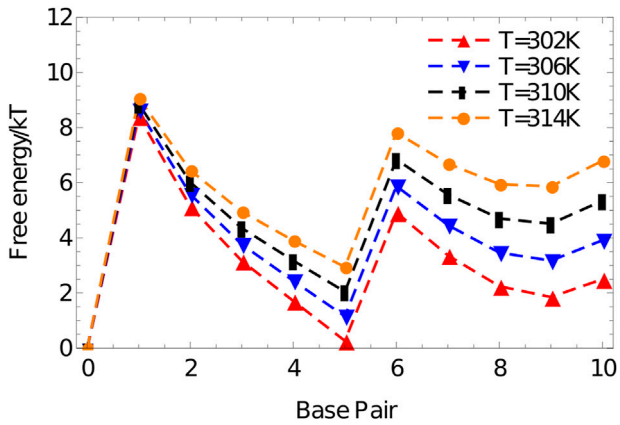


FIGURE 10 | Free energy vs number of base pairs formed for the complex 3′-AGCTTCCATG-5’/3′-AAGCTCATGG-5′ in a simulation volume of side length 20 oxDNA units.

7 DYNAMICAL SIMULATIONS

The simulations described hitherto probe static quantities obtained in the equilibrium ensemble. However, the dynamics of DNA-based systems can be equally important. In particular, the time required for reactions to happen is crucial when constructing complex self-assembling systems or functional circuits, particularly those that are intended to remain out of equilibrium, or exhibit an extremely slow relaxation to equilibrium (Dunn et al., 2015; Srinivas et al., 2017; Fern and Schulman, 2018; Cabello-Garcia et al., 2021).

Unlike the thermodynamic and structural properties, oxDNA has not been carefully parameterised to the dynamics of physical DNA. Coarse-graining is generally known to speed up timescales by smoothing free-energy landscapes (Murtola et al., 2009). Moreover, the explicitly dynamical algorithms (particularly the Andersen-like thermostat) give a fairly crude approximation to the dynamics expected from small molecules in solution. Neither the Anderson-like nor the Langevin thermostat incorporates cooperative hydrodynamics (an updated version of the Langevin thermostat developed to describe hydrodynamic effects (Davidchack et al., 2017) has not yet been implemented in LAMMPS), and both are typically run with large effective diffusion coefficients to enhance sampling (see **Supplementary Appendix A**).

Nonetheless, the dynamics of oxDNA is fundamentally constrained by the combination of its free-energy landscape and its embedding of that free-energy landscape in an explicit geometrical description. For comparison, it is surprisingly difficult to generate meaningful dynamics based on just the free-energy landscape predicted by the nearest neighbour model without an explicit geometrical representation (Srinivas et al., 2013; Schaeffer et al., 2015) [add commented-out citation].

As a result, dynamical simulations of oxDNA can provide useful insight into dynamical properties of physical DNA; the model has been particularly successful in describing toehold-mediated strand displacement (Srinivas et al., 2013; Machinek et al., 2014; Haley et al., 2020; Irmisch et al., 2020), one of the fundamental reactions of DNA nanotechnology. Importantly, the focus should always be on comparing the relative dynamics of two similar systems – for example, the dependence of strand displacement rates on toehold lengths. Unlike the thermodynamic and mechanical properties of oxDNA, absolute values of dynamical properties are largely irrelevant.

As an example, we simulate the dissociation kinetics of duplexes of length 4 (3′-ATAT-5’/3′-ATAT-5′), 5 (3′-ATATA-5’/3′-ATATA-5′) and 6 (3′-ATATAT-5’/3′-ATATAT-5′) at 320 K using the Anderson-like thermostat applied to oxDNA1.5, **Figure 11**. Example trajectories, showing the energy of the system per nucleotide, illustrate the two-state nature of the system discussed earlier in **Section 6**. The strands spend a substantial amount of time in states with an energy of approximately -1.0 in oxDNA units (duplex configurations), before suddenly transitioning to states with an

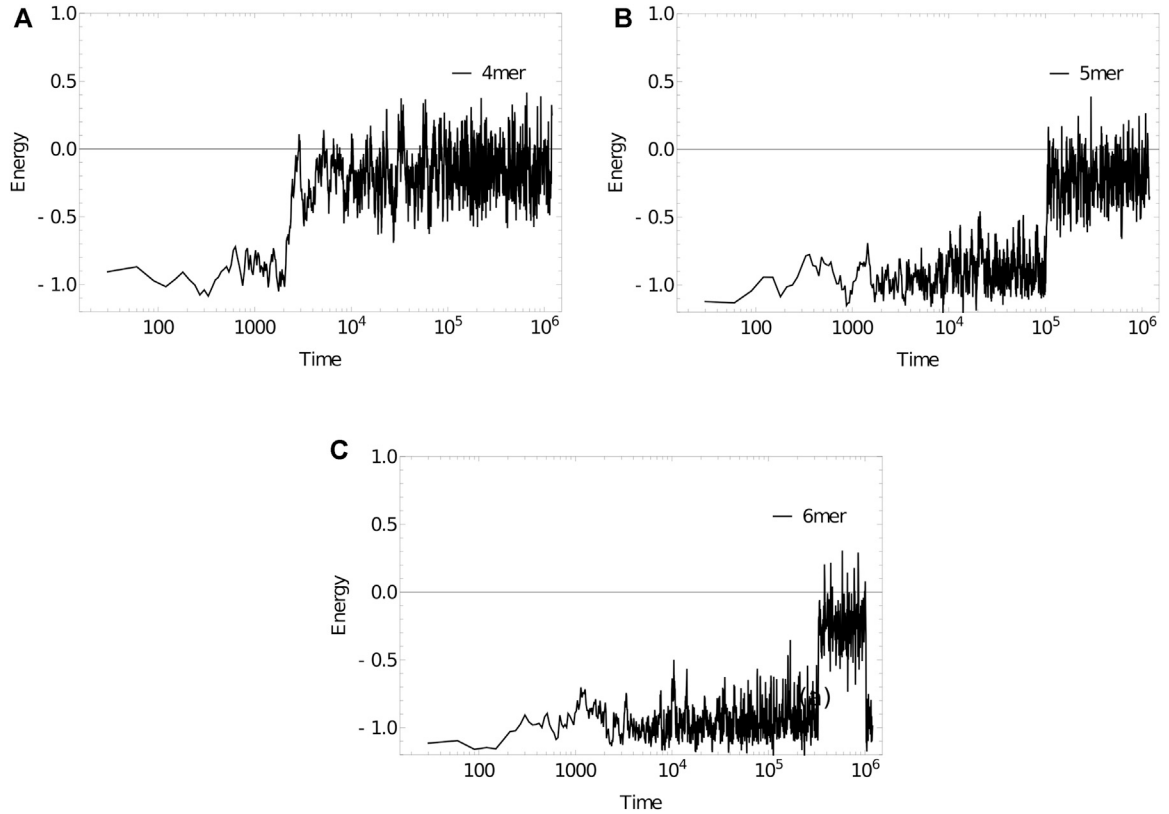


FIGURE 11 | Energy per nucleotide vs simulation time step for (a) 4mer, (b) 5mer, (c) 6mer duplexes at 320 K. Sudden transitions from low to high energy are indicative of rare-event melting.

energy around -0.2 (single-stranded states). As hinted at by these examples, longer strands take exponentially longer to dissociate (the simulation steps taken to reach an energy of -0.2 per nucleotide, averaged over 10 simulations for each length, are: 9117 ± 2883, 64032 ± 20249, 279744 ± 88463). This exponential suppression of the dissociation rate with strand length is consistent with dissociation being a rare event that requires the crossing of a free energy barrier whose height grows linearly with duplex length at fixed temperature (here, 320 K), as suggested by the free-energy profile in **Figure 7**.

In this case, all systems studied showed the required behaviour on relatively short time scales. Frequently, it is necessary to simulate much slower processes. We have found that the forward flux sampling (FFS) technique (Allen et al., 2009) is an effective tool for simulating dynamical processes with a longer timescale. However, FFS is trickier to implement than umbrella sampling, and is not yet built in to the released code in an optimal way.

8 Simulation of Large Structures

Another significant application area for oxDNA has been the simulation of large structures to assess their conformation, stability and flexibility (Fernandez-Castanon et al., 2016; Schreck et al., 2016; Sharma et al., 2017; Shi et al., 2017; Benson et al., 2018; Choi et al., 2018; Coronel et al., 2018; Berengut et al., 2019; Brady et al., 2019; Hoffecker et al., 2019; Snodin et al., 2019; Berengut et al., 2020; Chhabra et al., 2020; Poppleton et al., 2020; Tortora et al., 2020; Yao et al., 2020).

In this context, oxDNA represents an alternative to the CanDo model and simulation package (Castro et al., 2011). The added complexity of oxDNA has a computational cost, but means that it is better able to handle irregular systems. For such simulations, use of oxDNA2.0 is strongly recommended given its better representation of structure, particularly in the context of DNA origami. A more detailed primer on setting up these simulations can be fund in Ref (Doye et al., 2020); here we focus only on technical aspects of the simulations.

As briefly mentioned in **Section 4**, MD algorithms can facilitate the simulation of really large systems by allowing parallelisation across GPU threads or multiple CPUs. The oxDNA standalone code is GPU-enabled via the CUDA C API and supports runs on single CPUs and single GPUs, whereas the LAMMPS version of oxDNA uses the Message Passing Interface (MPI) and is optimised for parallel runs on multi-core CPUs and distributed memory architectures.

To provide benchmarks and examplar codes, we have performed large-scale simulations with both implementations on two different compute architectures, namely a NVIDIA V100 PCIe GPU with 5,120 CUDA cores at Arizona State University's High Performance Computing Facility, and the

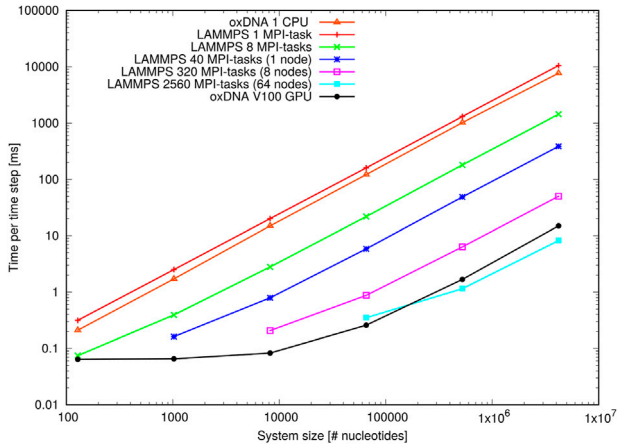


FIGURE 12 | Performance of the oligomer benchmark as time per time step for various system sizes: Shown are results of the oxDNA standalone code on a single CPU and NVIDIA V100 GPU and of the LAMMPS implementation of the oxDNA2 model at different CPU counts.

ARCHIE-WeSt HPC facility at the University of Strathclyde consisting of 64 Intel Xeon Gold 6138 (Skylake) processors @2.0GHz with 40 cores per node and 2,560 cores in total. The GPU and single-core CPU runs were performed with the oxDNA standalone code SVN version 6989. The GPU runs all used mixed precision and an edge-based approach (Rovigatti et al., 2015). The LAMMPS stable version from March 3, 2020 was used for the multi-core CPU runs (the small error in the LAMMPS implementation highlighted in **Supplementary Appendix B** is irrelevant for the purposes of these efficiency comparisons). All runs were performed with the oxDNA2.0 model featuring sequence-dependent stacking and hydrogen-bonding interactions.

Two different benchmarks were studied to analyse the performance of both implementations. The first one consisted of a varying number of double-stranded octamer duplexes and investigated the performance at different system sizes, ranging from 8 octamers with 128 nucleotides in total to 262,144 octamers with 4,194,304 nucleotides in total. The concentration of octamers was kept constant at one octamer per 20^3 oxDNA length units, whereas the temperature and salt concentration were set to $T = 293K$ and $[Na^+] = 500$ mM, respectively.

The second benchmark consisted of a DNA origami "pointer" structure (Bai et al., 2012) (15,238 nucleotides) and tested the performance at different salt concentrations between $[Na^+] =$ 100 mM and 1 M. The salt concentration is another performance-critical aspect in the simulation of nucleic acids that is often neglected. The reason is that the salt concentration affects the Debye screening length, which is proportional to the inverse square root of the salt concentration. The temperature of this second benchmark was fixed at $T = 293K$. The initial configuration was converted from the cadnano format using the TacoxDNA server (Suma et al., 2019), then relaxed using oxdna.org (Poppleton et al., 2021), implementing the protocol from (Doye et al., 2020) followed by a simulation at the respective salt concentration.

It is worth emphasising that origami structures such as the pointer are a setting in which the improved structural model of oxDNA2.0 is essential. Unless an accurate model is used, relatively small discrepancies can contribute strain that builds up across the structure, resulting in large scale distortion.

Figure 12 shows the results of the oligomer benchmark, which are expressed as time per integration time step in milliseconds. On a single Intel Xeon Gold CPU the standalone code implements a single timestep slightly faster than the LAMMPS implementation. Note, however, that the actual efficiency will depend on the choice of coefficients of coupling to the thermostats (see **Supplementray Appendix B**).

When deployed in parallel on more CPUs, the LAMMPS implementation offsets this disadvantage almost immediately. Its performance at the larger side of system sizes is more or less ideal as evidenced through the linear increase of time per integration step with system size. For smaller system sizes, and depending on how many CPUs were used, the performance levels off due to a build-up of MPI communication overheads. However, there is still a noticeable speed-up e.g. for 8,192 nucleotides on 320 MPI-tasks or 65,536 nucleotides on 2,560 MPI-tasks, which comes down to a very low 25 nucleotides per MPI-task. This unusually good performance of a parallel molecular dynamics code has been reported before (Henrich et al., 2018) and is owed to the rather complex oxDNA force field as the code spends a good deal of time carrying out the force calculation.

The GPU-implementation of the standalone code retains a significant advantage over the LAMMPS implementation for all but the largest benchmark sizes and runs on the full ARCHIE-WeSt system size (2,560 MPI-tasks) and its performance levels only off when the GPU becomes under-subscribed with threads at smaller system sizes. We can conclude that the LAMMPS implementation of oxDNA, besides its capability to run on a

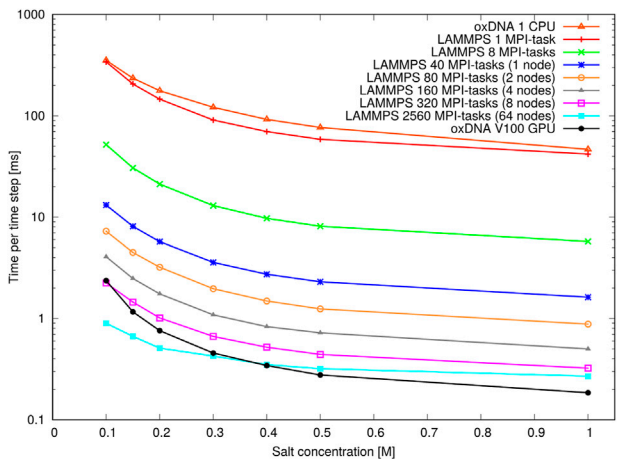


FIGURE 13 | Performance of the pointer benchmark as time per time step at various salt concentrations: Shown are results of the oxDNA standalone code on a single CPU and NVIDIA V100 GPU and of the LAMMPS implementation of the oxDNA2.0 model at different core counts.

variety of CPU architectures, is very suitable for studying small and intermediate system sizes, whereas the GPU-implementation has clearly the edge at large-scale simulations.

Figure 13 shows the performance with the pointer benchmark, again expressed as time per time step in milliseconds. This time the LAMMPS implementation is marginally faster than the oxDNA standalone code on a single CPU. Again, the GPU-runs of the standalone code features significantly shorter run times on all but the largest core counts and lowest salt concentrations. It appears the increase in run time between high and low salt concentration is slightly larger for the GPU-implementation of the standalone code. This could be due to a slightly better handling of neighbour lists in LAMMPS.

Most importantly, however, an increase in runtime by a factor 8–9 can be seen at all core counts when moving from high to moderate salt concentrations. This slowdown is in line with the increase in Debye length by about a factor 3 and reflects the longer cutoff radii and neighbour lists of the pair interactions. This large performance difference should be taken into account when choosing simulation parameters: For instance it is nearly always more convenient to perform relaxation runs to create well-initialized configurations at high salt concentrations (*e.g.* $[Na^+]$ = 1 M). Indeed, unless the response of the system to decreased salt concentration is of specific interest, we would generally recommend using high monovalent salt concentration such as $[Na^+]$ = 1 M for the actual data collection.

9 CONCLUSION

We have reviewed the properties of, and simulation methods available for, the oxDNA model. In the process we have created a well-documented library of examplar simulations available from (Sengar, 2021). Equally importantly, however, we have attempted to provide the necessary intuition both for successfully running oxDNA-based simulations, and also for identifying which systems would actually benefit from those simulations in the first place.

Having explored the model's strengths in some detail, it is worth noting a few natural directions for improvements. Although the model has well-parameterised sequence-dependent thermodynamics, and a good representation of average mechanical properties, it lacks sequence-dependent structure and mechanics. Incorporating this feature would be useful in and of itself, but would also be a useful first step towards building a model that could interface with other molecules such as proteins (Procyk et al., 2020).

AUTHOR CONTRIBUTIONS

AS: Ran simulations for **Figures 1–10** and partially wrote analysis and discussions for the corresponding sections. TO: Wrote introduction, analysis and discussion sections of the paper and ran simulations for **Figure 11**. OH: Ran pointer and oligomer benchmark simulations for the LAMPPS version of oxDNA for **Figures 12** and **13**. LR: Ran oligomer benchmark simulations for the standalone version of oxDNA for **Figure 12**. PŠ: Ran pointer benchmark simulations for the standalone version of oxDNA for **Figure 13**.

FUNDING

This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 851910). T.E.O. is supported by a Royal Society University Fellowship. OH acknowledges support from the EPSRC Early Career Research Software Engineer Fellowship Scheme (Grant No. EP/N019180/2). This work used the ARCHIE-WeSt High-Performance Computer (www.archie-west.ac.uk) based at the University of Strathclyde. PŠ acknowledges the use of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number TG-BIO210009.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.693710/full#supplementary-material

REFERENCES

Adleman, L. (1994). Molecular Computation of Solutions to Combinatorial Problems. *Science* 266, 1021–1024. doi:10.1126/science.7973651

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*. 4th ed. (New York: Garland Science).

Allen, R. J., Valeriani, C., and ten Wolde, P. R. (2009). Forward Flux Sampling for Rare Event Simulations. *J. Phys. Condens. Matter* 21, 463102. doi:10.1088/0953-8984/21/46/463102

Bae, J. H., Fang, J. Z., and Zhang, D. Y. (2020). High-throughput Methods for Measuring DNA Thermodynamics. *Nucleic Acids Res.* 48, e89. doi:10.1093/nar/gkaa521

Bai, X.-c., Martin, T. G., Scheres, S. H., and Dietz, H. (2012). Cryo-em Structure of a 3D DNA-Origami Object. *Proc. Natl. Acad. Sci.* 109, 20012–20017. doi:10.1073/pnas.1215713109

Benson, E., Mohammed, A., Rayneau-Kirkhope, D., Gådin, A., Orponen, P., and Högberg, B. (2018). Effects of Design Choices on the Stiffness of Wireframe DNA Origami Structures. *ACS nano* 12, 9291–9299. doi:10.1021/acsnano.8b04148

Berengut, J. F., Berengut, J. C., Doye, J. P., Prešern, D., Kawamoto, A., Ruan, J., et al. (2019). Design and Synthesis of Pleated DNA Origami Nanotubes with Adjustable Diameters. *Nucleic Acids Res.* 47, 11963–11975. doi:10.1093/nar/gkz1056

Berengut, J. F., Wong, C. K., Berengut, J. C., Doye, J. P., Ouldridge, T. E., and Lee, L. K. (2020). Self-limiting Polymerization of DNA Origami Subunits with Strain Accumulation. *ACS nano* 14, 17428–17441. doi:10.1021/acsnano.0c07696

Brady, R. A., Kaufhold, W. T., Brooks, N. J., Foderà, V., and Di Michele, L. (2019). Flexibility Defines Structure in Crystals of Amphiphilic DNA Nanostars. *J. Phys. Condensed Matter* 31, 074003. doi:10.1088/1361-648x/aaf4a1

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A Program for Macromolecular Energy,

Minimization, and Dynamics Calculations. *J. Comput. Chem.* 4, 187–217. doi:10.1002/jcc.540040211

Brown, R. F., Andrews, C. T., and Elcock, A. H. (2015). Stacking Free Energies of All DNA and RNA Nucleoside Pairs and Dinucleoside-Monophosphates Computed Using Recently Revised Amber Parameters and Compared with experiment. *J. Chem. Theor. Comput.* 11, 2315–2328. doi:10.1021/ct501170h

Brutzer, H., Luzzietti, N., Klaue, D., and Seidel, R. (2010). Energetics at the DNA Supercoiling Transition. *Biophysical J.* 98, 1267–1276. doi:10.1016/j.bpj.2009.12.4292

Bryant, Z., Stone, M. D., Gore, J., Smith, S. B., Cozzarelli, N. R., and Bustamante, C. (2003). Structural Transitions and Elasticity from Torque Measurements on DNA. *Nature* 424, 338–341. doi:10.1038/nature01810

Cabello-Garcia, J., Bae, W., Stan, G.-B. V., and Ouldridge, T. E. (2021). Handhold-mediated Strand Displacement: A Nucleic Acid Based Mechanism for Generating Far-From-Equilibrium Assemblies through Templated Reactions. *ACS nano* 15, 3272–3283. doi:10.1021/acsnano.0c10068

Cantor, C. R., and Schimmel, P. R. (1980). *Biophysical Chemistry: Part III: The Behavior of Biological Macromolecules*. San Francisco: Macmillan.

Caraglio, M., Skoruppa, E., and Carlon, E. (2019). Overtwisting Induces Polygonal Shapes in Bent DNA. *J. Chem. Phys.* 150, 135101. doi:10.1063/1.5084950

Castro, C. E., Kilchherr, F., Kim, D. N., Shiao, E. L., Wauer, T., Wortmann, P., et al. (2011). A Primer to Scaffolded DNA Origami. *Nat. Meth.* 8, 221–229. doi:10.1038/nmeth.1570

Chakraborty, D., Hori, N., and Thirumalai, D. (2018). Sequence-dependent Three Interaction Site Model for Single-And Double-Stranded DNA. *J. Chem. Theor. Comput.* 14, 3763–3779. doi:10.1021/acs.jctc.8b00091

Chen, H., Meisburger, S. P., Pabit, S. A., Sutton, J. L., Webb, W. W., and Pollack, L. (2012). Ionic Strength-dependent Persistence Lengths of Single-Stranded RNA and DNA. *Proc. Natl. Acad. Sci. United States America* 109, 799–804. doi:10.1073/pnas.1119057109

Cherry, K. M., and Qian, L. (2018). Scaling up Molecular Pattern Recognition with DNA-Based winner-take-all Neural Networks. *Nature* 559, 370–376. doi:10.1038/s41586-018-0289-6

Chhabra, H., Mishra, G., Cao, Y., Presern, D., Skoruppa, E., Tortora, M. M., et al. (2020). Computing the Elastic Mechanical Properties of Rodlike DNA Nanostructures. *J. Chem. Theor. Comput.* 16, 7748–7763. doi:10.1021/acs.jctc.0c00661

Choi, Y., Choi, H., Lee, A. C., and Kwon, S. (2018). Design and Synthesis of a Reconfigurable DNA Accordion Rack. *J. Vis. Exp.* 15, 58364. doi:10.3791/58364

Cocco, S., and Monasson, R. (1999). Statistical Mechanics of Torque Induced Denaturation of DNA. *Phys. Rev. Lett.* 83, 5178–5181. doi:10.1103/PhysRevLett.83.5178

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., et al. (1996). A Second Generation Force Field for the Simulation of Proteins. *Nucleic Acids Org. Mol. J. Am. Chem. Soc.* 117118, 5179–5197. doi:10.1021/ja955032e

Coronel, L., Suma, A., and Micheletti, C. (2018). Dynamics of Supercoiled DNA with Complex Knots: Large-Scale Rearrangements and Persistent Multi-Strand Interlocking. *Nucleic Acids Res.* 46, 7533–7541. doi:10.1093/nar/gky523

Craggs, T. D., Sustarsic, M., Plochowietz, A., Mosayebi, M., Kaju, H., Cuthbert, A., et al. (2019). Substrate Conformational Dynamics Facilitate Structure-specific Recognition of Gapped DNA by DNA Polymerase. *Nucleic Acids Res.* 47, 10788–10800. doi:10.1093/nar/gkz797

Crothers, D. M., Drak, J., Kahn, J. D., and Levene, S. D. (1992). DNA Bending, Flexibility, and Helical Repeat by Cyclization Kinetics. *Methods Enzymol.* 212, 3–29. doi:10.1016/0076-6879(92)12003-9

Dans, P. D., Walther, J., Gómez, H., and Orozco, M. (2016). Multiscale Simulation of DNA. *Curr. Opin. Struct. Biol.* 37, 29–45. doi:10.1016/j.sbi.2015.11.011

Dauxois, T., Peyrard, M., and Bishop, A. R. (1993). Dynamics and Thermodynamics of a Nonlinear Model for DNA Denaturation. *Phys. Rev. E* 47, 684–695. doi:10.1103/PhysRevE.47.684

Davidchack, R. L., Ouldridge, T. E., and Tretyakov, M. V. (2017). Geometric Integrator for Langevin Systems with Quaternion-Based Rotational Degrees of freedom and Hydrodynamic Interactions. *J. Chem. Phys.* 147, 224103. doi:10.1063/1.4999771

Davidchack, R. L., Ouldridge, T. E., and Tretyakov, M. V. (2015). New Langevin and Gradient Thermostats for Rigid Body Dynamics. *J. Chem. Phys.* 142 (14), 144114.

De Michele, C., Rovigatti, L., Bellini, T., and Sciortino, F. (2012). Self-assembly of Short DNA Duplexes: From a Coarse-Grained Model to Experiments through a Theoretical Link. *Soft Matter* 8, 8388–8398. doi:10.1039/c2sm25845e

Demurtas, D., Amzallag, A., Rawdon, E. J., Maddocks, J. H., Dubochet, J., and Stasiak, A. (2009). Bending Modes of DNA Directly Addressed by Cryo-Electron Microscopy of DNA Minicircles. *Nucleic Acids Res.* 37, 2882–2893. doi:10.1093/nar/gkp137

Desai, P. R., Brahmachari, S., Marko, J. F., Das, S., and Neuman, K. C. (2020). Coarse-grained Modelling of DNA Plectoneme Pinning in the Presence of Base-Pair Mismatches. *Nucleic Acids Res.* 48, 10713–10725. doi:10.1093/nar/gkaa836

Dessinges, M.-N., Maier, B., Zhang, Y., Peliti, M., Bensimon, D., and Croquette, V. (2002). Stretching Single Stranded DNA, a Model Polyelectrolyte. *Phys. Rev. Lett.* 89, 248102. doi:10.1103/physrevlett.89.248102

Douglas, S. M., Dietz, H., Liedl, T., Högberg, B., Graf, F., and Shih, W. M. (2009). Self-assembly of DNA into Nanoscale Three-Dimensional Shapes. *Nature* 459, 414–418. doi:10.1038/nature08016

Doye, J. P. K., Fowler, H., Prešern, D., Bohlin, J., Rovigatti, L., Romano, F., et al. (2020). *The oxDNA Coarse-Grained Model as a Tool to Simulate DNA Origami* (arXiv:2004.05052). Available at: https://www.arxiv.org/abs/2004.05052 (Accessed April 20, 2021).

Doye, J. P. K., Ouldridge, T. E., Louis, A. A., Romano, F., Šulc, P., Matek, C., et al. (2013). Coarse-graining DNA for Simulations of. *DNA nanotechnology* 15, 20395–20414. doi:10.1039/c3cp53545b

Du, Q., Kotlyar, A., and Vologodskii, A. (2008). Kinking the Double helix by Bending Deformation. *Nucleic Acids Res.* 36, 1120–1128. doi:10.1093/nar/gkm1125

Dunn, K. E., Dannenberg, F., Ouldridge, T. E., Kwiatkowska, M., Turberfield, A. J., and Bath, J. (2015). Guiding the Folding Pathway of DNA Origami. *Nature* 525, 82–86. doi:10.1038/nature14860

Engel, M. C., Romano, F., Louis, A. A., and Doye, J. P. (2020). Measuring Internal Forces in Single-Stranded DNA: Application to a DNA Force Clamp. *J. Chem. Theor. Comput.* 16, 7764–7775. doi:10.1021/acs.jctc.0c00286

Engel, M. C., Smith, D. M., Jobst, M. A., Sajfutdinow, M., Liedl, T., Romano, F., et al. (2018). Force-induced Unravelling of DNA Origami. *ACS nano* 12, 6734–6747. doi:10.1021/acsnano.8b01844

Fern, J., and Schulman, R. (2018). Modular DNA Strand-Displacement Controllers for Directing Material Expansion. *Nat. Commun.* 9, 1–8. doi:10.1038/s41467-018-06218-w

Fernandez-Castanon, J., Bomboi, F., Rovigatti, L., Zanatta, M., Paciaroni, A., Comez, L., et al. (2016). Small-angle Neutron Scattering and Molecular Dynamics Structural Study of Gelling DNA Nanostars. *J. Chem. Phys.* 145, 084910. doi:10.1063/1.4961398

Ferrenberg, A. M., and Swendsen, R. H. (1988). New Monte Carlo Technique for Studying Phase Transitions. *Phys. Rev. Lett.* 61, 2635–2638. doi:10.1103/physrevlett.61.2635

Fields, A. P., Meyer, E. A., and Cohen, A. E. (2013). Euler Buckling and Nonlinear Kinking of Double-Stranded DNA. *Nucleic Acids Res.* 41, 9881–9890. doi:10.1093/nar/gkt739

Flamm, C., Fontana, W., Hofacker, I. L., and Schuster, P. (2000). RNA Folding at Elementary Step Resolution. *RNA* 6, 325–338. doi:10.1017/s1355838200992161

Fonseca, P., Romano, F., Schreck, J. S., Ouldridge, T. E., Doye, J. P., and Louis, A. A. (2018). Multi-scale Coarse-Graining for the Study of Assembly Pathways in DNA-brick Self-Assembly. *J. Chem. Phys.* 148, 134910. doi:10.1063/1.5019344

Forth, S., Deufel, C., Sheinin, M. Y., Daniels, B., Sethna, J. P., and Wang, M. D. (2008). Abrupt Buckling Transition Observed during the Plectoneme Formation of Individual DNA Molecules. *Phys. Rev. Lett.* 100, 148301. doi:10.1103/physrevlett.100.148301

Fosado, Y. A. G., Landuzzi, F., and Sakaue, T. (2021). Twist Dynamics and Buckling Instability of Ring DNA: the Effect of Groove Asymmetry and Anisotropic Bending. *Soft Matter* 17, 1530–1537. doi:10.1039/d0sm01812k

Frenkel, D., and Smit, B. (2002). *Understanding Molecular Simulation: From Algorithms to Applications*. 2 edn. (Orlando, FL: Academic Press).

Fu, T. J., and Seeman, N. C. (1993). DNA Double-Crossover Molecules. *Biochemistry* 32, 3211. doi:10.1021/bi00064a003

Fujimoto, B. S., Brewood, G. P., and Schurr, J. M. (2006). Torsional Rigidities of Weakly Strained Dnas. *Biophysical J.* 91, 4166–4179. doi:10.1529/biophysj.106.087593

Fye, R. M., and Benham, C. J. (1999). Exact Method for Numerically Analyzing a Model of Local Denaturation in Superhelically Stressed DNA. *Phys. Rev. E* 59, 3408–3426. doi:10.1103/physreve.59.3408

Goodman, R. P., Sharp, I. A. T., Tardin, C. F., Erben, C. M., Berry, R. M., Schmidt, C. F., et al. (2005). Rapid Chiral Assembly of Rigid DNA Building Blocks for Molecular Nanofabrication. *Science* 310, 1661–1665. doi:10.1126/science.1120367

Gore, J., Bryant, Z., Nöllmann, M., Le, M. U., Cozzarelli, N. R., and Bustamante, C. (2006). DNA Overwinds when Stretched. *Nature* 442, 836–839. doi:10.1038/nature04974

Gross, P., Laurens, N., Oddershede, L. B., Bockelmann, U., Peterman, E. J., and Wuite, G. J. (2011). Quantifying How DNA Stretches, Melts and Changes Twist under Tension. *Nat. Phys.* 7, 731–736. doi:10.1038/nphys2002

Haley, N. E., Ouldridge, T. E., Ruiz, I. M., Geraldini, A., Louis, A. A., Bath, J., et al. (2020). Design of Hidden Thermodynamic Driving for Non-equilibrium Systems via Mismatch Elimination during DNA Strand Displacement. *Nat. Commun.* 11, 1–11. doi:10.1038/s41467-020-16353-y

Harrison, R. M., Romano, F., Ouldridge, T. E., Louis, A. A., and Doye, J. P. (2015). *Coarse-grained Modelling of strong DNA Bending I: Thermodynamics and Comparison to an Experimental" Molecular Vice"* (arXiv preprint arXiv:1506.09005). Available at: https://arxiv.org/abs/1506.09005 (Accessed April 20, 2021).

Harrison, R. M., Romano, F., Ouldridge, T. E., Louis, A. A., and Doye, J. P. K. (2019). Identifying Physical Causes of Apparent Enhanced Cyclization of Short DNA Molecules with a Coarse-Grained Model. *J. Chem. Theor. Comput.* 15, 4660–4672. doi:10.1021/acs.jctc.9b00112

Hatch, K., Danilowicz, C., Coljee, V., and Prentiss, M. (2008). Demonstration that the Shear Force Required to Separate Short Double-Stranded DNA Does Not Increase Significantly with Sequence Length for Sequences Longer Than 25 Base Pairs. *Phys. Rev. E* 78, 011920. doi:10.1103/PhysRevE.78.011920

Henning-Knechtel, A., Knechtel, J., and Magzoub, M. (2017). NAR Breakthrough Article DNA-Assisted Oligomerization of Pore-Forming Toxin Monomers into Precisely-Controlled Protein Channels. *Nucleic Acids Res.* 45, 12057–12068. doi:10.1093/nar/gkx990

Henrich, O., Gutiérrez Fosado, Y. A., Curk, T., and Ouldridge, T. E. (2018). Coarse-grained Simulation of DNA Using LAMMPS: An Implementation of the oxDNA Model and its Applications. *Eur. Phys. J. E* 41, 1–16. doi:10.1140/epje/i2018-11669-8

Herrero-Galán, E., Fuentes-Perez, M. E., Carrasco, C., Valpuesta, J. M., Carrascosa, J. L., Moreno-Herrero, F., et al. (2013). Mechanical Identities of RNA and DNA Double Helices Unveiled at the Single-Molecule Level. *J. Am. Chem. Soc.* 135, 122–131. doi:10.1021/ja3054755

Hinckley, D. M., Freeman, G. S., Whitmer, J. K., and De Pablo, J. J. (2013). An Experimentally-Informed Coarse-Grained 3-Site-Per-Nucleotide Model of Dna: Structure, Thermodynamics, and Dynamics of Hybridization. *J. Chem. Phys.* 139, 144903. doi:10.1063/1.4822042

Hobza, P., and Šponer, J. (1999). Structure, Energetics, and Dynamics of the Nucleic Acid Base Pairs: Nonempirical Ab Initio Calculations. *Chem. Rev.* 99, 3247–3276. doi:10.1021/cr9800255

Hoffecker, I. T., Chen, S., Gådin, A., Bosco, A., Teixeira, A. I., and Högberg, B. (2019). Solution-controlled Conformational Switching of an Anchored Wireframe DNA Nanostructure. *Small* 15, 1803628. doi:10.1002/smll.201803628

Hong, F., Jiang, S., Lan, X., Narayanan, R. P., Šulc, P., Zhang, F., et al. (2018). Layered-Crossover Tiles with Precisely Tunable Angles for 2D and 3D DNA Crystal Engineering. *J. Am. Chem. Soc.* 140, 14670–14676. doi:10.1021/jacs.8b07180

Huguet, J. M., Bizarro, C. V., Forns, N., Smith, S. B., Bustamante, C., and Ritort, F. (2010). Single-molecule Derivation of Salt Dependent Base-Pair Free Energies in DNA. *Proc. Natl. Acad. Sci. U.S.A.* 107. doi:10.1073/pnas.1001454107

Irmisch, P., Ouldridge, T. E., and Seidel, R. (2020). Modeling DNA-Strand Displacement Reactions in the Presence of Base-Pair Mismatches. *J. Am. Chem. Soc.* 142, 11451–11463. doi:10.1021/jacs.0c03105

Ivani, I., Dans, P. D., Noy, A., Pérez, A., Faustino, I., Hospital, A., et al. (2016). Parmbsc1: a Refined Force Field for DNA Simulations. *Nat. Methods* 13, 55. doi:10.1038/nmeth.3658

Ke, Y., Ong, L. L., Shih, W. M., and Yin, P. (2012). Three-dimensional Structures Self-Assembled from DNA Bricks. *Science* 338, 1177–1183. doi:10.1126/science.1227268

Kim, C., Lee, O.-c., Kim, J.-Y., Sung, W., and Lee, N. K. (2015). Dynamic Release of Bending Stress in Short Dsdna by Formation of a Kink and forks. *Angew. Chem.* 127, 9071–9075. doi:10.1002/ange.201502055

Kočar, V., Schreck, J. S., Čeru, S., Gradišar, H., Bašić, N., Pisanski, T., et al. (2016). Design Principles for Rapid Folding of Knotted DNA Nanostructures. *Nat. Commun.* 7, 1–8. doi:10.1038/ncomms10803

Korolev, N., Luo, D., Lyubartsev, A. P., and Nordenskiöld, L. (2014). A Coarse-Grained DNA Model Parameterized from Atomistic Simulations by Inverse Monte Carlo. *Polymers* 6, 1655–1675. doi:10.3390/polym6061655

Kratky, O., and Porod, G. (1949). Röntgenuntersuchung Gelöster Fadenmoleküle. *Recueil des Travaux Chimiques des Pays-Bas* 68, 1106–1122. doi:10.1002/recl.19490681203

Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. (1992). The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* 13, 1011–1021. doi:10.1002/jcc.540130812

LAMMPS Documentation (2021). USER-CGDNA package. Available at: https://docs.lammps.org/ (Accessed May 27, 2021).

Le, T. T., and Kim, H. D. (2014). Probing the Elastic Limit of DNA Bending. *Nucleic Acids Res.* 42, 10786–10794. doi:10.1093/nar/gku735

Lee, J. Y., Terakawa, T., Qi, Z., Steinfeld, J. B., Redding, S., Kwon, Y. H., et al. (2015). Base Triplet Stepping by the Rad51/RecA Family of Recombinases. *Science* 349, 977–981. doi:10.1126/science.aab2666

Lewis, M., Chang, G., Horton, N. C., Kercher, M. A., Pace, H. C., Schumacher, M. A., et al. (1996). Crystal Structure of the Lactose Operon Repressor and its Complexes with DNA and Inducer. *Science* 271, 1247–1254. doi:10.1126/science.271.5253.1247

Lionnet, T., Joubaud, S., Lavery, R., Bensimon, D., and Croquette, V. (2006). Wringing Out DNA. *Phys. Rev. Lett.* 96, 178102. doi:10.1103/physrevlett.96.178102

Machado, M. R., and Pantano, S. (2015). Exploring Laci–DNA Dynamics by Multiscale Simulations Using the SIRAH Force Field. *J. Chem. Theor. Comput.* 11, 5012–5023. doi:10.1021/acs.jctc.5b00575

Machinek, R. R., Ouldridge, T. E., Haley, N. E., Bath, J., and Turberfield, A. J. (2014). Programmable Energy Landscapes for Kinetic Control of DNA Strand Displacement. *Nat. Commun.* 5, 1–9. doi:10.1038/ncomms6324

Maciejczyk, M., Spasic, A., Liwo, A., and Scheraga, H. A. (2014). DNA Duplex Formation with a Coarse-Grained Model. *J. Chem. Theor. Comput.* 10, 5020–5035. doi:10.1021/ct4006689

Maffeo, C., and Aksimentiev, A. (2020). MrDNA: a Multi-Resolution Model for Predicting the Structure and Dynamics of DNA Systems. *Nucleic Acids Res.* 48, 5135–5146. doi:10.1093/nar/gkaa200

Maffeo, C., Ngo, T. T., Ha, T., and Aksimentiev, A. (2014). A Coarse-Grained Model of Unstructured Single-Stranded DNA Derived from Atomistic Simulation and Single-Molecule experiment. *J. Chem. Theor. Comput.* 10, 2891–2896. doi:10.1021/ct500193u

Matek, C., Ouldridge, T. E., Doye, J. P. K., and Louis, A. A. (2015). Plectoneme Tip Bubbles: Coupled Denaturation and Writhing in Supercoiled DNA. *Sci. Rep.* 5, 7655. doi:10.1038/srep07655

Matek, C., Ouldridge, T. E., Levy, A., Doye, J. P. K., and Louis, A. A. (2012). DNA Cruciform Arms Nucleate through a Correlated but Asynchronous Cooperative Mechanism. *J. Phys. Chem. B* 116, 11616–11625. doi:10.1021/jp3080755

Meng, H., Bosman, J., van der Heijden, T., and van Noort, J. (2014). Coexistence of Twisted, Plectonemic, and Melted DNA in Small Topological Domains. *Biophysical J.* 106, 1174–1181. doi:10.1016/j.bpj.2014.01.017

Metropolis, N., and Ulam, S. (1949). The Monte Carlo Method. *J. Am. Stat. Assoc.* 44, 335–341. doi:10.1080/01621459.1949.10483310

Mills, J. B., Vacano, E., and Hagerman, P. J. (1999). Flexibility of Single-Stranded DNA: Use of Gapped Duplex Helices to Determine the Persistence Lengths of Poly(dT) and Poly(dA). *J. Mol. Biol.* 285, 245–257. doi:10.1006/jmbi.1998.2287

Mosayebi, M., Louis, A. A., Doye, J. P., and Ouldridge, T. E. (2015). Force-induced Rupture of a DNA Duplex: from Fundamentals to Force Sensors. *ACS nano* 9, 11993–12003. doi:10.1021/acsnano.5b04726

Mosconi, F., Allemand, J. F., Bensimon, D., and Croquette, V. (2009). Measurement of the Torque on a Single Stretched and Twisted DNA Using

Magnetic Tweezers. *Phys. Rev. Lett.* 102, 078301. doi:10.1103/PhysRevLett.102.078301

Murphy, M. C., Rasnik, I., Chang, W., Lohman, T. M., and Ha, T. (2004). Probing Single-Stranded DNA Conformational Flexibility Using Flourescence Spectroscopy. *Biophys. J.* 86, 2530–2537. doi:10.1016/s0006-3495(04)74308-8

Murtola, T., Bunker, A., Vattulainen, I., Deserno, M., and Karttunen, M. (2009). Multiscale Modeling of Emergent Materials: Biological and Soft Matter. *Phys. Chem. Chem. Phys.* 11, 1869–1892. doi:10.1039/b818051b

Muscat, R. A., Bath, J., and Turberfield, A. J. (2011). A Programmable Molecular Robot. *Nano Lett.* 11, 982–987. doi:10.1021/nl1037165

Nguyen, H., Maier, J., Huang, H., Perrone, V., and Simmerling, C. (2014). Folding Simulations for Proteins with Diverse Topologies Are Accessible in Days with a Physics-Based Force Field and Implicit Solvent. *J. Am. Chem. Soc.* 136, 13959–13962. doi:10.1021/ja5032776

Nikolov, D. B., Chen, H., Halay, E. D., Hoffman, A., Roeder, R. G., and Burley, S. K. (1996). Crystal Structure of a Human Tata Box-Binding Protein/tata Element Complex. *Proc. Natl. Acad. Sci.* 93, 4862–4867. doi:10.1073/pnas.93.10.4862

Nisoli, C., and Bishop, A. R. (2011). Thermomechanics of DNA: Theory of thermal Stability under Load. *Phys. Rev. Lett.* 107, 068102. doi:10.1103/PhysRevLett.107.068102

Nomidis, S. K., Skoruppa, E., Carlon, E., and Marko, J. F. (2019). Twist-bend Coupling and the Statistical Mechanics of the Twistable Wormlike-Chain Model of DNA: Perturbation Theory and beyond. *Phys. Rev. E* 99, 032414. doi:10.1103/PhysRevE.99.032414

Odijk, T. (1995). Stiff Chains and Filaments under Tension. *Macromolecules* 28, 7016–7018. doi:10.1021/ma00124a044

Ouldridge, T. E., Hoare, R. L., Louis, A. A., Doye, J. P. K., Bath, J., and Turberfield, A. J. (2013a). Optimizing DNA Nanotechnology through Coarse-Grained Modeling: A Two-Footed DNA walker. *ACS Nano* 7, 2479–2490. doi:10.1021/nn3058483

Ouldridge, T. E. (2012). Inferring Bulk Self-Assembly Properties from Simulations of Small Systems with Multiple Constituent Species and Small Systems in the Grand Canonical Ensemble. *J. Chem. Phys.* 137, 144105. doi:10.1063/1.4757267

Ouldridge, T. E., Johnston, I. G., Louis, A. A., and Doye, J. P. K. (2009). The Self-Assembly of DNA Holliday Junctions Studied with a Minimal Model. *J. Chem. Phys.* 130, 065101. doi:10.1063/1.3055595

Ouldridge, T. E., Louis, A. A., and Doye, J. P. K. (2010a). DNA Nanotweezers Studied with a Coarse-Grained Model of DNA. *Phys. Rev. Lett.* 104, 178101. doi:10.1103/PhysRevLett.104.178101

Ouldridge, T. E., Louis, A. A., and Doye, J. P. K. (2010b). Extracting Bulk Properties of Self-Assembling Systems from Small Simulations. *J. Phys. Condensed Matter* 22, 104102. doi:10.1088/0953-8984/22/10/104102

Ouldridge, T. E., Louis, A. A., and Doye, J. P. K. (2011). Structural, Mechanical, and Thermodynamic Properties of a Coarse-Grained DNA Model. *J. Chem. Phys.* 134, 085101. doi:10.1063/1.3552946

Ouldridge, T. E., Šulc, P., Romano, F., Doye, J. P. K., and Louis, A. A. (2013b). DNA Hybridization Kinetics: Zippering, Internal Displacement and Sequence Dependence. *Nucleic Acids Res.* 41, 8886–8895. doi:10.1093/nar/gkt687

oxDNA wiki (2015). Available at: https://dna.physics.ox.ac.uk/ (Accessed April 20, 2021).

Park, G., Cho, M. K., and Jung, Y. (2021). Sequence-dependent Kink Formation in Short DNA Loops: Theory and Molecular Dynamics Simulations. *J. Chem. Theor. Comput.* 17, 1308–1317. doi:10.1021/acs.jctc.0c01116

Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, T. E., III, Laughton, C. A., et al. (2007). Refinement of the Amber Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophysical J.* 92, 3817–3829. doi:10.1529/biophysj.106.097782

Pérez, A., Noy, A., Lankas, F., Luque, F. J., and Orozco, M. (2004). The Relative Flexibility of B-DNA and A-RNA Duplexes: Database Analysis. *Nucleic Acids Res.* 32, 6144–6151. doi:10.1093/nar/gkh954

Podtelezhnikov, A. A., Mao, C., Seeman, N. C., and Vologodskii, A. (2000). Multimerization-cyclization of DNA Fragments as a Method of Conformational Analysis. *Biophysical J.* 79, 2692–2704. doi:10.1016/s0006-3495(00)76507-6

Poland, D., and Scheraga, H. A. (1966). Occurrence of a Phase Transition in Nucleic Acid Models. *J. Chem. Phys.* 45, 1464–1469. doi:10.1063/1.1727786

Poppleton, E., Bohlin, J., Matthies, M., Sharma, S., Zhang, F., and Šulc, P. (2020). Design, Optimization and Analysis of Large DNA and RNA Nanostructures through Interactive Visualization, Editing and Molecular Simulation. *Nucleic Acids Res.* 48, E72. doi:10.1093/nar/gkaa417

Poppleton, E., Romero, R., Mallya, A., Rovigatti, L., and Šulc, P. (2021). Oxdna.Org: a Public Webserver for Coarse-Grained Simulations of DNA and RNA Nanostructures. *Nucleic Acids Res.*

Procyk, J., Poppleton, E., and Šulc, P. (2020). Coarse-grained Nucleic Acid-Protein Model for Hybrid Nanotechnology. *Soft Matter* 17, 3586. doi:10.1039/d0sm01639j

Qian, L., Winfree, E., and Bruck, J. (2011). Neural Network Computation with DNA Strand Displacement Cascades. *Nature* 475, 368–372. doi:10.1038/nature10262

Qu, H., and Zocchi, G. (2011). The Complete Bending Energy Function for Nicked DNA. *EPL (Europhysics Letters)* 94, 18003. doi:10.1209/0295-5075/94/18003

Richmond, T. J., and Davey, C. A. (2003). The Structure of DNA in the Nucleosome Core. *Nature* 423, 145–150. doi:10.1038/nature01595

Rivetti, C., Walker, C., and Bustamante, C. (1998). Polymer Chain Statistics and Conformational Analysis of DNA Molecules with Bends or Sections of Different Flexibility. *J. Mol. Biol.* 280, 41–59. doi:10.1006/jmbi.1998.1830

Rocklin, G. J., Chidyausiku, T. M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., et al. (2017). Global Analysis of Protein Folding Using Massively Parallel Design, Synthesis, and Testing. *Science* 357, 168–175. doi:10.1126/science.aan0693

Romano, F., Chakraborty, D., Doye, J. P. K., Ouldridge, T. E., and Louis, A. A. (2013). Coarse-grained Simulations of DNA Overstretching. *J. Chem. Phys.* 138, 085101. doi:10.1063/1.4792252

Romano, F., Hudson, A., Doye, J. P. K., Ouldridge, T. E., and Louis, A. A. (2012). The Effect of Topology on the Structure and Free Energy Landscape of DNA Kissing Complexes. *J. Chem. Phys.* 136, 215102. doi:10.1063/1.4722203

Rothemund, P. W. (2006). Folding DNA to Create Nanoscale Shapes and Patterns. *Nature* 440, 297–302. doi:10.1038/nature04586

Rothemund, P. W., Papadakis, N., and Winfree, E. (2004). Algorithmic Self-Assembly of DNA Sierpinski Triangles. *Plos Biol.* 2, e424. doi:10.1371/journal.pbio.0020424

Rovigatti, L., Bomboi, F., and Sciortino, F. (2014). Accurate Phase Diagram of Tetravalent DNA Nanostars. *J. Chem. Phys.* 140, 154903. doi:10.1063/1.4870467

Rovigatti, L., Sulc, P., Reguly, I. Z., and Romano, F. (2015). A Comparison between Parallelization Approaches in Molecular Dynamics Simulations on GPUs. *J. Comput. Chem.* 36, 1–8. doi:10.1002/jcc.23763

Russo, J., Tartaglia, P., and Sciortino, F. (2009). Reversible Gels of Patchy Particles: Role of the Valence. *J. Chem. Phys.* 131, 014504. doi:10.1063/1.3153843

Salerno, D., Tempestini, A., Mai, I., Brogioli, D., Ziano, R., Cassina, V., et al. (2012). Single-molecule Study of the DNA Denaturation Phase Transition in the Force-Torsion Space. *Phys. Rev. Lett.* 109, 118303. doi:10.1103/physrevlett.109.118303

SantaLucia, J. (1998). A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics. *Proc. Natl. Acad. Sci. United States America* 95, 1460–1465. doi:10.1073/pnas.95.4.1460

SantaLucia, J., and Hicks, D. (2004). The Thermodynamics of DNA Structural Motifs. *Annu. Rev. Biophys. Biomol. Struct.* 33, 415–440. doi:10.1146/annurev.biophys.32.110601.141800

Savelyev, A. (2012). Do monovalent mobile Ions Affect Dna's Flexibility at High Salt Content?. *Phys. Chem. Chem. Phys.* 14, 2250–2254. doi:10.1039/c2cp23499h

Savelyev, A., and Papoian, G. A. (2009). Molecular Renormalization Group Coarse-Graining of Polymer Chains: Application to Double-Stranded DNA. *Biophysical J.* 96, 4044–4052. doi:10.1016/j.bpj.2009.02.067

Schaeffer, J. M., Thachuk, C., and Winfree, E. (2015). "Stochastic Simulation of the Kinetics of Multiple Interacting Nucleic Acid Strands," in *International Workshop on DNA-Based Computers* (Springer), 194–211. doi:10.1007/978-3-319-21999-8_13

Schreck, J. S., Romano, F., Zimmer, M. H., Louis, A. A., and Doye, J. P. (2016). Characterizing DNA star-tile-based Nanostructures Using a Coarse-Grained Model. *ACS nano* 10, 4236–4247. doi:10.1021/acsnano.5b07664

Seeman, N. C. (2003). *DNA a Mater. World* 421, 427–431. doi:10.1038/nature01406

Sengar, A. (2021). oxDNA Supporting Data. doi:10.5281/zenodo.4809769 (Accessed April 20, 2021).

Seol, Y., Skinner, G. M., Visscher, K., Buhot, A., and Halperin, A. (2007). Stretching of Homopolymeric RNA Reveals Single-Stranded Helices and Base-Stacking. *Phys. Rev. Lett.* 98, 158103. doi:10.1103/PhysRevLett.98.158103

Seol, Y., Skinner, G. M., and Visscher, K. (2004). Elastic Properties of a Single-Stranded Charged Homopolymeric Ribonucleotide. *Phys. Rev. Lett.* 93. doi:10.1103/PhysRevLett.93.118102

Sharma, R., Schreck, J. S., Romano, F., Louis, A. A., and Doye, J. P. (2017). Characterizing the Motion of Jointed DNA Nanostructures Using a Coarse-Grained Model. *ACS nano* 11, 12426–12435. doi:10.1021/acsnano.7b06470

Shi, Z., Castro, C. E., and Arya, G. (2017). Conformational Dynamics of Mechanically Compliant DNA Nanostructures from Coarse-Grained Molecular Dynamics Simulations. *ACS nano* 11, 4617–4630. doi:10.1021/acsnano.7b00242

Shin, J.-S., and Pierce, N. A. (2004). A Synthetic DNA walker for Molecular Transport. *J. Am. Chem. Soc.* 126, 10834–10835. doi:10.1021/ja047543j

Skoruppa, E., Laleman, M., Nomidis, S. K., and Carlon, E. (2017). DNA Elasticity from Coarse-Grained Simulations: The Effect of Groove Asymmetry. *J. Chem. Phys.* 146, 214902. doi:10.1063/1.4984039

Smith, S. B., Cui, Y., and Bustamante, C. (1996). Overstretching B-DNA: The Elastic Response of Individual Double-Stranded and Single-Stranded DNA Molecules. *Science* 271, 795–799. doi:10.1126/science.271.5250.795

Snodin, B. E., Randisi, F., Mosayebi, M., Šulc, P., Schreck, J. S., Romano, F., et al. (2015). Introducing Improved Structural Properties and Salt Dependence into a Coarse-Grained Model of DNA. *J. Chem. Phys.* 142, 234901. doi:10.1063/1.4921957

Snodin, B. E., Romano, F., Rovigatti, L., Ouldridge, T. E., Louis, A. A., and Doye, J. P. K. (2016). Direct Simulation of the Self-Assembly of a Small DNA Origami. *ACS Nano* 10, 1724–1737. doi:10.1021/acsnano.5b05865

Snodin, B. E., Schreck, J. S., Romano, F., Louis, A. A., and Doye, J. P. K. (2019). Coarse-grained Modelling of the Structural Properties of DNA Origami. *Nucleic Acids Res.* 47, 1585–1597. doi:10.1093/nar/gky1304

Šponer, J., Jurečka, P., and Hobza, P. (2004). Accurate Interaction Energies of Hydrogen-Bonded Nucleic Acid Base Pairs. *J. Am. Chem. Soc.* 126, 10142–10151. doi:10.1021/ja048436s

Šponer, J., Riley, K. E., and Hobza, P. (2008). Nature and Magnitude of Aromatic Stacking of Nucleic Acid Bases. *Phys. Chem. Chem. Phys.* 10, 2595–2610. doi:10.1039/b719370j

Srinivas, N., Ouldridge, T. E., Šulc, P., Schaeffer, J., Yurke, B., Louis, A. A., et al. (2013). On the Biophysics and Kinetics of Toehold-Mediated DNA Strand Displacement. *Nucl. Acids Res.* 41, 10641–10658. doi:10.1093/nar/gkt801

Srinivas, N., Parkin, J., Seelig, G., Winfree, E., and Soloveichik, D. (2017). Enzyme-free Nucleic Acid Dynamical Systems. *Science* 358, eaal2052. doi:10.1126/science.aal2052

Stoev, I. D., Cao, T., Caciagli, A., Yu, J., Ness, C., Liu, R., et al. (2020). On the Role of Flexibility in Linker-Mediated DNA Hydrogels. *Soft Matter* 16, 990–1001. doi:10.1039/c9sm01398a

Strick, T. R., Allemand, J. F., and Bensimon, D. (1996). The Elasticity of a Single Supercoiled DNA Molecule (Vol 271, Pg 1835, 1996). *Science* 272, 797. doi:10.1126/science.271.5257.1835

Šulc, P., Romano, F., Ouldridge, T. E., Doye, J. P. K., and Louis, A. A. (2014). A Nucleotide-Level Coarse-Grained Model of RNA. *J. Chem. Phys.* 140, 235102. doi:10.1063/1.4881424

Šulc, P., Romano, F., Ouldridge, T. E., Rovigatti, L., Doye, J. P. K., and Louis, A. A. (2012). Sequence-dependent Thermodynamics of a Coarse-Grained DNA Model. *J. Chem. Phys.* 137, 135101. doi:10.1063/1.4754132

Suma, A., Poppleton, E., Matthies, M., Šulc, P., Romano, F., Louis, A. A., et al. (2019). TacoxDNA: A User-Friendly Web Server for Simulations of Complex DNA Structures, from Single Strands to Origami. *J. Comput. Chem.* 40, 2586–2595. doi:10.1002/jcc.26029

Sutthibutpong, T., Matek, C., Benham, C., Slade, G. G., Noy, A., Laughton, C., et al. (2016). Long-range Correlations in the Mechanics of Small DNA Circles under Topological Stress Revealed by Multi-Scale Simulation. *Nucleic Acids Res.* 44, 9121–9130. doi:10.1093/nar/gkw815

Tee, S. R., and Wang, Z. (2018). How Well Can DNA Rupture DNA? Shearing and Unzipping Forces inside DNA Nanostructures. *ACS omega* 3, 292–301. doi:10.1021/acsomega.7b01692

Tempestini, A., Cassina, V., Brogioli, D., Ziano, R., Erba, S., Giovannoni, R., et al. (2013). Magnetic Tweezers Measurements of the Nanomechanical Stability of DNA against Denaturation at Various Conditions of Ph and Ionic Strength. *Nucleic Acids Res.* 41, 2009–2019. doi:10.1093/nar/gks1206

Tikhomirov, G., Petersen, P., and Qian, L. (2017). Fractal Assembly of Micrometre-Scale DNA Origami Arrays with Arbitrary Patterns. *Nature* 552, 67–71. doi:10.1038/nature24655

Tomov, T. E., Tsukanov, R., Glick, Y., Berger, Y., Liber, M., Avrahami, D., et al. (2017). DNA Bipedal Motor Achieves a Large Number of Steps Due to Operation Using Microfluidics-Based Interface. *Acs Nano* 11, 4002–4008. doi:10.1021/acsnano.7b00547

Torrie, G. M., and Valleau, J. P. (1977). Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comp. Phys.* 23, 187–199. doi:10.1016/0021-9991(77)90121-8

Tortora, M. M., Mishra, G., Prešern, D., and Doye, J. P. (2020). Chiral Shape Fluctuations and the Origin of Chirality in Cholesteric Phases of DNA Origamis. *Sci. Adv.* 6, eaaw8331. doi:10.1126/sciadv.aaw8331

Uusitalo, J. J., Ingolfsson, H. I., Akhshi, P., Tieleman, D. P., and Marrink, S. J. (2015). Martini Coarse-Grained Force Field: Extension to DNA. *J. Chem. Theor. Comput.* 11, 3932–3945. doi:10.1021/acs.jctc.5b00286

Vafabakhsh, R., and Ha, T. (2012). Extreme Bendability of DNA Less Than 100 Base Pairs Long Revealed by Single-Molecule Cyclization. *Science* 337, 1097–1101. doi:10.1126/science.1224139

van Mameren, J., Gross, P., Farge, G., Hooijman, P., Modesti, M., Falkenberg, M., et al. (2009). Unraveling the Structure of DNA during Overstretching by Using Multicolor, Single-Molecule Fluorescence Imaging. *Proc. Natl. Acad. Sci.* 106, 18231–18236. doi:10.1073/pnas.0904322106

Wagenbauer, K. F., Sigl, C., and Dietz, H. (2017). Gigadalton-scale Shape-Programmable DNA Assemblies. *Nature* 552, 78–83. doi:10.1038/nature24651

Wang, M. D., Yin, H., Landick, R., Gelles, J., and Block, S. M. (1997). Stretching DNA with Optical Tweezers. *Biophysical J.* 72, 1335–1346. doi:10.1016/s0006-3495(97)78780-0

Wang, Q., Irobalieva, R. N., Chiu, W., Schmid, M. F., Fogg, J. M., Zechiedrich, L., et al. (2017). Influence of DNA Sequence on the Structure of Minicircles under Torsional Stress. *Nucleic Acids Res.* 45, 7633–7642. doi:10.1093/nar/gkx516

Wang, Q., Myers, C. G., and Pettitt, B. M. (2015). Twist-induced Defects of the P-SSP7 Genome Revealed by Modeling the Cryo-EM Density. *J. Phys. Chem. B* 119, 4937–4943. doi:10.1021/acs.jpcb.5b00865

Wang, Q., and Pettitt, B. M. (2014). Modeling DNA Thermodynamics under Torsional Stress. *Biophysical J.* 106, 1182–1193. doi:10.1016/j.bpj.2014.01.022

Wang, Q., and Pettitt, B. M. (2016). Sequence Affects the Cyclization of DNA Minicircles. *J. Phys. Chem. Lett.* 7, 1042–1046. doi:10.1021/acs.jpclett.6b00246

Wang, X., and Ha, T. (2013). Defining Single Molecular Forces Required to Activate Integrin and Notch Signaling. *Science* 340, 991–994. doi:10.1126/science.1231041

Watson, J. D., and Crick, F. H. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* 171, 737–738. doi:10.1038/171737a0

Whitelam, S., Feng, E. H., Hagan, M. F., and Geissler, P. L. (2009). The Role of Collective Motion in Examples of Coarsening and Self-Assembly. *Soft Matter* 5, 1251–1262. doi:10.1039/b810031d

Whitelam, S., and Geissler, P. L. (2007). Avoiding Unphysical Kinetic Traps in Monte Carlo Simulations of Strongly Attractive Particles. *J. Chem. Phys.* 127, 154101. doi:10.1063/1.2790421

Wickham, S. F. J., Bath, J., Katsuda, Y., Endo, M., Hidaka, K., Sugiyama, H., et al. (2012). A DNA-Based Molecular Motor that Can Navigate a Network of Tracks. *Nat. Nanotechnol.* 7, 169–173. doi:10.1038/nnano.2011.253

Widom, J. (2001). Role of DNA Sequence in Nucleosome Stability and Dynamics. *Q. Rev. Biophys.* 34, 269. doi:10.1017/s0033583501003699

Woods, D., Doty, D., Myhrvold, C., Hui, J., Zhou, F., Yin, P., et al. (2019). Diverse and Robust Molecular Algorithms Using Reprogrammable DNA Self-Assembly. *Nature* 567, 366–372. doi:10.1038/s41586-019-1014-9

Xayaphoummine, A., Bucher, T., and Isambert, H. (2005). Kinefold Web Server for RNA/DNA Folding Path and Structure Prediction Including Pseudoknots and Knots. *Nucleic Acids Res.* 33, W605–W610. doi:10.1093/nar/gki447

Yamakawa, H. (1977). A Hypothesis on Polymer Chain Configurations. Helical Wormlike Chains. *Macromolecules* 10, 692–696. doi:10.1021/ma60057a039

Yao, G., Zhang, F., Wang, F., Peng, T., Liu, H., Poppleton, E., et al. (2020). Meta-dna Structures. *Nat. Chem.* 12, 1067–1075. doi:10.1038/s41557-020-0539-8

Yoo, J., and Aksimentiev, A. (2012). Improved Parametrization of Li+, Na+, K+, and Mg2+ Ions for All-Atom Molecular Dynamics Simulations of Nucleic Acid Systems. *J. Phys. Chem. Lett.* 3, 45–50. doi:10.1021/jz201501a

Yurke, B., Turberfield, A. J., Mills, A. P., Simmel, F. C., and Neumann, J. L. (2000). A DNA-Fuelled Molecular Machine Made of DNA. *Nature* 406, 605–608. doi:10.1038/35020524

Zhang, D. Y., and Seelig, G. (2011). Dynamic DNA Nanotechnology Using Strand-Displacement Reactions. *Nat. Chem.* 3, 103–113. doi:10.1038/nchem.957

Zhang, F., Jiang, S., Wu, S., Li, Y., Mao, C., Liu, Y., et al. (2015). Complex Wireframe DNA Origami Nanostructures with Multi-Arm junction Vertices. *Nat. Nanotechnol.* 10, 779. doi:10.1038/nnano.2015.162

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 © 2021 Sengar, Ouldridge, Henrich, Rovigatti and Šulc. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 23 June 2021
doi: 10.3389/fmolb.2021.669897

Elucidating Axonal Injuries Through Molecular Modelling of Myelin Sheaths and Nodes of Ranvier

Marzieh Saeedimasine[1†], *Annaclaudia Montanino*[2], *Svein Kleiven*[2] *and Alessandra Villa*[3*]

[1]Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden, [2]Division of Neuronic Engineering, KTH-Royal Institute of Technology, Stockholm, Sweden, [3]PDC-Center for High Performance Computing, KTH-Royal Institute of Technology, Stockholm, Sweden

Around half of the traumatic brain injuries are thought to be axonal damage. Disruption of the cellular membranes, or alternatively cytoskeletal damage has been suggested as possible injury trigger. Here, we have used molecular models to have a better insight on the structural and mechanical properties of axon sub-cellular components. We modelled myelin sheath and node of Ranvier as lipid bilayers at a coarse grained level. We built ex-novo a model for the myelin. Lipid composition and lipid saturation were based on the available experimental data. The model contains 17 different types of lipids, distributed asymmetrically between two leaflets. Molecular dynamics simulations were performed to characterize the myelin and node-of-Ranvier bilayers at equilibrium and under deformation and compared to previous axolemma simulations. We found that the myelin bilayer has a slightly higher area compressibility modulus and higher rupture strain than node of Ranvier. Compared to the axolemma in unmyelinated axon, mechanoporation occurs at 50% higher strain in the myelin and at 23% lower strain in the node of Ranvier in myelinated axon. Combining the results with finite element simulations of the axon, we hypothesizes that myelin does not rupture at the thresholds proposed in the literature for axonal injury while rupture may occur at the node of Ranvier. The findings contribute to increases our knowledge of axonal sub-cellular components and help to understand better the mechanism behind axonal brain injury.

Keywords: coarse grained model, molecular dynamics simulations (MD simulations), brain injury, axonal membrane, mechanoporation

OPEN ACCESS

Edited by:
Fabio Trovato,
Freie Universität Berlin, Germany

Reviewed by:
Sudipta Kumar Sinha,
Indian Institute of Technology Ropar, India
Conrado Pedebos,
University of Southampton, United Kingdom

****Correspondence:***
Alessandra Villa
avilla@kth.se

†Present address:
Marzieh Saeedimasine,
Department of Materials and Environmental Chemistry, Stockholm University, Stockholm, Sweden

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *19 February 2021*
Accepted: *22 April 2021*
Published: *23 June 2021*

Citation:
Saeedimasine M, Montanino A, Kleiven S and Villa A (2021) Elucidating Axonal Injuries Through Molecular Modelling of Myelin Sheaths and Nodes of Ranvier. Front. Mol. Biosci. 8:669897. doi: 10.3389/fmolb.2021.669897

1 INTRODUCTION

Traumatic brain injury (TBI) (Menon et al., 2010) is an injury to the brain caused by an external force. Causes include falls, vehicle collisions and violence. One of the most common consequence of a TBI is diffuse axonal injury, a multifocal damage to white matter (Johnson et al., 2013). Such an injury is invisible to conventional brain imaging, and can only be histologically diagnosed and one of its hallmarks is the presence of axonal swellings (Hill et al., 2016).

Axons are long projection of the nerve cell surrounded by a membrane, called axolemma. Axolemma separates the interior of the axon from the outside environments. In the nervous system, axons may be myelinated, or unmyelinated. When the axon is myelinated, a extra multilamellar membrane, called myelin sheath, insulates segments of axon. The sheath consists of repeating units of double bilayers separated by 3–4 nm-thick aqueous layers that alternate between the

cytoplasmic and extracellular faces of membranes (Inouye and Kirschner, 1988a). Dehydrated myelin is unusual in that it is composed of 75–80% lipid and 20–25% protein by weight, compared with around 50% of lipids in most other cell membranes (Williams et al., 1993). Multiple lipids make up the myelin sheath, and each sheath contributes to the structure, adhesive stability, and possibly the pathogenesis of the myelin membrane. The asymmetric distribution of lipid composition on the cytoplasmic and extracellular faces likely also plays an important role (Inouye and Kirschner, 1988b). Axolemma and myelin compositions differ both in lipid type and degree of saturation. **Table 1** summarizes the available experimental data on lipid composition for axolemma and myelin. The most striking feature of myelin lipid composition is the enrichment in glycolipids together with long-chain fatty acids (DeVries et al., 1981).

Short unmyelinated segments, nodes of Ranvier, occur periodically between segments of the myelin sheath in myelineted axons. In nodes of Ranvier, the axolemma is directly exposed to the extracellular space. The nodes are uninsulated and highly enriched in ion channels (Westenbroek et al., 1989; Catterall et al., 2005), allowing them to participate in the exchange of ions required to regenerate the action potential. One of the voltage-gated ion channels, embedded in myelinated axons, is sodium channel protein type subunit alpha, Nav1.1 (Duflocq et al., 2008). In axonal membrane, the density of sodium channels varies between 5 and 3,000 channels/μm^2: lower densities are observed in unmyelinated axons, while higher densities are found in the nodal portions of myelinated axons (Wann, 1993; Hu and Jonas, 2014). Interestingly, sodium channels have also been proposed to influence the injury response (Iwata, 2004).

Experimental observations have so far led to the formulation of two main theories regarding the cellular primary injury mechanism. Disruption of the axolemma (Pettus and Povlishock, 1996; LaPlaca and Thibault, 1997; Fitzpatrick et al., 1998), or alternatively cytoskeletal damage (Tang-Schomer et al., 2012) has been suggested mainly as injury trigger. However, using a purely mechanical approach we discarded microtubule damage as injury trigger and revealed instead high level of strains on the axonal membrane (Montanino and Kleiven, 2018). To further investigate the molecular level effects of such strains, we bridged the finite element model of the axon with a molecular-based membrane model (Montanino et al., 2020). Despite the approximation of the models, we showed that in a typical injury scenario, the axonal cortex sustains deformations large enough to entail pore formation in the adjoining lipid bilayer. The observed axonal deformation at which poration occurs (10–12%) agrees well with the thresholds obtained both with *in-vitro* (Hemphill et al., 2011; Nakadate et al., 2017) and *in-vivo*/*ex-vivo* (Bain and Meaney, 2000; Singh et al., 2015) stretch injury experiments and allows us to provide quantitative evidences that do not exclude pore formation in the membrane as a result of trauma.

When investigating the progression of axonal injury in the white matter, several studies have observed impairments at the myelin level in the form of delamination of the myelin lamellae (Maxwell, 2013) and general loss of myelin, or demyelination, (Mierzwa et al., 2015; Mu et al., 2019). While white matter degeneration is undoubtedly a distinctive feature of traumatic brain injury, the causal relationships between myelin disruption and the multitude of events associated with axonal damage is still unclear. Computational models of the axon can potentially clarify the damage-causality chain, provided that a mechanical description of myelin is obtained.

Here, we aim to describe membrane-component of myelinated axon at molecular level and to elucidate if and where membrane rupture can occur as a result of an injury. To achieve this, we have used molecular-based model for myelin sheath and for node of Ranvier in line with experimental composition (**Table 1**) and a coarse grained (CG) description. For myelin, we have built ex-novo the model based on experimental lipid composition and saturation from central nervous system (O'Brien and Rouser, 1964; Manzoli et al., 1970; Inouye and Kirschner, 1988b; Bosio et al., 1998). For the node of Ranvier, we have used an already available plasma membrane model (Ingólfsson et al., 2014), since too few experimental data on the axolemma composition for central nervous system are available. Molecular dynamics (MD) simulations have been used to characterize the molecular system at equilibrium and under deformation. Finally the deformation results from molecular simulations are combined with the axonal deformation obtained using an axonal finite element (FE) model (Montanino et al., 2020).

2 METHODS

2.1 Molecular Model

2.1.1 Model for Myelin Sheath

The myelin is described as lipid bilayer. The membrane model contains 17 different types of lipids, distributed asymmetrically

TABLE 1 | Experimental and modelled lipid composition in weight fraction for different types of membrane.

Lipid type Membrane	PC	PE	SM	PS	Glycolipid	CHOL	Other
Myelin (exp)[a]	0.11	0.17	0.06	0.06	0.28	0.28	0.04
	(0.01)	(0.01)	(0.03)	(0.01)	(0.03)	(0.01)	(0.02)
Myelin model	0.12	0.22	0.03	0.09	0.23	0.29	0.02
Axolemma (exp)[b]	0.20	0.19	0.08	0.06	0.20	0.24	0.10
	(0.09)	(0.07)	(0.05)	(0.03)	(0.05)	(0.03)	(0.07)
Plasma model[c]	0.30	0.17	0.17	0.06	0.05	0.17	0.08

For composition observed experimentally the average values is reported together with the maximum deviations from the mean values (in parentheses).

[a]*data for the central nervous system of human, bovine and rat from Rasband and Macklin (2012) and references inside.*

[b]*data from central and peripheral nervous system from Camejo et al. (1969); Zambrano et al. (1971); DeVries et al. (1981); DeVries et al. (1999).*

[c]*plasma membrane model of Ingólfsson et al. (2014) was used to describe the axolemma.*

between two leaflets, labelled as extracellular and cytoplasmic leaflet. The extracellular leaflet is characterized by having phosphatidylcholine (PC) (7%), phosphatidylethanolamine (PE) (8%), cholesterol (CHOL) (43%), cerebroside (39%), sphingomyelin (SM) (2%), phosphatidylserine (PS) (1%) and other anionic lipids, while the cytoplasmic leaflet contains PC (11%), PE (27%), CHOL (44%), SM (3%), PS (12%) and other anionic lipids. Note that the reported values are in mol%. **Supplementary Table S1** lists all the lipid types used and their disposition in the leaflets together with the level of saturation of fatty acids. The membrane model was built accounting the available experimental data: we use the values derived by Inouye and Kirschner (Inouye and Kirschner, 1988b) for myelin in central neurons system to define the lipid disposition in the extracellular and cytoplasmic leaflet; the work of Manzoli and coworkers (Manzoli et al., 1970) to define the saturation of the lipid tails for phospholipids, while the saturation values for cerebroside have obtained from O'Brien and Rouser (1964) and Bosio et al. (1998).

2.1.2 Model for Node of Ranvier

To describe the region of Node of Ranvier, we use a lipid bilayer embedded with ion channels. A channel concentration of 3,086 channels/was used to describe the node. The mammalian plasma membrane model designed by Ingólfsson et al. (2014) and deposited on MARTINI webpage (http://www.cgmartini.nl/) was used to describe the membrane. The model contains 63 different types of lipids distributed asymmetrically between two leaflets. The extracellular leaflet has a higher level of tails saturation and contains PC (36%), PE (6%), CHOL (31%), SM (19%), glycolipids (6%), and other lipids (2%). The cytoplasmic leaflet, which has a higher level of polyunsaturation, contains PC (17%), PE (25%), CHOL (29%), SM (9%), PS (11%), anionic phosphatidylinositol (PI) (2%), and other lipids (7%).

As model for the protein we used the *Homo sapiens* Nav1.1. The three-dimensional (3D) structure of Nav1.1 has not been yet resolved experimentally, however, the encoding gene is known (SCN1A gene) (Malo et al., 1994). We used the 3D structure obtained previously using homology modeling (for details see (Montanino et al., 2020)). As template, the cryo-electron microscopy structure of putative sodium channel from American cockroach, NavPaS (PDB ID: 5X0M) (Shen et al., 2017) was used. The structure had 100% confidence (matching probability) and 48% sequence identity (identical residues) with NavPaS. The missing terminal domains and linkers were added and the structure was energy minimized and shortly equilibrated in the PC lipid bilayers at the atomistic level using CHARMM36 force field (Pastor and MacKerell, 2011; Best et al., 2012) before building the CG model.

2.1.3 Force Field

The membrane and protein systems were described at the coarse-grained level using the MARTINI2.2 force field (Marrink et al., 2004; Marrink et al., 2007; Monticelli et al., 2008; de Jong et al., 2013) together with one-bead non-polar water model (Marrink et al., 2007). In the MARTINI model, small groups of atoms (3–4 heavy atoms) are united into beads which interact with each other by means of empirical potentials.

2.2 Systems Setup

2.2.1 Myelin Model

The myelin was modelled as a bilayer having the composition reported in **Table 1**. The initial lipid coordinates were constructed using the 2,000 lipids and an initial box of 24 × 24 × 12 nm. The equilibrated 20 nm bilayer (at 16 μs) was used as template to build a 40 nm bilayer and a multi-bilayers system. The 40 nm bilayer system has a total of 8,000 lipids and was placed in a cubic box of 41 × 41 × 14 nm and solvated by about 138,000 CG water beads. The multi-bilayers system was built using a distance of 1.7 and 1.6 nm between extracellular layers and cytoplasmic layers, respectively, and was placed in a cubic box of 21 × 21 × 48 nm and solvated by about 106,000 CG water beads. To all the systems NaCl was added to mimic the ionic strength at physiological condition (150 mM NaCl).

2.2.2 Node-of-Ranvier Model

A channel concentration of 3,086 channels/μm^2 was used to describe the node. The bilayer (containing a total of 6,500 lipids) was placed in a cubic box (42 × 42 × 18 nm) and solvated by about 190,000 CG water beads. NaCl was added to mimic the ionic strength at physiological condition (150 mM NaCl). To achieve a concentration of 3,086 channels/μm^2, 16 copies of proteins were embedded in a larger bilayer (containing 15,000 lipids in a cubic box of 69 × 69 × 20 nm). The proteins were located at 5 nm distance from each other.

2.3 Molecular Dynamics Simulations

All MD simulations were performed using the GROMACS simulation package, version 2016 (Abraham et al., 2015) (manual.gromacs.org/2016). The bilayer systems were equilibrated at constant temperature (37°C) and pressure (1 bar). The temperature was held constant using velocity rescale thermostat (Bussi et al., 2009) with a time constant of 1.0 ps. The pressure was held constant using semi-isotropic Parrinello-Rahman barostat (Parrinello and Rahman, 1981) with a time constant of 12 ps (compressibility of 3×10^{-4} bar^{-1}). The Verlet cutoff scheme (Páll and Hess, 2013) and a timestep of 10 fs (for myelin system) and 15 fs (for protein-membrane system) were used. Periodic boundary conditions were applied. Non-bonded interactions were calculated between all beads using a cutoff of 1.1 nm. Long-range electrostatic interactions were treated using a reaction field potential (Tironi et al., 1995) with switching distance of 1.1 nm in line with MARTINI setting. After a short equilibration, 16 μs were performed for both node-of-Ranvier and 20 nm myelin models, while 25 μs MD simulations were performed for 40 nm and multi-layer myelin model. **Supplementary Figure S1** shows equilibration of box dimension and energy within the first 16 μs. Note, the protein positions were kept fixed in only the simulations at equilibrium to allow the relaxation of the lipid distribution.

To evaluate membrane behavior under mechanical stress we performed 6 μs simulations at constant areal strains (NP_zAT ensemble) for the node-of-Ranvier and 40-nm myelin model.

No position constrains was applied to the proteins Values < 0.05 were considered for areal strain. The last 4 μs are used for data production. To monitor the pore formation, we first fast pre-deformed the bilayers in x direction up to 30% strain (using a deformation speed of 1×10^{-5} nm/ps), then starting from 30% strain two independent simulations were performed using slower deformation speeds: 2.5×10^{-6} nm/ps and 25×10^{-6} nm/ps. Stretching simulations were performed for the node-of-Ranvier and myelin (40 nm and multi-layer) model. No position constrains was applied to the proteins.

2.4 Simulation Analysis

To describe the structure of lipid bilayers, we calculated bilayer thickness and number of lipid contacts. Bilayer thickness was calculated using the method proposed by Pandit et al. (2003), Pandit et al. (2004) and implemented in APL@Voro software (Lukat et al., 2013). Using Voronoi tessellation, the methods identifies the normal distance between phosphorus atoms from two leaflets that are "vertical neighbors" of each other.

Lipid contacts were calculated by counting the number of neighboring lipids within 1.5 nm of proteins or lipids by considering the first tail bead after the headgroup of lipid or polar group of cholesterol molecules. Enrichment factors were calculated as (Corradi et al., 2018):

$$Enrichment\,(L) = \frac{Ratio\,(L)_x}{Ratio\,(L)_{bulk}} \tag{1}$$

In which $\text{Ratio}\,(L)_x = \frac{(\text{no.}L)_x}{(\text{tot.no.lipids})_x}$ and $\text{Ratio}\,(L)_{bulk} = \frac{\text{tot.no.}\,(L)}{\text{tot.no.lipids}}$ where x was chosen 0.7 and 2.1 nm for lipids around the protein in node-of-Ranvier model and 1.5 nm for lipid distribution around each lipid type in myelin model.

The partition coefficient $K_{mem/wat}$ was calculated as $[solute]_{mem}/[solute]_{wat}$, where [solute] denotes the concentration of water molecules in the lipid bilayers and water solution, respectively. $[water]_{mem}$ was obtained by dividing the average number of water molecules by the lipid bilayers volume, while for the water concentration in water the experimental value (55.5 mol/L) was used. The membrane volume was corrected to account for the presence of the protein.

The software VMD (Humphrey et al., 1996) was used for graphical representations. Reported values were averaged on 5 μs and the errors were obtained by dividing the data production into five parts and calculating the standard error between them, if it was not specified differently.

2.5 Mechanical Properties

Area compressibility modulus (K_A) is defined as the derivative of surface tension as function of the areal strain,

$$K_A = \left(\frac{\partial \gamma}{\partial \epsilon_A}\right) T \tag{2}$$

where γ is the surface tension and the areal strain (ϵ_A) is defined as $\epsilon_A = \left(\frac{A}{A_0}\right) - 1$. **Eq. 2** was chosen to be in line with experimental approach (Needham and Nunn, 1990; Rawicz et al., 2000) used on biological membrane. Evans et al. (1976) and Needham and Nunn (1990) observed a linear relation between surface tension for small areal change (<0.05 areal change) in their study on red blood cell and chlosterol-rich membrane, respectively.

To calculate area compressibility modulus, we used NP_zAT simulations at three strain values ($\epsilon_A < 0.05$). We divided production data in 4 parts and calculated the average γ and ϵ_A values γ is calculated according to the following

$$\gamma = l_z\left(P_{zz} - \frac{P_{xx}}{2} - \frac{P_{yy}}{2}\right) \tag{3}$$

where P_{xx}, P_{yy}, and P_{zz} are diagonal elements of pressure matrix and l_z is the box height along z.

We performed linear regression between the obtained γ and ϵ_A values, the slope of the regression line is K_A and the standard error of the slope is the standard error of K_A.

2.6 Multiscale Approach

To provide insights into the initiation of axonal damage, we combined the FE model of a generic portion of an axon with the molecular-based myelin and node-of-Ranvier models. The node-of-Ranvier model was used to exemplify the approach in **Figure 1**. Firstly, the axon FE model was utilized to simulate typical stretch injury scenarios, then as a result of axonal deformation, maximum local deformations happening at the cortex level was extracted and applied to the molecular models. The multiscale approach together with details on axon FE simulations was previously described and published in Montanino et al. (2020). The axon FE model (**Figure 1C**) was validated and described in Montanino and Kleiven (2018). The axon FE model is a 8 μm long representative volume of an axon consisting of three main compartments: a microtubule (MT) bundle, the neurofilament network and the axolemma-cortex complex wrapping the entire structure. To create a general axonal behavior, 10 different FE axon models were generated by randomly moving the MTs discontinuities locations, while keeping the average MTs length of the original model. Details regarding the element and material modelling choices can be found in Montanino and Kleiven (2018). All simulations were performed in LSDYNA using an implicit dynamic solver and the 1st and 2nd principal strains (ϵ_x, ϵ_y) in the cortex plane were extracted as function of axonal strain for axonal strain rates of 1, 20, and 40/s.

3 RESULTS AND DISCUSSION

3.1 Structural Features of Myelin Model

The myelin membrane was modelled as a lipid bilayers, having 17 different types of lipids, distributed asymmetrically between two leaflets, labelled as extracellular and cytoplasmic leaflet (**Figure 2**). The lipids distribution and saturation were assigned accounting for a collection of experimental data on leaflet mole composition, fatty acid length and saturation. In particular, we used the values derived by Inouye and Kirschner

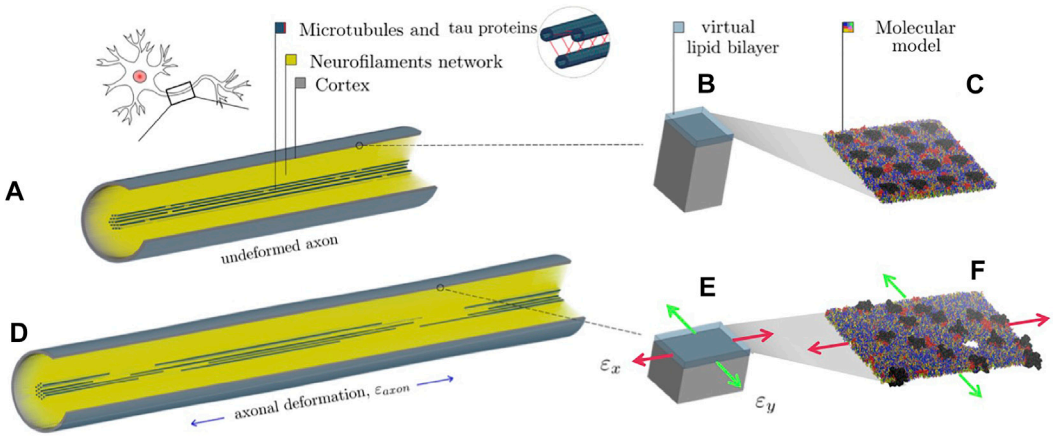


FIGURE 1 | Representation of the combined finite element-molecular based modeling approach (Montanino et al., 2020). **(A)** Finite element axonal model where a quarter of the model was removed to reveal the arrangement of the inner structures. **(B)** Single undeformed cortex element and a juxtaposed virtual lipid bilayer. **(C)** molecular bilayer model. **(D)** The axon finite element model is deformed of a quantity (ε_{axon}). **(E)** Maximum localized deformations ($\varepsilon_x \varepsilon_y$) are extracted and applied to the molecular model. **(F)** Visualization of the induced pore upon deformation of the molecular model.

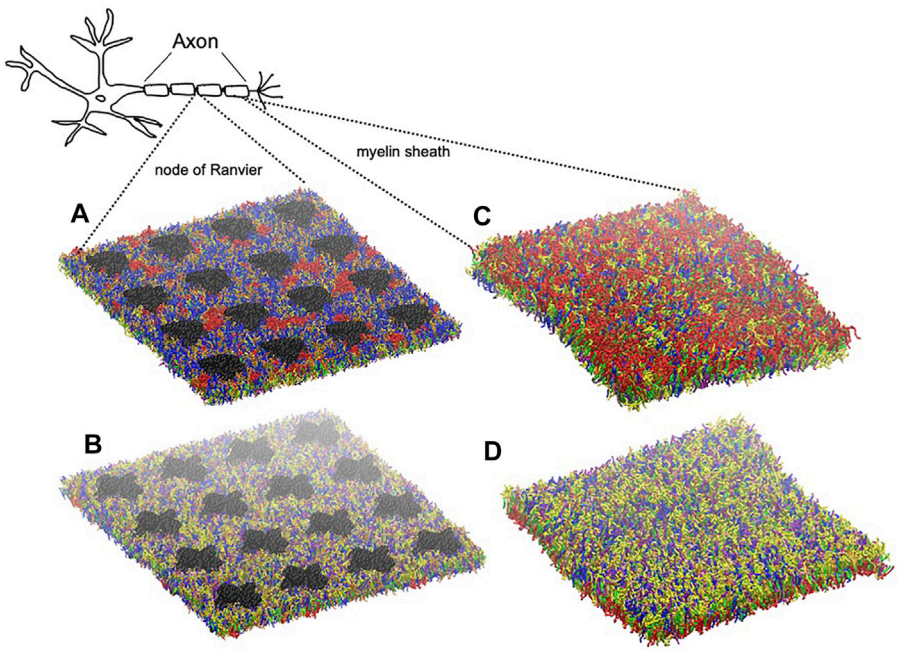


FIGURE 2 | Graphical representation of node-of-Ranvier bilayer model (70 nm) at 16 μs **(A, B)** and myelin bilayer model (40 nm) at 25 μs **(C, D)** together with a draw of a neuron. Both up and down leaflets are shown. Molecules are colored as follows: PC lipids are in blue, PE in yellow, CHOL in green, SM in orange, PS in purple, glycolipids in red, anionic lipids in pink, the rest of lipids in gray, protein in black VDW representation, and water is not visualized for clarity.

(1988b) for myelin in central neurons system to define the lipid disposition in the extracellular and cytoplasmic leaflet; the work of Manzoli et al. (1970) to define the saturation of the lipid tails for phospholipids, while the saturation values for cerebroside were obtained from O'Brien and Rouser (1964) and Bosio et al. (1998). **Supplementary Table S1** reports details on the lipid composition of the myelin model together with the experimental reference. In **Table 1** we compare the model's lipids abundance with a wide collection of data on lipid composition observed for myelin sheath in the central nervous system of diverse species (human, bovine, rat). The myelin model has a composition in line with experimental data, small deviations are observed for PE, PS and glycolipid, whose weight fraction is between myelin and axolemma experimental composition.

TABLE 2 | Bilayers structural and mechanical properties. In parenthesis standard error.

Properties Simulated Models	Area compressibility Modulus K_A (mN/m)	Thickness Coefficient (*nm*)	Water partition $logK_{mem/wat}$	Rupture Strain
model for unmyel.axon	352 (3)[a]	4.168 (0.001)[a]	−1.75 (0.01)[a]	0.47[a]; 0.51[a]
node-of-Ranvier model	252 (20)	4.032 (0.001)	−1.50 (0.01)	0.36[b]; 0.40[c]
myelin model	339 (12)	4.506 (0.015)	−1.98 (0.01)	0.70[b]; 0.82[c]

Lowest observed rupture strains are reported for each deformation rate.
[a]*Data from Montanino et al. (2020).*
[b]*deformation rate 2.5 × 10^{-6} nm/ps.*
[c]*deformation rate 25 × 10^{-6} nm/ps.*

TABLE 3 | Lipd enrichment factor in myelin bilayer. Number of lipids within 1.5 nm of a reference lipid type are reported for the extracellular and cytoplasmic leaflet separately. Standard errors are less than 0.1

Lipids type	PC	PE	PS	SM	PI	Glycolipid	CHOL
Extracellular leaflet							
PC	1.15	1.16	1.15	1.11	1.14	0.84	1.08
PE	1.16	1.14	1.08	1.03	1.07	0.85	1.07
PS	1.12	1.04	0.89	1.17	0.91	0.88	1.07
SM	1.13	1.05	1.23	1.08	1.13	0.88	1.06
PI	1.13	1.06	0.94	1.10	0.38	0.89	1.07
Glycolipid	0.77	0.78	0.83	0.79	0.82	1.03	1.07
Cytosplasmic leaflet							
PC	0.99	1.00	1.03	1.00	1.00	-	1.00
PE	1.00	0.99	1.04	0.97	1.04	-	0.99
PS	1.02	1.02	0.95	1.02	0.94	-	1.00
SM	1.03	0.99	1.06	1.01	1.02	-	0.98
PI	1.01	1.04	0.96	1.01	0.92	-	0.99

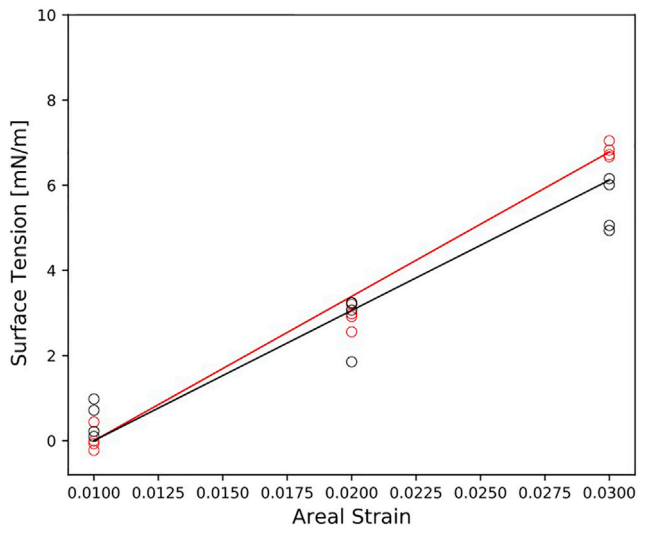


FIGURE 3 | Surface tension as function of the areal strain for myelin and node-of-Ranvier model, colored in red and black respectively. Reported data (circles) corresponds to the average on 1 μs time-window. The slope of the line is K_A.

The myelin bilayers were simulated for 16 μs. Box dimensions and total energy converged in the first microseconds (**Supplementary Figure S1**). To check the convergence of the lipid distribution, we compare the numbers of lipid neighboring for the 20 and 40 nm bilayers: the values over the last 5 μs agrees within the fluctuations (**Supplementary Figure S2**).

The equilibrated myelin bilayer has a thickness of 4.51 nm (**Table 2**), larger than axolemma bilayers and comparable with the value derived using cable theory and the mean dielectric constant of squid axon myelin (around 4.5 nm) (Min et al., 2009). **Table 3** reports the lipid enrichment factor. In the extracellular leaflet a depletion of phospholipids (in particular PC and PE, and SM) is observed around glycolipids. A deplection of PI around PI and enrichment of PS around SM are also reported. In the cytoplasmic leaflet no relevant enrichment/depletion is observed, most of the values are with 5% from an homogeneous distribution.

3.2 Node-of-Ranvier Model

Node of Ranvier was modelled as lipid bilayer with embedded ion channels. We use a plasma membrane lipid bilayers composition. **Table 1** compares the composition with lipid abundance observed for axolemma in mammalian neuron system. The values are closer to the composition of axolemma than to the one of myelin. The larger deviation from experimental values is observed for the glycolipids. We have also to note that we can not exclude that the experimental composition for axolemma may be contaminated by myelin lipids (DeVries et al., 1981). In comparison with the composition used to describe myelin sheath, model for node-of-Ranvier is characterized by a lower content of glycolipids and CHOL molecules and higher content of PC and SM lipids, in line with the trend shown by experimental values. The model is also characterized by having shorter long-chain fatty acids than myelin model in line with experimental observation (DeVries et al., 1981).

To describe the proteins in the node, we use sodium channel protein type subunit alpha (Nav1.1), a characteristic protein of nodal portions of myelinates axons, belonging to the family of voltage-gated ion channels (Duflocq et al., 2008). It is known that several different ion channels are present, but currently information are available to model solely sodium channel protein type subunit alpha (Nav1.1). 16 proteins were embedded in the bilayers (equivalent to 3,000 channels/μm^2) to mimic a value of ion channel concentration in line with the

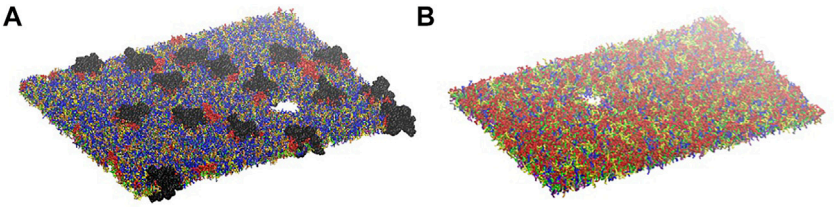


FIGURE 4 | Node-of-Ranvier **(A)** and myelin **(B)** bilayers at 0.36 and 0.70 strain, respectively. For color code see **Figure 2**.

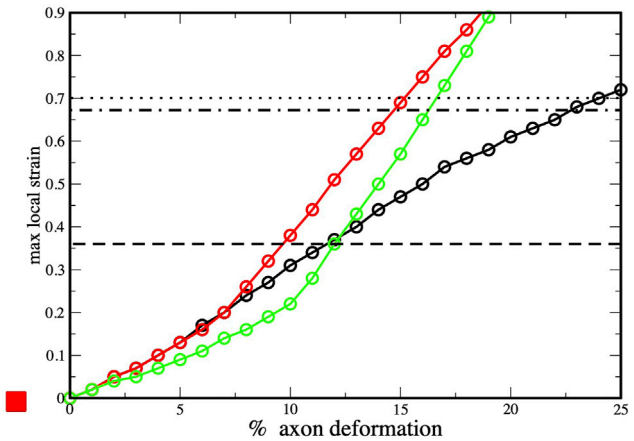


FIGURE 5 | Axonal deformation versus maximum local strains at three strain rates: 1 s^{-1} (black), 10 s^{-1} (red) and 40 s^{-1} (green). Values obtained from finite element simulations of axonal stretch (Montanino et al., 2020). The dotted line corresponds to rupture strain observed for myelin model (and dashed-dot lines to the myelin multi-layer model), while the dashed line corresponds the to rupture strain observed for the node-of-Ranvier model.

experimental observation for the node of Ranvier (Wann, 1993; Hu and Jonas, 2014). Below we compare the results with previously performed simulations of a bilayer, having the same lipid composition as the node but only one embedded protein, representing a condition closer to axolemma in unmyelinated axon (Montanino et al., 2020).

We checked the convergence of lipid distribution around the proteins by monitoring the lipid-protein contacts (**Supplementary Figure S3**). In general, the lipid-protein contacts are in overall agreement with the contact recorded for one protein embedded in the bilayer. The node-of-Ranvier bilayer has a thickness of around 4.0 nm (**Table 2**). High protein density makes the bilayer slightly thinner: a values of 4.2 was calculated for the bilayer with one embedded protein. The difference in lipid composition (both in tail length and head-group) between the node-of-Ranvier and myelin bilayers results difference in the thickness and the water permeability, given the proportionality between the partition coefficient $K_{mem/wat}$ and the permeability (P_m). The myelin is thicker and is less permeable to water than the node of Ranvier. Now we can proceed to evaluate what is the effect of the composition on the mechanical feature.

3.3 Bilayer Models Under Deformation

To study the bilayers response to mechanical deformation, we performed simulations at three different areal strains. **Figure 3** reports the surface tension as a function of the areal strain for each membrane model. In general, the surface tension increases linearly for small areal strain. The myelin model shows the steepest slope of surface tension-areal strain (**Figure 3**). That means that the myelin model is more resistant to change its area than node-of-Ranvier model.

The area compressibility modulus, K_A, was estimated for small areal strain values (less than 0.05) and reported in **Table 2**. A larger K_A values is observed for the myelin model than for node-of-Ranvier models: a value of 339 mN/m and 252 mN/m, respectively for myelin and node-of-Ranvier bilayers, was obtained. At this point we can not distinguish if the difference is due to the protein concentration, lipid composition or a combination of them. The value for the axolemma model in unmyelinated (plasma membrane with one embedded protein) is in line with the area compressibility modulus measured for red blood cell membranes (375 ± 60 mN/m) at 37°C by Waugh and Evans (1979). Unlucky a validation for all the models toward experiments is not possible due to lack of experimental data. A previous study (Saeedimasine et al., 2019) shows that atomistic and CG simulations give different K_A values, but similar trend upon change in lipid composition. That gives us trust in the observed trend, even if the absolute values can not be validated.

Simulations at different deformation rate were performed to identify the rupture points. We define the rupture point when at least one pore is formed in the bilayer. To detect the formation we monitored the surface tension: a jump in the surface tension corresponds to the formation of at least one pore. In the myelin model pores were observed at 0.7–0.8 strain while in the node-of-Ranvier model at around 0.4 (**Figure 4**). Poration in both bilayers occurs in region lacking protein and glycolipids. This type of lipids are mainly found in the outer leaflet of membranes and are known to make the bilayer more resistant to deformation thanks to interactions between sugar headgroups (Saeedimasine et al., 2019).

To provide insights into the initiation of axonal damage, we combined the FE model of a generic portion of an axon with the molecular-based myelin and node-of-Ranvier models. Firstly, the axon FE model was utilized to simulate typical stretch injury scenarios, then as a result of axonal deformation, maximum local deformations happening at the cortex level was extracted and applied to the molecular models (**Figure 1**). **Figure 5** shows the relation between the applied axon deformation and the resulting maximum local deformation observed at different strain rates.

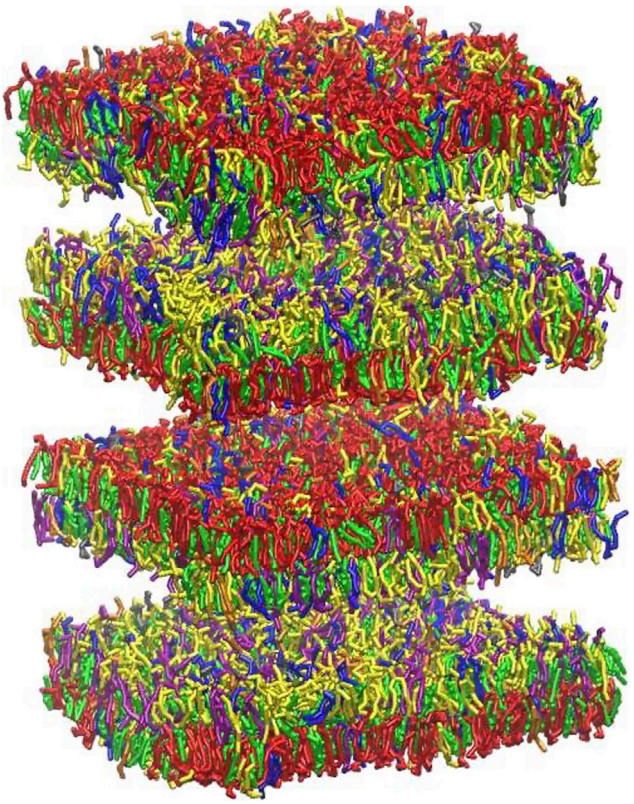


FIGURE 6 | Graphical representation of multi-layer myelin at 25 μs. The multi-bilayers system was built using a distance of around 1.7 nm between extracellular and cytoplasmic layers. For lipid molecules color code see **Figure 2** and for bilayer composition see **Table 2**.

We found that higher strain rates do not lead to higher local maximum strains for axonal strains less than 12%. This is due to the viscoelastic properties of tau proteins and of the cortex itself resulting in less cortex deformation at strain rate 40/s compared to 1 or 20/s (Montanino et al., 2020).

We connect the bilayer strain at which local events occurs in molecular model (e.i pore formation) with the maximum local strain observed in the finite element simulations during axonal deformation. This allows to understand which axonal strain correspond to the bilayer rupture. The node-of-Ranvier model can withstand an applied deformation up to 36%, corresponding to axonal deformation of 9–11% (see dashed line in **Figure 5**). At higher strains pore formation is observed. While poration occurs at a larger cortex strain (70%) for the myelin model and this corresponds to an axonal deformation of 15–23%.

Myelin sheath consists of not one bilayer but of repeating units of double bilayers separated by aqueous layers. To verify the effect of the multi layers structure on the rupture threshold, we have built a double bilayers myelin model (**Figure 6**), equilibrated and deformed it as it was done for the single bilayer model. The distances between the bilayers were taken from X-ray diffraction data (Caspar and Kirschner, 1971; Kirschner et al., 1989): 1.6 and 1.7 nm were reported as widths of the spaces between membranes at the cytoplasmic and extracellular appositions, respectively, for the central nervous system. Starting from a strain value of 0.66 pore formation can occur in one of the bilayers (**Figure 6**). The first pore formation is followed by formation of multiple pores in different layers (at strain 0.73). This corresponds to an axonal strain of 14–22%, in line of what observed for the single-bilayer myelin model.

All in all, our results show that mechanoporation occurs at 49–74% higher strain for the myelin and at 15–23% lower strain at the node of Ranvier compared with axolemma with one embedded protein. *In vivo* and *ex vivo* experiments on uniaxially oriented neuronal cell and spinal nerve showed injurious changes above 10% applied strain (Singh et al., 2015; Nakadate et al., 2017). Experiments conducted on non-oriented neuronal culture indicated that mechanoporation occurs only when strains higher than 10% were uniaxially applied to the culture (Hemphill et al., 2011). If mechanico-poration trigger axonal brain injury, our results indicate that such events in myelined axon will occur at the level of the node of Ranvier. Indeed, according to our models, poration in the node-of-Ranvier occurs at axonal strains going from 9% to 11%, while myelin rupture occurs at axonal strains larger than 15%.

Although our modeling approach brings some insights into the axonal injury mechanism, note that we do not aim to describe the injury event, since those events occur at time scale (fraction of seconds) not accessible to current molecular simulations. Moreover, the simulated bilayers are simplified models of cellular membrane: they do not account for all the different embedded proteins, they are described by a coarse grained model (where a group of atoms is described by a particle). Finally, the model does not account for possible protein structural change at different strains, since the protein is described as a semi-rigid body.

4 CONCLUSION

In this study, we combine CG molecular dynamics simulations and FE method to improve our understanding of what can trigger axonal brain injury. To achieve this we used molecular models to describe sub-cellular components of the myelinated axon, in particular myelin sheath and node of Ranvier, and a FE simulations of the axon.

The ex-novo built myelin model well reproduces the myelin experimental composition and structural feature of myelin. The model is more glycolipid-rich with longer fatty acid tails than the model used to describe the node-of-Ranvier. The myelin bilayer has a higher thickness, rupture point and area compressibility modulus than node of Ranvier. The results support a scenario where node-of-ranvier is most vulnerable component, myelin has more of a supporting role given that it likely will not porate before the axolemma, in unmyelined axon, and node-of-Ranvier do.

Combining the results with finite element simulations of the axon model, we provided quantitative evidences that

mechanoporation of the axon membranes is an event that cannot be excluded in a typical axonal injury scenario. Our results indicate that mechanoporation may occur at the node of Ranvier at the thresholds proposed in the literature for axonal injury, while no myelin rupture is observed at the injury threshold. Poration occurs at more than 50% higher strain for the myelin compared with node-of-Ranvier and axolemma (in unmyelinated axon). The findings contribute to increase our knowledge of axonal sub-cellular components and help to understand better the mechanism behind axonal brain injury. Finally, our work shows the power of combining FE and MD to describe complex biological scenario, that requires the description of different length scale.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author. Topology and coordinate files for myelin model are deposited in https://github.com/alevil-gmx/myelin_model.

AUTHOR CONTRIBUTIONS

All the authors designed the study. MS performed and analyse molecular dynamics simulation. AV analysed the results and prepared the manuscript. All authors critically reviewed the manuscript.

FUNDING

The authors thank the Swedish Research Council (VR-2016-05314 and VR-2020-04496), the European Union Horizon 2020 Research and Innovation Framework Programme under the Marie Skłodowska-Curie (grant agreement No. 642662), and the Swedish National Infrastructure for Computing (SNIC2017-1-491 and SNIC2018-3-548) for the support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.669897/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

Bain, A. C., and Meaney, D. F. (2000). Tissue-Level Thresholds for Axonal Damage in an Experimental Model of Central Nervous System White Matter Injury. *J. Biomechanical Eng.* 122, 615–622. doi:10.1115/1.1324667

Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., et al. (2012). Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. *J. Chem. Theor. Comput.* 8, 3257–3273. doi:10.1021/ct300400x

Bosio, A., Binczek, E., Haupt, W. F., and Stoffel, W. (1998). Composition and Biophysical Properties of Myelin Lipid Define the Neurological Defects in Galactocerebroside- and Sulfatide-Deficient Mice. *J. Neurochem.* 70, 308–315. doi:10.1046/j.1471-4159.1998.70010308.x

Bussi, G., Zykova-Timan, T., and Parrinello, M. (2009). Isothermal-isobaric Molecular Dynamics Using Stochastic Velocity Rescaling. *J. Chem. Phys.* 130, 074101. doi:10.1063/1.3073889

Camejo, G., Villegas, G. M., Barnola, F. V., and Villegas, R. (1969). Characterization of Two Different Membrane Fractions Isolated from the First Stellar Nerves of the Squid Dosidicus Gigas. *Biochim. Biophys. Acta (Bba) - Biomembranes* 193, 247–259. doi:10.1016/0005-2736(69)90186-2

Caspar, D. L. D., and Kirschner, D. A. (1971). Myelin Membrane Structure at 10 Å Resolution. *Nat. New Biol.* 231, 46–52. doi:10.1038/newbio231046a0

Catterall, W. A., Goldin, A. L., and Waxman, S. G. (2005). International Union of Pharmacology. XLVII. Nomenclature and Structure-Function Relationships of Voltage-Gated Sodium Channels. *Pharmacol. Rev.* 57, 397–409. doi:10.1124/pr.57.4.4

Corradi, V., Mendez-Villuendas, E., Ingólfsson, H. I., Gu, R.-X., Siuda, I., Melo, M. N., et al. (2018). Lipid-Protein Interactions Are Unique Fingerprints for Membrane Proteins. *ACS Cent. Sci.* 4, 709–717. doi:10.1021/acscentsci.8b00143

de Jong, D. H., Singh, G., Bennett, W. F. D., Arnarez, C., Wassenaar, T. A., Schäfer, L. V., et al. (2013). Improved Parameters for the Martini Coarse-Grained Protein Force Field. *J. Chem. Theor. Comput.* 9, 687–697. doi:10.1021/ct300646g

DeVries, G. H., Campbell, B., and Saunders, R. (1999). Isolation and Characterization of Unmyelinated Axolemma from Bovine Splenic Nerve. *J. Neurosci. Res.* 57, 670–679. doi:10.1002/(sici)1097-4547(19990901)57:5<670::aid-jnr9>3.0.co;2-b

DeVries, G. H., Zetusky, W. J., Zmachinski, C., and Calabrese, V. P. (1981). Lipid Composition of Axolemma-Enriched Fractions from Human Brains. *J. Lipid Res.* 22, 208–216. doi:10.1016/s0022-2275(20)35364-5

Duflocq, A., Le Bras, B., Bullier, E., Couraud, F., and Davenne, M. (2008). Nav1.1 Is Predominantly Expressed in Nodes of Ranvier and Axon Initial Segments. *Mol. Cell Neurosci.* 39, 180–192. doi:10.1016/j.mcn.2008.06.008

Evans, E. A., Waugh, R., and Melnik, L. (1976). Elastic Area Compressibility Modulus of Red Cell Membrane. *Biophysical J.* 16, 585–595. doi:10.1016/S0006-3495(76)85713-X

Fitzpatrick, M. O., Maxwell, W. L., and Graham, D. I. (1998). The Role of the Axolemma in the Initiation of Traumatically Induced Axonal Injury. *J. Neurol. Neurosurg. Psychiatry* 64, 285–287. doi:10.1136/jnnp.64.3.285

Hemphill, M. A., Dabiri, B. E., Gabriele, S., Kerscher, L., Franck, C., Goss, J. A., et al. (2011). A Possible Role for Integrin Signaling in Diffuse Axonal Injury. *PLOS ONE* 6, e22899–11. doi:10.1371/journal.pone.0022899

Hill, C. S., Coleman, M. P., and Menon, D. K. (2016). Traumatic Axonal Injury: Mechanisms and Translational Opportunities. *Trends Neurosciences* 39, 311–324. doi:10.1016/j.tins.2016.03.002

Hu, H., and Jonas, P. (2014). A Supercritical Density of Na+ Channels Ensures Fast Signaling in GABAergic Interneuron Axons. *Nat. Neurosci.* 17, 686–693. doi:10.1038/nn.3678

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual Molecular Dynamics. *J. Mol. Graphics* 14, 33–38. doi:10.1016/0263-7855(96)00018-5

Ingólfsson, H. I., Melo, M. N., Van Eerden, F. J., Arnarez, C., Lopez, C. A., Wassenaar, T. A., et al. (2014). Lipid Organization of the Plasma Membrane. *J. Am. Chem. Soc.* 136, 14554–14559. doi:10.1021/ja507832e

Inouye, H., and Kirschner, D. A. (1988b). Membrane Interactions in Nerve Myelin: II. Determination of Surface Charge from Biochemical Data. *Biophysical J.* 53, 247–260. doi:10.1016/S0006-3495(88)83086-8

Inouye, H., and Kirschner, D. A. (1988a). Membrane Interactions in Nerve Myelin. I. Determination of Surface Charge from Effects of pH and Ionic

Strength on Period. *Biophysical J.* 53, 235–245. doi:10.1016/S0006-3495(88)83085-6

Iwata, A. (2004). Traumatic Axonal Injury Induces Proteolytic Cleavage of the Voltage-Gated Sodium Channels Modulated by Tetrodotoxin and Protease Inhibitors. *J. Neurosci.* 24, 4605–4613. doi:10.1523/jneurosci.0515-03.2004

Johnson, V. E., Stewart, W., and Smith, D. H. (2013). Axonal Pathology in Traumatic Brain Injury. *Exp. Neurol.* 246, 35–43. doi:10.1016/j.expneurol.2012.01.013

Kirschner, D. A., Inouye, H., Ganser, A. L., and Mann, V. (1989). Myelin Membrane Structure and Composition Correlated: A Phylogenetic Study. *J. Neurochem.* 53, 1599–1609. doi:10.1111/j.1471-4159.1989.tb08558.x

LaPlaca, M. C., and Thibault, L. E. (1997). Anin Vitro Traumatic Injury Model to Examine the Response of Neurons to a Hydrodynamically-Induced Deformation. *Ann. Biomed. Eng.* 25, 665–677. doi:10.1007/BF02684844

Lukat, G., Krüger, J., and Sommer, B. (2013). APL@voro: A Voronoi-Based Membrane Analysis Tool for GROMACS Trajectories. *J. Chem. Inf. Model.* 53, 2908–2925. doi:10.1021/ci400172g

Malo, M. S., Blanchard, B. J., Andresen, J. M., Srivastava, K., Chen, X.-N., Li, X., et al. (1994). Localization of a Putative Human Brain Sodium Channel Gene (SCN1A) to Chromosome Band 2q24. *Cytogenet. Cel Genet* 67, 178–186. doi:10.1159/000133818

Manzoli, F. A., Stefoni, S., Manzoli-Guidotti, L., and Barbieri, M. (1970). The Fatty Acids of Myelin Phospholipids. *FEBS Lett.* 10, 317–320. doi:10.1016/0014-5793(70)80462-8

Marrink, S. J., de Vries, A. H., and Mark, A. E. (2004). Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B* 108, 750–760. doi:10.1021/jp036508g

Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., and De Vries, A. H. (2007). The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* 111, 7812–7824. doi:10.1021/jp071097f

Maxwell, W. (2013). Damage to Myelin and Oligodendrocytes: a Role in Chronic Outcomes Following Traumatic Brain Injury? *Brain Sci.* 3, 1374–1394. doi:10.3390/brainsci3031374

Menon, D. K., Schwab, K., Wright, D. W., and Maas, A. I. (2010). Position Statement: Definition of Traumatic Brain Injury. *Arch. Phys. Med. Rehabil.* 91, 1637–1640. doi:10.1016/j.apmr.2010.05.017

Mierzwa, A. J., Marion, C. M., Sullivan, G. M., McDaniel, D. P., and Armstrong, R. C. (2015). Components of Myelin Damage and Repair in the Progression of White Matter Pathology after Mild Traumatic Brain Injury. *J. Neuropathol. Exp. Neurol.* 74, 218–232. doi:10.1097/nen.0000000000000165

Min, Y., Kristiansen, K., Boggs, J. M., Husted, C., Zasadzinski, J. a., and Israelachvili, J. (2009). Interaction Forces and Adhesion of Supported Myelin Lipid Bilayers Modulated by Myelin Basic Protein. *Proc. Natl. Acad. Sci.* 106, 3154–3159. doi:10.1073/pnas.0813110106

Montanino, A., and Kleiven, S. (2018). Utilizing a Structural Mechanics Approach to Assess the Primary Effects of Injury Loads onto the Axon and its Components. *Front. Neurol.* 9, 643. doi:10.3389/fneur.2018.00643

Montanino, A., Saeedimasine, M., Villa, A., and Kleiven, S. (2020). Localized Axolemma Deformations Suggest Mechanoporation as Axonal Injury Trigger. *Front. Neurol.* 11, 25. doi:10.3389/fneur.2020.00025

Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., and Marrink, S.-J. (2008). The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theor. Comput.* 4, 819–834. doi:10.1021/ct700324x

Mu, J., Li, M., Wang, T., Li, X., Bai, M., Zhang, G., et al. (2019). Myelin Damage in Diffuse Axonal Injury. *Front. Neurosci.* 13, 217. doi:10.3389/fnins.2019.00217

Nakadate, H., Kurtoglu, E., Furukawa, H., Oikawa, S., Aomura, S., Kakuta, A., et al. (2017). "Strain-rate Dependency of Axonal Tolerance for Uniaxial Stretching," in 61st Stapp Car Crash Conference, November 8–10, 2021, Denver, CO (The Stapp Association). doi:10.4271/2017-22-0003

Needham, D., and Nunn, R. S. (1990). Elastic Deformation and Failure of Lipid Bilayer Membranes Containing Cholesterol. *Biophysical J.* 58, 997–1009. doi:10.1016/S0006-3495(90)82444-9

O'Brien, J. S., and Rouser, G. (1964). The Fatty Acid Composition of Brain Sphingolipids: Sphingomyelin, Ceramide, Cerebroside, and Cerebroside Sulfate. *J. Lipid Res.* 5, 339–342.

Páll, S., and Hess, B. (2013). A Flexible Algorithm for Calculating Pair Interactions on SIMD Architectures. *Computer Phys. Commun.* 184, 2641–2650. doi:10.1016/j.cpc.2013.06.003

Pandit, S. A., Bostick, D., and Berkowitz, M. L. (2003). An Algorithm to Describe Molecular Scale Rugged Surfaces and its Application to the Study of a Water/lipid Bilayer Interface. *J. Chem. Phys.* 119, 2199–2205. doi:10.1063/1.1582833

Pandit, S. A., Vasudevan, S., Chiu, S. W., Jay Mashl, R., Jakobsson, E., and Scott, H. L. (2004). Sphingomyelin-cholesterol Domains in Phospholipid Membranes: Atomistic Simulation. *Biophysical J.* 87, 1092–1100. doi:10.1529/biophysj.104.041939

Parrinello, M., and Rahman, A. (1981). Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* 52, 7182–7190. doi:10.1063/1.328693

Pastor, R. W., and MacKerell, A. D. (2011). Development of the CHARMM Force Field for Lipids. *J. Phys. Chem. Lett.* 2, 1526–1532. doi:10.1021/jz200167q

Pettus, E. H., and Povlishock, J. T. (1996). Characterization of a Distinct Set of Intra-axonal Ultrastructural Changes Associated with Traumatically Induced Alteration in Axolemmal Permeability. *Brain Res.* 722, 1–11. doi:10.1016/0006-8993(96)00113-8

Rasband, M. N., and Macklin, W. B. (2012). "Myelin Structure and Biochemistry," in *Basic Neurochemistry*. Editors S. T. Brady, G. J. Siegel, R. W. Albers, and D. L. B. T. Price Eighth Edition (New York: Academic Press), 180, 199. doi:10.1016/B978-0-12-374947-5.00010-9

Rawicz, W., Olbrich, K. C., McIntosh, T., Needham, D., and Evans, E. (2000). Effect of Chain Length and Unsaturation on Elasticity of Lipid Bilayers. *Biophysical J.* 79, 328–339. doi:10.1016/S0006-3495(00)76295-3

Saeedimasine, M., Montanino, A., Kleiven, S., and Villa, A. (2019). Role of Lipid Composition on the Structural and Mechanical Features of Axonal Membranes: a Molecular Simulation Study. *Sci. Rep.* 9, 8000. doi:10.1038/s41598-019-44318-9

Shen, H., Zhou, Q., Pan, X., Li, Z., Wu, J., and Yan, N. (2017). Structure of a Eukaryotic Voltage-Gated Sodium Channel at Near-Atomic Resolution. *Science* 355, eaal4326. doi:10.1126/science.aal4326

Singh, S., Pelegri, A. A., and Shreiber, D. I. (2015). Characterization of the Three-Dimensional Kinematic Behavior of Axons in Central Nervous System White Matter. *Biomech. Model. Mechanobiol* 14, 1303–1315. doi:10.1007/s10237-015-0675-z

Tang-Schomer, M. D., Johnson, V. E., Baas, P. W., Stewart, W., and Smith, D. H. (2012). Partial Interruption of Axonal Transport Due to Microtubule Breakage Accounts for the Formation of Periodic Varicosities after Traumatic Axonal Injury. *Exp. Neurol.* 233, 364–372. doi:10.1016/j.expneurol.2011.10.030

Tironi, I. G., Sperb, R., Smith, P. E., and van Gunsteren, W. F. (1995). A Generalized Reaction Field Method for Molecular Dynamics Simulations. *J. Chem. Phys.* 102, 5451–5459. doi:10.1063/1.469273

Wann, K. T. (1993). Neuronal Sodium and Potassium Channels: Structure and Function. *Br. J. Anaesth.* 71, 2–14. doi:10.1093/bja/71.1.2

Waugh, R., and Evans, E. A. (1979). Thermoelasticity of Red Blood Cell Membrane. *Biophysical J.* 26, 115–131. doi:10.1016/S0006-3495(79)85239-X

Westenbroek, R. E., Merrick, D. K., and Catterall, W. A. (1989). Differential Subcellular Localization of the Ri and Rii Na+ Channel Subtypes in Central Neurons. *Neuron* 3, 695–704. doi:10.1016/0896-6273(89)90238-9

Williams, K. A., Deber, C. M., and Klrschner, O. A. (1993). The Structure and Function of Central Nervous System Myelin. *Crit. Rev. Clin. Lab. Sci.* 30, 29–64. doi:10.3109/10408369309084665

Zambrano, F., Cellino, M., and Canessa-Fischer, M. (1971). The Molecular Organization of Nerve Membranes. *J. Membrain Biol.* 6, 289–303. doi:10.1007/BF02116575

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 © 2021 Saeedimasine, Montanino, Kleiven and Villa. 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 Molecular Biosciences

MINI REVIEW
published: 28 June 2021
doi: 10.3389/fmolb.2021.688700

Quantitative Modeling of Protein Synthesis Using Ribosome Profiling Data

Vandana Yadav[1†], *Inayat Ullah Irshad*[2†], *Hemant Kumar*[3] *and Ajeet K. Sharma*[2*]

[1]Department of Physics, Indian Institute of Technology Madras, Chennai, India, [2]Department of Physics, Indian Institute of Technology Jammu, Jammu, India, [3]School of Basic Sciences, Indian Institute of Technology Bhubaneswar, Bhubaneswar, India

Quantitative prediction on protein synthesis requires accurate translation initiation and codon translation rates. Ribosome profiling data, which provide steady-state distribution of relative ribosome occupancies along a transcript, can be used to extract these rate parameters. Various methods have been developed in the past few years to measure translation-initiation and codon translation rates from ribosome profiling data. In the review, we provide a detailed analysis of the key methods employed to extract the translation rate parameters from ribosome profiling data. We further discuss how these approaches were used to decipher the role of various structural and sequence-based features of mRNA molecules in the regulation of gene expression. The utilization of these accurate rate parameters in computational modeling of protein synthesis may provide new insights into the kinetic control of the process of gene expression.

Keywords: ribosome profiling, ribo-seq, translation-initiation rate, codon translation rate, codon usage bias, quantitative modeling of protein synthesis

OPEN ACCESS

Edited by:
Fabio Trovato,
Freie Universität Berlin, Germany

Reviewed by:
Alexander Schmidt,
University of Basel, Switzerland
Cédric Gobet,
École Polytechnique Fédérale de Lausanne, Switzerland

***Correspondence:**
Ajeet K. Sharma
ajeet.sharma@iitjammu.ac.in

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *31 March 2021*
Accepted: *25 May 2021*
Published: *28 June 2021*

Citation:
Yadav V, Ullah Irshad I, Kumar H and Sharma AK (2021) Quantitative Modeling of Protein Synthesis Using Ribosome Profiling Data. Front. Mol. Biosci. 8:688700. doi: 10.3389/fmolb.2021.688700

1 INTRODUCTION

Protein molecules carry out a vast array of biological functions. Indeed, almost every cellular process, from genome regulation to energy metabolism, requires a unique set of proteins with their precise concentration in a cell (Berg et al., 2002; Miyazaki and Esser, 2009). Therefore, the abundance of proteins in a cell is tightly regulated by various mechanisms acting at the level of transcription and translation (Curtis et al., 1995; Wu and Belasco, 2008). Understanding this regulation of protein synthesis remains one of the active areas of research from the last few decades (Merrick, 1992; Proud, 2006; Dever et al., 2016; Kummer and Ban, 2021). It was previously believed that cellular protein levels are primarily determined by mRNA copy number (Greenbaum et al., 2003; Lu et al., 2007). However, recent studies demonstrated that the translational regulation of protein synthesis also contributes significantly to *in vivo* protein abundance (Vogel et al., 2010; Li et al., 2014). Translational regulation is achieved by various structural and sequence-based features of mRNA molecules that can kinetically control the rate of protein synthesis (Kudla et al., 2009; Vogel et al., 2010; Tuller et al., 2010). Therefore, the knowledge of the rates at which different steps of translation occur and their connection with those mRNA features would provide key information concerning how the expression of an individual gene is regulated.

Single-molecule experiments provided significant insights into the details of the process of protein synthesis (Munro et al., 2008; Volkov and Johansson, 2018). However, little was known on how different mRNA features can control protein synthesis until the Next Generation Sequencing (NGS) experiments started uncovering various kinetic properties of this process (Ingolia et al., 2019). Several methods have been developed in the last decade that allows the accurate extraction of translation rate

parameters using ribosome profiling data (Weinberg et al., 2016; Duc and Song, 2018; Sharma et al., 2019). The analysis of these translation rate parameters and their use in protein synthesis simulations have started unraveling the sophisticated mechanism that nature has developed to optimize the use of cellular resources (Tuller et al., 2010; Diament et al., 2018; Lyu et al., 2020).

In this mini-review, we provide the brief overview of a few recently developed methods that extract translation rate parameters from ribosome profiling data. We explain the technical details and assumptions made in these approaches, and also comment on their accuracy in several different contexts. We also highlight some of the recent results that use ribosome profiling data to provide new insights into the kinetic control of protein synthesis. This mini-review aims to promote the use of big biological data sets among biophysicists, biophysical chemists and system biologist to achieve greater accuracy and reliability in the quantitative modeling of protein synthesis.

2 QUANTITATIVE MODELING OF PROTEIN SYNTHESIS

The first mathematical model of protein synthesis was developed by MacDonald et al. (1968), and since then various similar models and their extensions were proposed to understand the different aspects of mRNA translation (Garai et al., 2009; Sharma and Chowdhury, 2012; Ciandrini et al., 2013; Margaliot and Tuller, 2013). Among these, the totally asymmetric simple exclusion process (TASEP) is a model which incorporates key steps regulating *in vivo* protein production and is extensively used to stimulate protein synthesis. In this model, an mRNA is considered as a one-dimensional lattice where each site in that lattice represents a single codon. A ribosome in this model is like an extended particle that covers ten consecutive codon positions of a transcript where its location is usually identified by the position of its A-site (Sharma and O'Brien, 2017; Ingolia et al., 2009). The ribosome A-site is located at the sixth codon from the $5'$ end of the ribosome. Therefore, a ribosome at the j^{th} position covers $j-5$ to $j+4$ codons (Ahmed et al., 2019). Protein synthesis in this model is divided into three sub-steps: initiation, elongation and termination (Merrick, 1992) (**Figure 1A**). Initiation occurs when a ribosome assembles at the start codon of the i^{th} transcript with rate $\alpha(i)$ (Kozak, 1999; Merrick and Pavitt, 2018). A ribosome initiates protein synthesis with its A-site at the second codon position of the transcript, therefore the translation initiation occurs when the first six codons are not occupied by another ribosome. The limited availability of ribosomes in a cell makes translation-initiation a rate-limiting step of the protein synthesis (Shah et al., 2013). The ribosome then starts taking a series of stochastic steps toward the stop codon. On the i^{th} transcript, a ribosome slides from codon position j to $j+1$ with rate $\omega(j,i)$**.** In each of such steps, a ribosome selects cognate aa-tRNA molecule, forms a peptide bond and then moves to the next codon position **(**Sharma and Chowdhury, 2011; Sharma and Chowdhury, 2012**)** Note that multiple ribosomes simultaneously move on a single transcript where each of them synthesizes a separate copy of protein. Therefore, due to mutual exclusion, a ribosome cannot move to the next codon if its passage is blocked by another downstream ribosome. After arriving at the stop codon, the ribosome terminates the process, and releases a fully synthesize protein with rate $\beta(i)$. The termination of protein synthesis also leads to the disassembly of the ribosome at the stop codon (Hellen, 2018).

The TASEP with an uniform elongation rate, unitary particle size, and infinite lattice size is exactly solved by Derrida et al. (1992). Later, Lazarescu and Mallick (2011) extended this work by solving the TASEP with finite lattice size. Adding to this, Kolomeisky (1998) provided an analytical solution for TASEP with local inhomogeneities. For extended particles, Shaw et al. (2004) solved the TASEP model under mean-field approximation for uniform elongation rate. Recently, Szavits-Nossan et al. (2018) provided an analytical solution for the TASEP with non-uniform elongation rate and extended particle size; however, it ignores the higher-order terms of power series solution of the model. Therefore, in the absence of any exact analytical solution, computer simulation of the TASEP model is the best possible approach for making reliable quantitative predictions on protein synthesis.

Multiple approaches have been used in the past to simulate protein synthtesis on TASEP model, including kinetic Monte-Carlo (Zia et al., 2011), next reaction method (Gibson and Bruck, 2000; Duc and Song, 2018), Gillespie's method (Gillespie, 1977), etc. The Gillespie's method is one of the very efficient approach for studying stochastic systems and is commonly used to simulate the TASEP model (Gillespie, 1977). Solving TASEP using this method requires to calculate the parameter

$$R = \sum_{j=1}^{L_i} T(j) \tag{1}$$

which is the sum total of rates for all transitions that lead to a new state from the current state of the translation system. (Note that every unique arrangement of ribosomes on a transcript is a separate state.) In **Eq. 1**, $T(1) = \alpha(i)$, $T(j) = \omega(j,i)\delta(j)$ for $2 \le j < L_i - 1$ and $T(L_i) = \beta(i)\delta(L_i)$; $\delta(j)$ is one when a ribosome occupies the j^{th} codon position of the transcript, otherwise it is zero, and L_i is the total number of codons in the i^{th} transcript. The TASEP model assumes that all transitions in translation system are Markovian therefore, the dwell time in a given state is exponentially distributed with a mean value of $\frac{1}{R}$. This quantity is calculated by generating an exponentially distributed random number $\tau = \frac{1}{R} ln(r_1)$, where r_1 is a random number that is uniformly distributed between 0 and 1. The next transition is randomly chosen according to the relative contributions of all possible transitions in R (**Eq. 1**). For this, another random number r_2, which is uniformly distributed between 0 and R, is generated, and the next transition is identified according to the selection criteria given in **Table 1** (A flow-chart explaining the various steps of protein synthesis simulations is shown in **Figure 1B**). Repeating this procedure generates a trajectory of protein synthesis on a given mRNA transcript, which can be used to calculate various quantities: rate of protein synthesis, average and codon-specific ribosome density, in-silico ribosome profiles, etc.

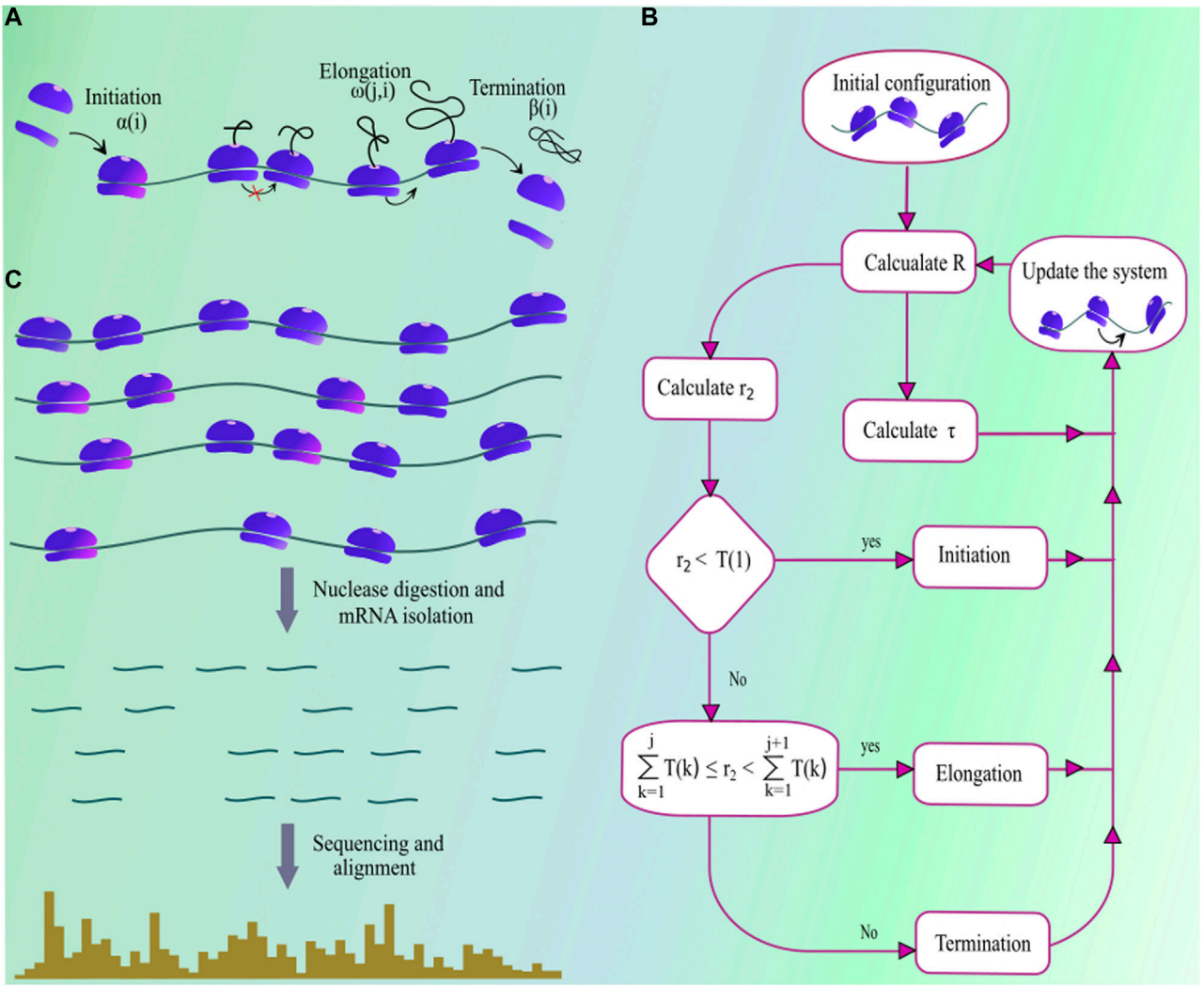


FIGURE 1 | (A) A pictorial illustration of the steps involved in the process of protein synthesis. Ribosome subunits assemble at the start codon with rate $\alpha(i)$ when no ribosome occupies the first six codons of the transcript. Then, the ribosome starts moving toward the stop codon by a single codon at a time. It hops from codon j to $j+1$ with rate $\omega(j,i)$. Note well, a ribosome cannot move to the next codon if it is occupied by another ribosome. After arriving at the stop codon, the ribosome terminates protein synthesis and releases fully synthesized protein with rate $\beta(i)$ **(B)** The flow-chart describes the computational algorithm of protein synthesis simulations using the Gillespie's method. For details, see the main text and **table (1) (C)** explains various experimental and computational steps involved in preparing ribosome profiles.

3 METHODS OF EXTRACTING TRANSLATION RATE PARAMETERS

Accurate quantitative predictions for protein production critically depends on having precise estimates of translational rate parameters. Therefore, in this section, we discuss common approaches to extract translation rate parameters from ribosome profiles which provides the access to a plethora of information about the protein synthesis process. In ribosome profiling experiment, translation elongation is arrested by the treatment with drug cycloheximide (**Figure 1C**). Then, after the cell lysis, the portions of the mRNA molecule that are not protected by ribosomes are digested by nucleases. The remaining ribosome-protected mRNA fragments are subsequently sequenced and aligned to a reference genome. The typical length of a ribosome-protected mRNA fragment is 28–31 nucleotides; therefore, fragments outside this range are excluded from the analysis (Pop et al., 2014). Note that a ribosome translates the codon present at its A-site. Therefore, the position of A-site is identified on those ribosome-protected mRNA fragments (Ahmed et al., 2019), referred to as the read aligned to the A-site codon. Within a transcript, more ribosome-reads aligned to a codon means a longer translation time for that particular codon. This steady-state profile of ribosome occupancy can be used in extracting translation initiation and codon translation rates. Many computational methods have been developed in the past to extract these rate parameters from

TABLE 1 | A list of the types of transitions and conditions in simulation algorithm.

Conditions	Transitions
$r_2 < T(1)$, and the first six codon positions of the transcripts are unoccupied	Translation-initiation
$\sum_{k=1}^{j} T(k) \le r_2 < \sum_{k=1}^{j+1} T(k)$, a ribosome is at the j^{th} codon position with no ribosome at $(j+10)$	Transition of the ribosome from codon position j to $j+1$
$\sum_{k=1}^{L_i-1} T(k) \le r_2 < \sum_{k=1}^{L_i} T(k)$, a ribosome is at the L_i^{th} codon	Translation-termination

ribosome profiling data. We categorize these methods into three different groups.

3.1 Optimization Based Methods

These methods take advantage of some known characteristics of the translation process and extract codon translation rates by optimizing or fitting a function of ribosome profiling reads. For example, Dana and Tuller (2014) proposed a method for extracting the "typical" translation time of a codon by fitting a function that characterizes ribosome dwell time distribution. To do that, Dana and Tuller (2014) first normalize the ribosome footprint counts by the average reads aligned to the transcript. That is,

$$\hat{r}(j,i) = \frac{r(j,i)}{\frac{1}{L_i-40}\sum_{j=21}^{L_i-20} r(j,i)} \tag{2}$$

$r(j,i)$ in **Eq. 2** is the number of ribosome profiling reads aligned to the j^{th} codon of the i^{th} transcript. To minimize any sampling error, the authors recommend excluding the genes with median read counts less than one. In addition to that, the first and the last 20 codons were removed from the analysis as unusually high ribosome footprint density was found in those regions (Ingolia et al., 2009). Then, the distribution of normalized footprint count for a codon type "c"(*e.g.*, CUU, AUG) is computed using the data collected from the whole transcriptome. It was found that this distribution is a superposition of the normal and exponential distributions (Dana and Tuller 2014; Dana and Tuller , 2015). Authors rationalize this observation by proposing that the normal distribution characterizes the typical decoding time for a codon type (τ_1) whereas the exponential distribution reflects the delay (τ_2) caused by rare translation pauses and ribosomal interference. Therefore, the distribution of codon translation time (τ), a sum of random variables τ_1 and τ_2, is the convolution of the normal and exponential distribution, and has the following functional form.

$$f(\tau, \mu_c, \sigma_c, \lambda_c) = \frac{\lambda_c}{2} e^{\frac{\lambda_c}{2}\left(2\mu_c + \lambda_c\sigma_c^2 - 2\tau\right)} \operatorname{erf}\left[\frac{\mu_c + \lambda_c\sigma_c^2 - \tau}{\sqrt{2}\sigma_c}\right] \tag{3}$$

In **Eq. 3**, μ_c and σ_c are the mean and standard deviation of the normal distribution, respectively. λ_c is the coefficient for exponential distribution whose inverse is the average time delay caused by rare translation pauses and ribosomal interference. The distribution in **Eq. 3** was then fitted with normalized read distribution of codon "c", providing μ_c, the "typical" codon translation time. Using this analysis, the authors calculate the 'typical' decoding time for all 61 sense codons.

Pop et al. (2014) proposed a similar method that calculates gene-specific and globally averaged translation time of a codon type "c" (*i.e.*, μ_c^i and μ_c, respectively) by maximizing the following objective function.

$$max(\mu_c^i, \mu_c)\left[log\prod_{i,j}\mu_c^{i\left(r(j,i)/J_i\right)} e^{-\mu_c^i} - \sum_{i,c} w_c^i \left(log\mu_c^i - log\mu_c\right)^2\right] \tag{4}$$

The first term in the objective function represents the likelihood of observed ribosome profiles for a specific μ_c^i whereas the second term minimizes the difference between the global and gene-specific average translation time. w_c^i in **Eq. 4** is a ratio of the total number of "c" codons in the i^{th} transcript to all transcripts. Therefore, the genes with more "c" codons will have greater weightage in the objective function. J_i is the ribosome flux at transcript i which was fixed to $\sum_{j \in i} \frac{r(j,i)}{L_i}$. Then, L-BFGS algorithm (Byrd et al., 1995) is applied to search the numerical values of μ_c^i and μ_c that maximizes the objective function in **Eq. 4**. The inverse of the μ_c is the translation rate of codon "c".

In another optimization based approach, Szavits-Nossan and Ciandrini (2020) have used the non equilibrium analysis of ribosome profiling data to infer the ratio of elongation to initiation rate *i.e.* $k(j,i) = \frac{\omega(j,i)}{\alpha(i)}$. In this method, $k(j,i)$s were calculated by minimizing the difference between experimentally measured and numerically computed codon-specific ribosome density of a transcript. For this, an objective function

$$S = \sum_{j=2}^{L_i} \left(\rho^{ana}\{k(j,i)\} - \rho(j,i)\right)^2 \tag{5}$$

is minimised by using the Least-Squared optimization technique. $\rho^{ana}\{k(j,i)\}$ in **Eq. 5** is numerically computed by using an expression derived by Szavits-Nossan et al. (2018) whereas $\rho(j,i)$s were calculated by distributing the experimentally measured polysome density according to the distribution of ribosome profiling reads along a transcript. Optimizing the objective function in **Eq. 5** yields the numerical values of $k(j,i)$. Unlike Pop et al. (2014) and Dana and Tuller (2014), this method does not put any constraint on the variations in the translation rate of a codon type. Therefore, the normalized rates obtained from this method can precisely capture the local variation in codon translation rates along a transcript.

3.2 Simulation Based Methods

Simulation based methods follow an iterative procedure, where in each iteration, translation rate parameters are altered until the

simulation output converges to its experimentally measured counterparts. For example, Gritsenko et al. (2015) compared the experimentally measured and simulated ribosome densities on mRNA segments of different lengths. This comparison on short mRNA segments allows capturing the local variation in codon translation rates whereas in long segments it increase the reliability of measured rate parameters. These mRNA segments were constructed by dividing the whole transcript into two parts in such a way that both of them get an equal number of combined ribosome profiling and RNA-seq reads. The daughter segments were further divided recursively using the same approach until a reliable estimate of ribosome density can be made.

To implement this method, Gritsenko et al. (2015) used an experimental observation that ribosome density on a mRNA segment is distributed log-normally among all its replicates. This means the probability of finding a specific value of ribosome density in a single observation can be expressed as

$$P\left(C, N_k^i \middle| r_k^i, \sigma_k^i\right) \propto \frac{1}{N_k^i\sqrt{2\pi}} e^{\frac{lnN_k^i + lnC - r_k^i}{2\left(\sigma_k^i\right)^2}} \tag{6}$$

r_k^i and σ_k^i in **Eq. 6** are the mean and standard deviation of ribosome density for the k^{th} segment of transcript i; N_k^i is the number of ribosomes on the same mRNA segment which was observed in a single snapshot of protein synthesis simulations. The experimental ribosome density was calculated by taking the ratio of the normalized ribosome profiling reads with the RNA-Seq reads aligned to the same segment. Since this quantity is measured in arbitrary units, a parameter C was introduced to scale it to the simulated ribosome density. r_k^i in **Eq. 6** was calculated by taking the mean of ribosome density from the data collected from all replicates of the same experiment. The shape parameter σ_k^i was calculated for a group of segments with the same length as it was not possible to reliably calculate σ_k^i from a very small number of replicates. Then, using **Eq. 6**, Gritsenko et al. (2015) define an objective function Ψ that quantifies how well the simulation model predicts experimentally measured ribosome densities.

$$\Psi = \sum_i \sum_k \left[-\frac{1}{2\sigma_k^i}\left(lnN_k^i - r_k^i + lnC\right)^2 - lnN_k^i \right] \tag{7}$$

Gritsenko et al. (2015) carry out the transcriptome-wide protein synthesis simulations by supplying some initial translation rate parameters to the model. Snapshots taken from protein synthesis simulations produced N_k^i for each mRNA segment which were used to compute Ψ in the model evaluation step. Then, using the numerical value of Ψ, a genetic algorithm (Covariance Matrix Adaptation Evolutionary Strategy) proposes new initiation and codon translation rates, which were further used to simulate protein synthesis on the whole transcriptome. This process is repeated until Ψ is maximized which produces the translation initiation and codon translation rates.

In an another simulation based study, Duc and Song (2018) measured translation rate parameters by comparing the normalized *in vivo* ribosome profiling reads (**Eq. 2**) with those obtained from simulations. This method also requires initial translation rate parameters to generate in-silico ribosome profiles which will be refined in every iteration of the method. Initial codon translation rates were estimated as follows.

$$\omega(j,i) = \begin{cases} min\left(\omega_{max}, \dfrac{\sum_{j=1}^{L_i} r(j,i)}{r(j,i)}\right), & \text{if } r(j,i) \neq 0 \\ \omega_{max}, & \text{otherwise} \end{cases} \tag{8}$$

ω_{max} here is a crude guess of the maximum codon translation rate in *S. cerevisiae*. The initial translation initiation rates were calculated by following a method developed by Ciandrini et al. (2013). In that method, initiation rate of a transcript is varied until the simulated average ribosome density matches with what has been measured in polysome profiling experiments. These initial translation rate parameters were then used to simulate protein synthesis and generate *in silico* ribosome profiles. In each step, simulated and *in vivo* ribosome profiles were compared and error ε is computed for each codon of a transcript.

$$\varepsilon(j) = \left| \frac{r(j,i)}{\sum_j r(j,i)} - \frac{r^{sim}(j,i)}{\sum_j r^{sim}(j,i)} \right| \tag{9}$$

Authors also defined codon positions with significant error where it was larger than $\frac{10\sum_j \varepsilon(j)}{L_i}$. It is very likely that ribosome profiling reads at such positions may have been influenced by extensive ribosome traffic-jams which increases the dwell time of downstream ribosomes. Therefore, for such cases, translation rate of neighboring codons were also updated in the next iteration cycle. Authors used the following updating rules in each iteration of the method.

$$\omega(j,i) = \begin{cases} \lambda_1 \omega(j,i) & \text{for codon } j \text{ where } \varepsilon(j) < \dfrac{10\sum_j \varepsilon(j)}{L_i} \\ \lambda_2 \omega(j,i), & \text{for codons } j-30 \text{ to } j-10 \text{ when } \varepsilon(j) > \dfrac{10\sum_j \varepsilon(j)}{L_i} \\ \omega(j,i) & \text{otherwise} \end{cases} \tag{10}$$

λ_1 and λ_2 are chosen using the golden section search algorithm. After updating the codon translation rates, a transcriptome-wide simulations of protein synthesis were carried out and then the same procedure is repeated until no error sites were detected. This provides the translation initiation and codon translation rates that generates *in silico* ribosome profiles similar to the *in vivo* profiles.

3.3 Chemical Kinetic Based Methods

Chemical kinetic-based methods do not require extensive simulations of protein synthesis. Instead, they rely on analytical expressions of translation rate parameters that use ribosome profiling and RNA-Seq data as input variables. A recent publication used the mean-field and steady-state assumptions, and derived the following analytical expressions for translation-initiation and codon translation rates (Sharma et al., 2019).

$$\alpha(i) = \frac{\langle \rho(i) \rangle (L_i - 1)}{\langle TT(i) \rangle \left[1 - \sum_{k=2}^{11} \rho(k,i)\right]} \tag{11}$$

$$\omega(j,i) = \frac{\sum_{j=2}^{L_i-1}\rho(j,i)}{\rho(j,i)\langle TT(i)\rangle} \quad (12)$$

In **Eq. 11**, **Eq. 12** $\langle TT(i)\rangle$ is the time a ribosome takes to move from the start to stop codon whereas $\rho(i)$ is the average ribosome density on the transcript. $\langle TT(i)\rangle$ can be calculated by using a scaling relation between gene translation time and the number of codons in a transcript (Sharma et al., 2018) if gene-specific reliable estimates of protein synthesis times are not available. The average ribosome density on a transcript is proportional to the ratio of the number of ribosome profiling and RNA-Seq reads aligned to that transcript. This proportionality was used to estimate the average ribosome density on a transcript (Sharma et al., 2019). $\rho(j,i)$s were calculated by distributing the $\rho(i)$s according to the variations in ribosome profiling reads across the transcript *i*. This new analysis method neither relies on heuristic and ad-hoc approaches nor requires extensive protein synthesis simulations, and implementing this equation-based method is much easier than others.

4 STATISTICAL NOISE AND SEQUENCE BIASES IN MEASURED TRANSLATION RATE PARAMETERS

Ribosome profiles provide single codon resolution to the protein synthesis process. However, these data sets are very noisy and are subjected to numerous biases associated with various steps of the ribosome profiling experiment, including the amplification of ribosome footprints by RT-PCR, nuclease digestion, cell lysis, etc. (O'Connor et al., 2016; Mohammad et al., 2019; Xiao et al., 2016; Hussmann et al., 2015). Such statistical errors and biases will also be reflected in the measured translation rate parameters. Dana and Tuller (2014) minimize their impact by ignoring any variation in the translation rate of a codon type. This approach drastically reduces the total number of parameters to be extracted from the ribosome profiling data. It minimizes the statistical uncertainty in the measurement of codon translation rates and also averages out various sequence biases. However, a major drawback of this approach is that it does not account for the context-dependent variations in codon translation rates. The other extreme approach taken by Szavits-Nossan and Ciandrini (2020) measures the translation rate for each codon in an mRNA transcript but such measurements are subjected to a higher degree of stochastic noise. A few probabilistic and machine learning models have also been applied to minimize the effect of noise and biases in the identification of A-site position on ribosome footprints (Fang et al., 2018; Tunney et al., 2018; Gobet et al., 2020). These models have successfully captured the context-dependent variation in codon translation rates and also performed well in transcripts with low abundance (Liu et al., 2020; Michel et al., 2016).

5 MOLECULAR DETERMINANTS OF TRANSLATION RATE PARAMETERS

A closer look at the measured translation rate parameters unraveled the molecular determinants of translation-initiation and codon translation rates (Sharma et al., 2019; Duc et al., 2018). For example, tRNA pool hypothesis proposed codon translation rates to be proportional to the availability of cognate tRNA molecules (Ikemura, 1981; Ikemura, 1985). However, this hypothesis was never explicitly tested as there was no method that allowed translation rate measurement for all codons. Codon translation rates measured by ribosome profiling experiments supported this hypothesis in *S. cerevisiae* and *E. coli* (Dana and Tuller, 2014; Sharma et al., 2019). However, no such behaviour was observed in mouse cell lines (Gobet et al., 2020; Ingolia et al., 2011). Interestingly, a strong correlation between codon usage and tRNA abundance is observed in mammalian cells (Gobet et al., 2020; Gobet and Naef, 2017; Neelagandan et al., 2020), suggesting that tRNA levels are tuned according to their requirement in a cell.

In addition, an analysis of measured codon translation rate has shown that mRNA structures downstream to the A-site codon increase their translation time (Sharma et al., 2019). The reason for this increase is that the ribosome has to first unfold the structure to proceed to the next codon (Qu et al., 2011). The average increase of 6.7% in codon translation time was reported due to the presence of a structure in mRNA molecule. However, depending upon the stability of that structure, it may vary from one codon to another. Similarly, the presence of proline amino acid on the ribosome P-site increases the median translation time of a codon by 19% (Sharma et al., 2019) because the stereochemistry of proline amino acid delays the peptide bond formation with adjacent amino acid (Pavlov et al., 2009). Furthermore, Duc and Song (2018) have discovered that the aqueous environment inside the ribosome exit tunnel leads to a faster translation of codons when hydrophobic amino acid residues are present inside the tunnel. Electrostatic charges on the nascent-protein have been shown to modulate the translation elongation rate (Duc and Song, 2018; Riba et al., 2019). The ribosome exit tunnel is negatively charged. Therefore, the presence of positively charged amino acid into the tunnel decreases the translation rate of downstream codons (**Figure 2A**). The identity of the amino acid at the ribosome P-site also affects the translation rate of the A-site codon. In *S. cerevisiae*, eighty six different pairs of amino acids at the A- and P-sites speed up the elongation rate whereas it is slowed down in the case of eighty one other pairs of amino acids (Ahmed et al., 2020). A similar behaviour was also observed in mouse liver cells where different codon combinations at A- and P-sites, E- and P-sites, and E- and A-sites can speed up or slow down the translation elongation (Gobet et al., 2020). Post-translational modifications of tRNA molecules also enhances the rate of translation elongation in *S. cerevisiae*, *N. crassa* and *C. elegans* (Lyu et al., 2020; Nedialkova and Leidel, 2015). In addition to these molecular factors, patterns of slow and fast codons can cause ribosome traffic-jams on a transcript, and can significantly affect the time a ribosome spends at a given codon position (Diament et al., 2018; Duc et al., 2018).

Translation initiation is another rate-limiting step (Shah et al., 2013) and sets an upper bound to the rate at which proteins are produced from a single transcript (Szavits-Nossan and Ciandrini, 2020). Initiation rates measured from ribosome profiling

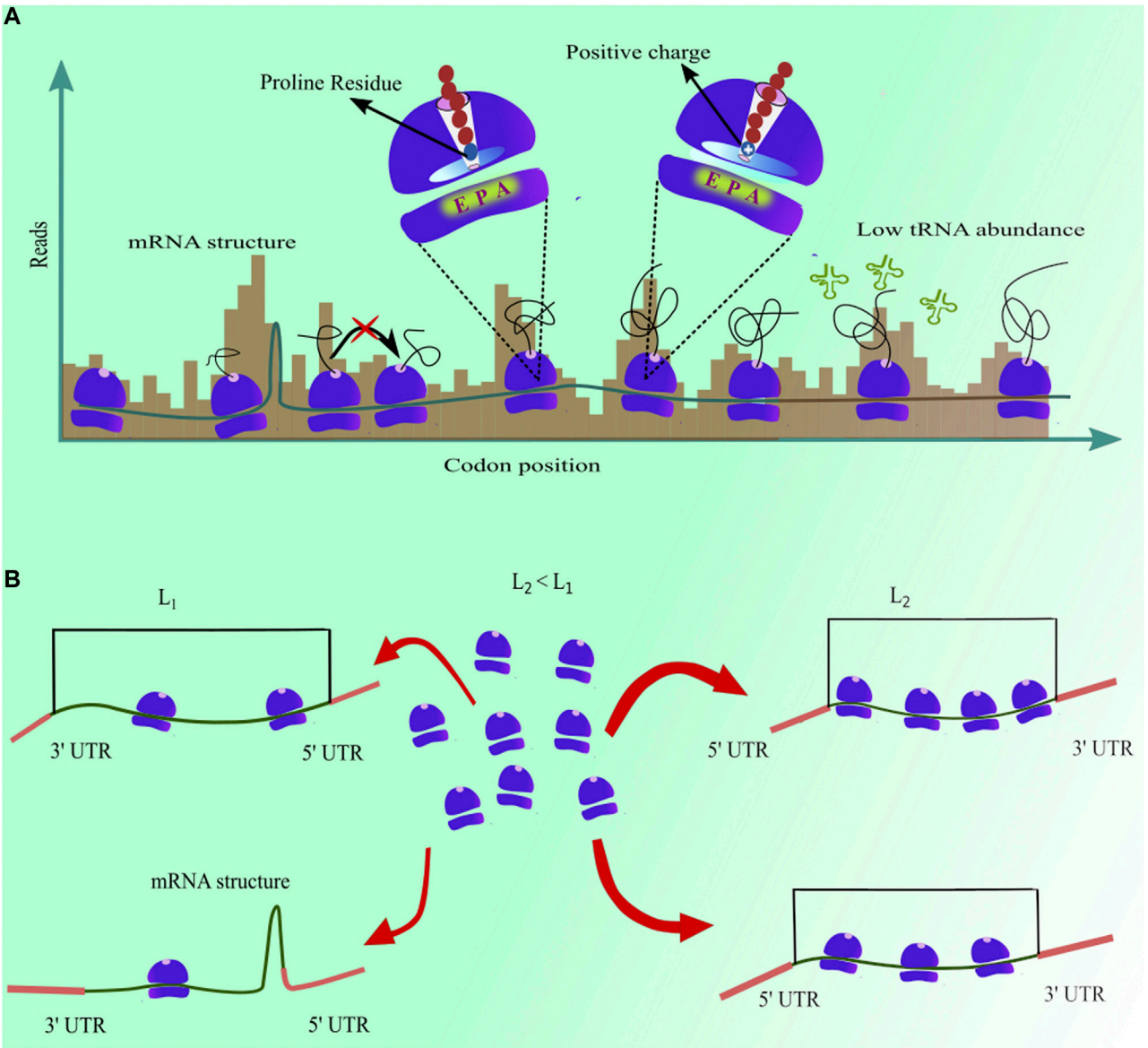


FIGURE 2 | (A) mRNA structure, low tRNA abundance, and the presence of proline and positively charged amino acids in ribosome exit tunnel increase ribosome dwell time, resulting in higher ribosome profiling reads for those codon positions **(B)** Presence of mRNA structure at the 5′ end and length of the coding sequence in a transcript anti-correlate with initiation rate.

experiments also identified its molecular determinants (**Figure 2B**). For example, significant negative correlations of initiation rate with the free energy of mRNA folding near the start codon and transcript length were observed in multiple studies (Weinberg et al., 2016; Duc et al., 2018; Sharma et al., 2019). A stable structure near the start codon makes this region inaccessible to a ribosome for initiating the process of protein synthesis, thus decreasing the translation initiation rate. Indeed, many mRNA sequence design algorithms minimize folding energy in this region to enhance the production heterologous proteins (Salis et al., 2009; Angov, 2011). Similarly, the length of the coding sequence of a transcript is inversely proportional to the translation initiation rate (Duc and Song, 2018; Sharma et al., 2019). It is because the ribosomes completing protein synthesis at the termination end can easily diffuse to the start codon in shorter mRNA transcripts (Fernandes et al., 2017). These faster initiation rates in shorter transcripts help them in producing more proteins. Moreover, the presence of AUG codons upstream to the start codon can interfere in the recruitment of ribosomes for translation initiation, thus resulting in a decrease in the translation initiation rate (Sharma et al., 2019). Furthermore, the presence of KOZAK sequence in *S. cerevisiae* transcripts leads to a faster initiation as it serves as a stable binding site for the small ribosome subunit (Sharma et al., 2019). These new findings

have demonstrated that a combination of several molecular factors work in tandem to finely modulate the translation-intiation and codon translation rates.

6 CONCLUDING REMARKS AND FUTURE DIRECTIONS

The development of ribosome profiling has allowed access to relative ribosome occupancy at single codon resolution (Ingolia et al., 2009; Ingolia et al., 2019). Many computational tools can convert this time-independent steady-state information into the kinetic rate parameters of protein synthesis (Dana and Tuller 2014; Pop et al., 2014; Szavits-Nossan and Ciandrini, 2020). Analysis of these rate parameters and their use in protein synthesis simulations give significant insight into the translational regulation of an individual gene (Shah et al., 2013; Lyu et al., 2020). These rate parameters also help identify various structural and sequence-based mRNA features that control the rate of protein synthesis (Weinberg et al., 2016; Duc et al., 2018; Sharma et al., 2019). Furthermore, the knowledge of these rates offers an unprecedented opportunity to explore and model other parallel and downstream processes influenced by translation-elongation kinetics, including co-translational protein folding, mRNA degradation, protein translocation through a membrane, chaperone binding, post-translational modifications, etc (Sharma and O'Brien, 2018; Radhakrishnan et al., 2016).

Analysis of measured translation rate parameters demonstrated that a combination of multiple molecular factors determines the translation-initiation and codon translation rates (Sharma et al., 2019). The strength with which these molecular factors act on translation rate parameters can vary from one place to another. Therefore, the same codon at two different locations can be translated at different rates. Current approaches can capture this context-dependent variation in codon translation rates (Sharma et al., 2019; Szavits-Nossan et al., 2018). However, none of them can quantify the impact of each molecular factor on translation-initiation and codon translation rates. Decoupling their effects would enable the scientists to make reliable predictions on translation rate parameters by only looking at mRNA sequence features, thus providing deeper insights into the context-dependent variation in codon translation rates.

Reliable predictions on protein synthesis, from a single transcript to the whole-cell level, are required for numerous synthetic biology applications (Purcell et al., 2013; Burke et al., 2020). For example, carefully placing the molecular determinants of translation rate parameters may help in designing heterologous genes and synthetic biological circuits. Moreover, in the absence of reliable gene expression models, synthetic biology relies heavily on the trial and error approach (Purcell et al., 2013; El Karoui et al., 2019). Therefore, the accurate predictions made by the quantitative models of protein synthesis will speed up the whole process of designing synthetic biology products with potential applications in areas such as drug delivery, cellular engineering, next-generation drugs, deployable medical devices, etc (Salis et al., 2009; Goldberg et al., 2018; Macklin et al., 2014).

AUTHOR CONTRIBUTIONS

AS conceptualised the project. VY, IR, HK, and AS wrote the manuscript.

REFERENCES

Ahmed, N., Sormanni, P., Ciryam, P., Vendruscolo, M., Dobson, C. M., and O'Brien, E. P. (2019). Identifying A- and P-Site Locations on Ribosome-Protected mRNA Fragments Using Integer Programming. *Sci. Rep.* 9, 6256. doi:10.1038/s41598-019-42348-x

Ahmed, N., Friedrich, U. A., Sormanni, P., Ciryam, P., Altman, N. S., Bukau, B., et al. (2020). Pairs of Amino Acids at the P- and A-Sites of the Ribosome Predictably and Causally Modulate Translation-Elongation Rates. *J. Mol. Biol.* 432, 166696. doi:10.1016/j.jmb.2020.10.030

Angov, E. (2011). Codon Usage: Nature's Roadmap to Expression and Folding of Proteins. *Biotechnol. J.* 6, 650–659. doi:10.1002/biot.201000332

Berg, J. M., Tymoczko, J. L., and Stryer, L. (2002). *Biochemistry*, 38. New York: 5thWH Freeman, 76.

Burke, P. E. P., Campos, C. B. d., Quiles, M. G., and Quiles, M. G. (2020). A Biochemical Network Modeling of a Whole-Cell. *Sci. Rep.* 10, 13303. doi:10.1038/s41598-020-70145-4

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.* 16, 1190–1208. doi:10.1137/0916069

Ciandrini, L., Stansfield, I., and Romano, M. C. (2013). Ribosome Traffic on Mrnas Maps to Gene Ontology: Genome-wide Quantification of Translation Initiation Rates and Polysome Size Regulation. *Plos Comput. Biol.* 9, e1002866. doi:10.1371/journal.pcbi.1002866

Curtis, D., Lehmann, R., and Zamore, P. D. (1995). Translational Regulation in Development. *Cell.* 81, 171–178. doi:10.1016/0092-8674(95)90325-9

Dana, A., and Tuller, T. (2014). The Effect of Trna Levels on Decoding Times of Mrna Codons. *Nucleic Acids Res.* 42, 9171–9181. doi:10.1093/nar/gku646

Dana, A., and Tuller, T. (2015). Mean of the Typical Decoding Rates: a New Translation Efficiency index Based on the Analysis of Ribosome Profiling Data. *G3: Genes, Genomes, Genet.* 5, 73–80. doi:10.1534/g3.114.015099

Dao Duc, K., Saleem, Z. H., and Song, Y. S. (2018). Theoretical Analysis of the Distribution of Isolated Particles in Totally Asymmetric Exclusion Processes: Application to Mrna Translation Rate Estimation. *Phys. Rev. E.* 97, 012106. doi:10.1103/PhysRevE.97.012106

Dao Duc, K., and Song, Y. S. (2018). The Impact of Ribosomal Interference, Codon Usage, and Exit Tunnel Interactions on Translation Elongation Rate Variation. *Plos Genet.* 14, e1007166. doi:10.1371/journal.pgen.1007166

Derrida, B., Domany, E., and Mukamel, D. (1992). An Exact Solution of a One-Dimensional Asymmetric Exclusion Model with Open Boundaries. *J. Stat. Phys.* 69, 667–687. doi:10.1007/bf01050430

Dever, T. E., Kinzy, T. G., and Pavitt, G. D. (2016). Mechanism and Regulation of Protein Synthesis in saccharomyces Cerevisiae. *Genetics*. 203, 65–107. doi:10.1534/genetics.115.186221

Diament, A., Feldman, A., Schochet, E., Kupiec, M., Arava, Y., and Tuller, T. (2018). The Extent of Ribosome Queuing in Budding Yeast. *Plos Comput. Biol.* 14, e1005951. doi:10.1371/journal.pcbi.1005951

El Karoui, M., Hoyos-Flight, M., and Fletcher, L. (2019). Future Trends in Synthetic Biology-A Report. *Front. Bioeng. Biotechnol.* 7, 175. doi:10.3389/fbioe.2019.00175

Fang, H., Huang, Y.-F., Radhakrishnan, A., Siepel, A., Lyon, G. J., and Schatz, M. C. (2018). Scikit-ribo Enables Accurate Estimation and Robust Modeling of

Translation Dynamics at Codon Resolution. *Cel Syst.* 6, 180–191. doi:10.1016/j.cels.2017.12.007

Fernandes, L. D., Moura, A. P. S., and Ciandrini, L. (2017). Gene Length as a Regulator for Ribosome Recruitment and Protein Synthesis: Theoretical Insights. *Sci. Rep.* 7, 17409–17411. doi:10.1038/s41598-017-17618-1

Garai, A., Chowdhury, D., Chowdhury, D., and Ramakrishnan, T. V. (2009). Stochastic Kinetics of Ribosomes: Single Motor Properties and Collective Behavior. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 80, 011908. doi:10.1103/PhysRevE.80.011908

Gibson, M. A., and Bruck, J. (2000). Efficient Exact Stochastic Simulation of Chemical Systems with many Species and many Channels. *J. Phys. Chem. A.* 104, 1876–1889. doi:10.1021/jp993732q

Gillespie, D. T. (1977). Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.* 81, 2340–2361. doi:10.1021/j100540a008

Gobet, C., and Naef, F. (2017). Ribosome Profiling and Dynamic Regulation of Translation in Mammals. *Curr. Opin. Genet. Development*. 43, 120–127. doi:10.1016/j.gde.2017.03.005

Gobet, C., Weger, B. D., Marquis, J., Martin, E., Neelagandan, N., Gachon, F., et al. (2020). Robust Landscapes of Ribosome Dwell Times and Aminoacyl-Trnas in Response to Nutrient Stress in Liver. *Proc. Natl. Acad. Sci. USA*. 117, 9630–9641. doi:10.1073/pnas.1918145117

Goldberg, A. P., Szigeti, B., Chew, Y. H., Sekar, J. A., Roth, Y. D., and Karr, J. R. (2018). Emerging Whole-Cell Modeling Principles and Methods. *Curr. Opin. Biotechnol.* 51, 97–102. doi:10.1016/j.copbio.2017.12.013

Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003). Comparing Protein Abundance and Mrna Expression Levels on a Genomic Scale. *Genome Biol.* 4, 117–118. doi:10.1186/gb-2003-4-9-117

Gritsenko, A. A., Hulsman, M., Reinders, M. J. T., and de Ridder, D. (2015). Unbiased Quantitative Models of Protein Translation Derived from Ribosome Profiling Data. *Plos Comput. Biol.* 11, e1004336. doi:10.1371/journal.pcbi.1004336

Hellen, C. U. T. (2018). Translation Termination and Ribosome Recycling in Eukaryotes. *Cold Spring Harb Perspect. Biol.* 10, a032656. doi:10.1101/cshperspect.a032656

Hussmann, J. A., Patchett, S., Johnson, A., Sawyer, S., and Press, W. H. (2015). Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. *Plos Genet.* 11, e1005732. doi:10.1371/journal.pgen.1005732

Ikemura, T. (1981). Correlation between the Abundance of escherichia Coli Transfer Rnas and the Occurrence of the Respective Codons in its Protein Genes. *J. Mol. Biol.* 146, 1–21. doi:10.1016/0022-2836(81)90363-6

Ikemura, T. (1985). Codon Usage and Trna Content in Unicellular and Multicellular Organisms. *Mol. Biol. Evol.* 2, 13–34. doi:10.1093/oxfordjournals.molbev.a040335

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. (2009). Genome-wide Analysis *In Vivo* of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* 324, 218–223. doi:10.1126/science.1168978

Ingolia, N. T., Hussmann, J. A., and Weissman, J. S. (2019). Ribosome Profiling: Global Views of Translation. *Cold Spring Harb. Perspect. Biol.* 11, a032698. doi:10.1101/cshperspect.a032698

Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011). Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell*. 147, 789–802. doi:10.1016/j.cell.2011.10.002

Kolomeisky, A. B. (1998). Asymmetric Simple Exclusion Model with Local Inhomogeneity. *J. Phys. A: Math. Gen.* 31, 1153–1164. doi:10.1088/0305-4470/31/4/006

Kozak, M. (1999). Initiation of Translation in Prokaryotes and Eukaryotes. *Gene*. 234, 187–208. doi:10.1016/s0378-1119(99)00210-3

Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. (2009). Coding-sequence Determinants of Gene Expression in escherichia Coli. *Science* 324, 255–258. doi:10.1126/science.1170160

Kummer, E., and Ban, N. (2021). Mechanisms and Regulation of Protein Synthesis in Mitochondria. *Nat. Rev. Mol. Cel. Biol.* 22, 1–19. doi:10.1038/s41580-021-00332-2

Lazarescu, A., and Mallick, K. (2011). An Exact Formula for the Statistics of the Current in the Tasep with Open Boundaries. *J. Phys. A: Math. Theor.* 44, 315001. doi:10.1088/1751-8113/44/31/315001

Li, G.-W., Burkhardt, D., Gross, C., and Weissman, J. S. (2014). Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell* 157, 624–635. doi:10.1016/j.cell.2014.02.033

Liu, Q., Shvarts, T., Sliz, P., and Gregory, R. I. (2020). Ribotoolkit: an Integrated Platform for Analysis and Annotation of Ribosome Profiling Data to Decode Mrna Translation at Codon Resolution. *Nucleic Acids Res.* 48, W218–W229. doi:10.1093/nar/gkaa395

Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E. M. (2007). Absolute Protein Expression Profiling Estimates the Relative Contributions of Transcriptional and Translational Regulation. *Nat. Biotechnol.* 25, 117–124. doi:10.1038/nbt1270

Lyu, X., Yang, Q., Li, L., Dang, Y., Zhou, Z., Chen, S., et al. (2020). Adaptation of Codon Usage to Trna I34 Modification Controls Translation Kinetics and Proteome Landscape. *Plos Genet.* 16, e1008836. doi:10.1371/journal.pgen.1008836

MacDonald, C. T., Gibbs, J. H., and Pipkin, A. C. (1968). Kinetics of Biopolymerization on Nucleic Acid Templates. *Biopolymers* 6, 1–25. doi:10.1002/bip.1968.360060102

Macklin, D. N., Ruggero, N. A., and Covert, M. W. (2014). The Future of Whole-Cell Modeling. *Curr. Opin. Biotechnol.* 28, 111–115. doi:10.1016/j.copbio.2014.01.012

Margaliot, M., and Tuller, T. (2013). Ribosome Flow Model with Positive Feedback. *J. R. Soc. Interf.* 10, 20130267. doi:10.1098/rsif.2013.0267

Merrick, W. C. (1992). Mechanism and Regulation of Eukaryotic Protein Synthesis. *Microbiol. Mol. Biol. Rev.* 56, 291–315. doi:10.1128/mmbr.56.2.291-315.1992

Merrick, W. C., and Pavitt, G. D. (2018). Protein Synthesis Initiation in Eukaryotic Cells. *Cold Spring Harb. Perspect. Biol.* 10, a033092. doi:10.1101/cshperspect.a033092

Michel, A. M., Mullan, J. P. A., Velayudhan, V., O'Connor, P. B. F., Donohue, C. A., and Baranov, P. V. (2016). Ribogalaxy: a Browser Based Platform for the Alignment, Analysis and Visualization of Ribosome Profiling Data. *RNA Biol.* 13, 316–319. doi:10.1080/15476286.2016.1141862

Miyazaki, M., and Esser, K. A. (2009). Cellular Mechanisms Regulating Protein Synthesis and Skeletal Muscle Hypertrophy in Animals. *J. Appl. Physiol.* 106, 1367–1373. doi:10.1152/japplphysiol.91355.2008

Mohammad, F., Green, R., and Buskirk, A. R. (2019). A Systematically-Revised Ribosome Profiling Method for Bacteria Reveals Pauses at Single-Codon Resolution. *Elife*. 8, e42591. doi:10.7554/eLife.42591

Munro, J. B., Vaiana, A., Sanbonmatsu, K. Y., and Blanchard, S. C. (2008). A New View of Protein Synthesis: Mapping the Free Energy Landscape of the Ribosome Using Single-Molecule Fret. *Biopolymers* 89, 565–577. doi:10.1002/bip.20961

Nedialkova, D. D., and Leidel, S. A. (2015). Optimization of Codon Translation Rates via Trna Modifications Maintains Proteome Integrity. *Cell* 161, 1606–1618. doi:10.1016/j.cell.2015.05.022

Neelagandan, N., Lamberti, I., Carvalho, H. J. F., Gobet, C., and Naef, F. (2020). What Determines Eukaryotic Translation Elongation: Recent Molecular and Quantitative Analyses of Protein Synthesis. *Open Biol.* 10, 200292. doi:10.1098/rsob.200292

O'Connor, P. B., Andreev, D. E., and Baranov, P. V. (2016). Comparative Survey of the Relative Impact of Mrna Features on Local Ribosome Profiling Read Density. *Nat. Commun.* 7, 12915. doi:10.1038/ncomms12915

Pavlov, M. Y., Watts, R. E., Tan, Z., Cornish, V. W., Ehrenberg, M., and Forster, A. C. (2009). Slow Peptide Bond Formation by Proline and Other N-Alkylamino Acids in Translation. *Proc. Natl. Acad. Sci.* 106, 50–54. doi:10.1073/pnas.0809211106

Pop, C., Rouskin, S., Ingolia, N. T., Han, L., Phizicky, E. M., Weissman, J. S., et al. (2014). Causal Signals between Codon Bias, Mrna Structure, and the Efficiency of Translation and Elongation. *Mol. Syst. Biol.* 10, 770. doi:10.15252/msb.20145524

Proud, C. G. (2006). Regulation of Protein Synthesis by Insulin. *Biochem. Soc. Trans.* 34, 213–216. doi:10.1042/bst0340213

Purcell, O., Jain, B., Karr, J. R., Covert, M. W., and Lu, T. K. (2013). Towards a Whole-Cell Modeling Approach for Synthetic Biology. *Chaos* 23, 025112. doi:10.1063/1.4811182

Qu, X., Wen, J.-D., Lancaster, L., Noller, H. F., Bustamante, C., and Tinoco, I. (2011). The Ribosome Uses Two Active Mechanisms to Unwind Messenger Rna during Translation. *Nature* 475, 118–121. doi:10.1038/nature10126

Radhakrishnan, A., Chen, Y.-H., Martin, S., Alhusaini, N., Green, R., and Coller, J. (2016). The Dead-Box Protein Dhh1p Couples Mrna Decay and Translation by Monitoring Codon Optimality. *Cell* 167, 122–132. doi:10.1016/j.cell.2016.08.053

Riba, A., Di Nanni, N., Mittal, N., Arhné, E., Schmidt, A., and Zavolan, M. (2019). Protein Synthesis Rates and Ribosome Occupancies Reveal Determinants of Translation Elongation Rates. *Proc. Natl. Acad. Sci. USA* 116, 15023–15032. doi:10.1073/pnas.1817299116

Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009). Automated Design of Synthetic Ribosome Binding Sites to Control Protein Expression. *Nat. Biotechnol.* 27, 946–950. doi:10.1038/nbt.1568

Shah, P., Ding, Y., Niemczyk, M., Kudla, G., and Plotkin, J. B. (2013). Rate-limiting Steps in Yeast Protein Translation. *Cell* 153, 1589–1601. doi:10.1016/j.cell.2013.05.049

Sharma, A. K., Ahmed, N., and O'Brien, E. P. (2018). Determinants of Translation Speed Are Randomly Distributed across Transcripts Resulting in a Universal Scaling of Protein Synthesis Times. *Phys. Rev. E.* 97, 022409. doi:10.1103/PhysRevE.97.022409

Sharma, A. K., and Chowdhury, D. (2011). Distribution of Dwell Times of a Ribosome: Effects of Infidelity, Kinetic Proofreading and Ribosome Crowding. *Phys. Biol.* 8, 026005. doi:10.1088/1478-3975/8/2/026005

Sharma, A. K., and Chowdhury, D. (2012). Template-directed Biopolymerization: Tape-Copying Turing Machines. *Biophys. Rev. Lett.* 07, 135–175. doi:10.1142/s1793048012300083

Sharma, A. K., and O'Brien, E. P. (2017). Increasing Protein Production Rates Can Decrease the Rate at Which Functional Protein Is Produced and Their Steady-State Levels. *J. Phys. Chem. B.* 121, 6775–6784. doi:10.1021/acs.jpcb.7b01700

Sharma, A. K., and O'Brien, E. P. (2018). Non-equilibrium Coupling of Protein Structure and Function to Translation-Elongation Kinetics. *Curr. Opin. Struct. Biol.* 49, 94–103. doi:10.1016/j.sbi.2018.01.005

Sharma, A. K., Sormanni, P., Ahmed, N., Ciryam, P., Friedrich, U. A., Kramer, G., et al. (2019). A Chemical Kinetic Basis for Measuring Translation Initiation and Elongation Rates from Ribosome Profiling Data. *Plos Comput. Biol.* 15, e1007070. doi:10.1371/journal.pcbi.1007070

Shaw, L. B., Kolomeisky, A. B., and Lee, K. H. (2004). Local Inhomogeneity in Asymmetric Simple Exclusion Processes with Extended Objects. *J. Phys. A: Math. Gen.* 37, 2105–2113. doi:10.1088/0305-4470/37/6/010

Szavits-Nossan, J., and Ciandrini, L. (2020). Inferring Efficiency of Translation Initiation and Elongation from Ribosome Profiling. *Nucleic Acids Res.* 48, 9478–9490. doi:10.1093/nar/gkaa678

Szavits-Nossan, J., Ciandrini, L., and Romano, M. C. (2018). Deciphering Mrna Sequence Determinants of Protein Production Rate. *Phys. Rev. Lett.* 120, 128101. doi:10.1103/physrevlett.120.128101

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., et al. (2010). An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell.* 141, 344–354. doi:10.1016/j.cell.2010.03.031

Tunney, R., McGlincy, N. J., Graham, M. E., Naddaf, N., Pachter, L., and Lareau, L. F. (2018). Accurate Design of Translational Output by a Neural Network Model of Ribosome Distribution. *Nat. Struct. Mol. Biol.* 25, 577–582. doi:10.1038/s41594-018-0080-2

Vogel, C., de Sousa Abreu, R., Ko, D., Le, S. Y., Shapiro, B. A., Burns, S. C., et al. (2010). Sequence Signatures and mRNA Concentration Can Explain Two-thirds of Protein Abundance Variation in a Human Cell Line. *Mol. Syst. Biol.* 6, 400. doi:10.1038/msb.2010.59

Volkov, I. L., and Johansson, M. (2018). Single-molecule Tracking Approaches to Protein Synthesis Kinetics in Living Cells. *Biochemistry*. 58, 7–14. doi:10.1021/acs.biochem.8b00917

Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B., and Bartel, D. P. (2016). Improved Ribosome-Footprint and Mrna Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cel. Rep.* 14, 1787–1799. doi:10.1016/j.celrep.2016.01.043

Wu, L., and Belasco, J. G. (2008). Let Me Count the Ways: Mechanisms of Gene Regulation by Mirnas and Sirnas. *Mol. Cel.* 29, 1–7. doi:10.1016/j.molcel.2007.12.010

Xiao, Z., Zou, Q., Liu, Y., and Yang, X. (2016). Genome-wide Assessment of Differential Translations with Ribosome Profiling Data. *Nat. Commun.* 7, 11194. doi:10.1038/ncomms11194

Zia, R. K. P., Dong, J. J., and Schmittmann, B. (2011). Modeling Translation in Protein Synthesis with Tasep: A Tutorial and Recent Developments. *J. Stat. Phys.* 144, 405–428. doi:10.1007/s10955-011-0183-1

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 © 2021 Yadav, Ullah Irshad, Kumar and Sharma. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 22 July 2021
doi: 10.3389/fmolb.2021.644133

Generating Chromosome Geometries in a Minimal Cell From Cryo-Electron Tomograms and Chromosome Conformation Capture Maps

*Benjamin R. Gilbert[1], Zane R. Thornburg[1], Vinson Lam[2], Fatema-Zahra M. Rashid[3,4], John I. Glass[5], Elizabeth Villa[2], Remus T. Dame[3,4] and Zaida Luthey-Schulten[1]**

[1]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [2]Division of Biological Sciences, University of California San Diego, San Diego, CA, United States, [3]Leiden Institute of Chemistry, Leiden University, Leiden, Netherlands, [4]Center for Microbial Cell Biology, Leiden University, Leiden, Netherlands, [5]Synthetic Biology Group, J. Craig Venter Institute, La Jolla, CA, United States

OPEN ACCESS

Edited by:
Joanna Trylska,
University of Warsaw, Poland

Reviewed by:
Slavica Jonic,
UMR7590 Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie (IMPMC), France
Jean-Charles Walter,
UMR5221 Laboratoire Charles Coulomb (L2C), France
Ivan Junier,
UMR5525 Techniques de l'Ingénierie Médicale et de la Complexité Informatique, Mathématiques et Applications, Grenoble (TIMC-IMAG), France

***Correspondence:**
Zaida Luthey-Schulten
zan@illinois.edu

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: 20 December 2020
Accepted: 14 May 2021
Published: 22 July 2021

Citation:
Gilbert BR, Thornburg ZR, Lam V, Rashid F-ZM, Glass JI, Villa E, Dame RT and Luthey-Schulten Z (2021) Generating Chromosome Geometries in a Minimal Cell From Cryo-Electron Tomograms and Chromosome Conformation Capture Maps. Front. Mol. Biosci. 8:644133. doi: 10.3389/fmolb.2021.644133

JCVI-syn3A is a genetically minimal bacterial cell, consisting of 493 genes and only a single 543 kbp circular chromosome. Syn3A's genome and physical size are approximately one-tenth those of the model bacterial organism *Escherichia coli*'s, and the corresponding reduction in complexity and scale provides a unique opportunity for whole-cell modeling. Previous work established genome-scale gene essentiality and proteomics data along with its essential metabolic network and a kinetic model of genetic information processing. In addition to that information, whole-cell, spatially-resolved kinetic models require cellular architecture, including spatial distributions of ribosomes and the circular chromosome's configuration. We reconstruct cellular architectures of Syn3A cells at the single-cell level directly from cryo-electron tomograms, including the ribosome distributions. We present a method of generating self-avoiding circular chromosome configurations in a lattice model with a resolution of 11.8 bp per monomer on a 4 nm cubic lattice. Realizations of the chromosome configurations are constrained by the ribosomes and geometry reconstructed from the tomograms and include DNA loops suggested by experimental chromosome conformation capture (3C) maps. Using ensembles of simulated chromosome configurations we predict chromosome contact maps for Syn3A cells at resolutions of 250 bp and greater and compare them to the experimental maps. Additionally, the spatial distributions of ribosomes and the DNA-crowding resulting from the individual chromosome configurations can be used to identify macromolecular structures formed from ribosomes and DNA, such as polysomes and expressomes.

Keywords: cryo-electron tomography, chromosome conformation capture (3C) maps, computational modeling, whole-cell models, chromosome modeling, ribosome distribution, bacterial minimal cell, JCVI-syn3A

1 INTRODUCTION

JCVI-syn3A is a genetically minimal bacterial cell, consisting of 493 genes and a single short 543 kbp circular chromosome derived from a Gram-positive bacterium, *Mycoplasma mycoides*. Previous work established genome-scale gene essentiality and proteomics data along with its essential metabolic network (Hutchison et al., 2016; Breuer et al., 2019) and a kinetic model of genetic

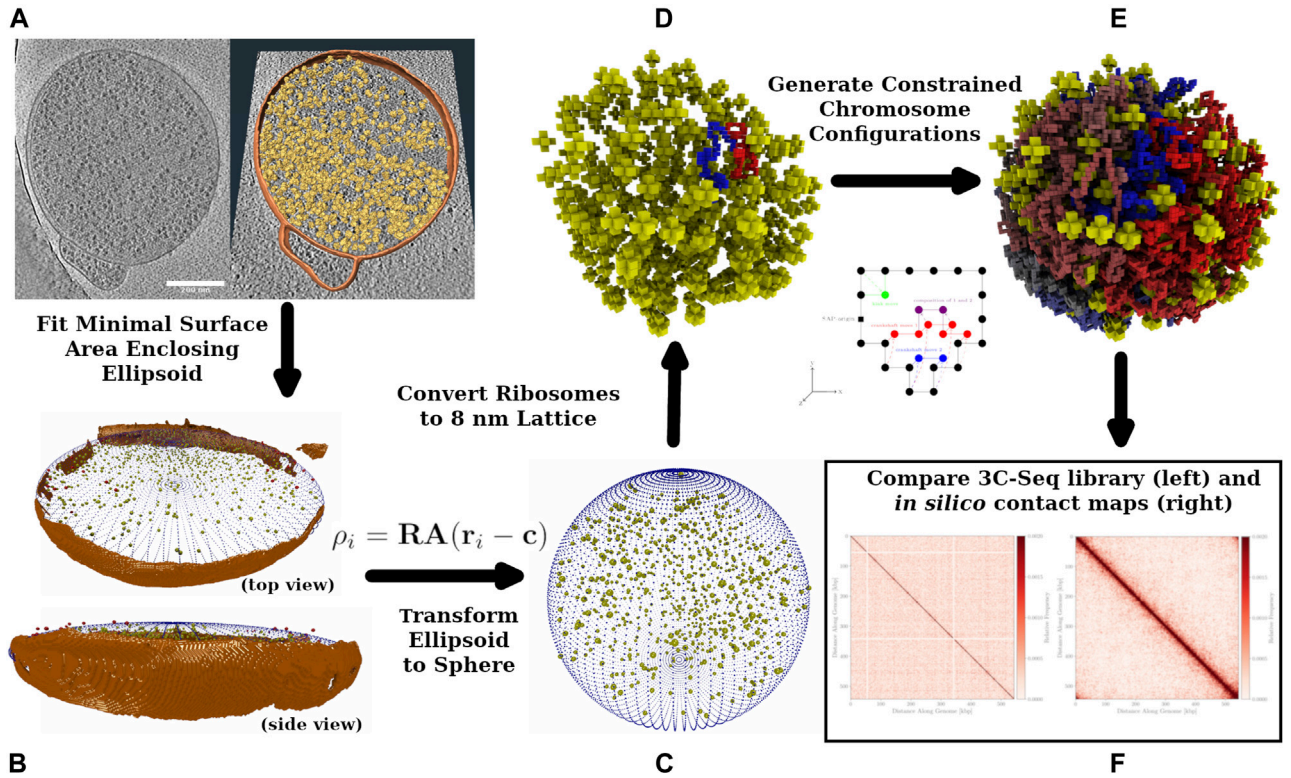


FIGURE 1 | Workflow Diagram: Cryo-ET data is used to reconstruct spherical Syn3A cells, constrained chromosome configurations are generated within the reconstructed cells and resulting *in silico* chromosome contact maps are compared to the experimental 3C-Seq map. **(A)**: The left side is a single z-slice of the tomographic reconstruction. The right side are the ribosomes (yellow) identified using template-matching and the membrane segmentation (orange) superimposed on the z-slice. **(B)**: The shape of the blotted cells is approximated by an ellipsoid that is manually compared to the tomographic membrane segmentation (orange). In an iterative procedure, a series of minimal surface area enclosing ellipsoids (MSAEE) are fit around the ribosome coordinates while extraneous ribosomes assessed to be outside of the true membrane are removed within each iteration. The final fitted ellipsoid is shown in blue and the extraneous ribosomes in red. **(C)**: As Syn3A cells are known to have a spherical morphology, the ribosome coordinates and ellipsoidal membrane approximation are then transformed to a sphere with equivalent surface area. **(D)**: The continuum representation is then converted to an 8 nm cubic lattice representation used for whole-cell simulations with LM. **(E)**: Circular and self-avoiding chromosome configurations are generated as a lattice polymer on a 4 nm cubic lattice. The 4 nm lattice is coincident with the 8 nm cubic lattice and the chromosomes are constrained to avoid the ribosomes and remain within the membrane. In the representative DNA configuration, monomers are colored red and blue on opposite arms of the chromosome. **(F)**: *In silico* contact maps from ensembles of generated DNA configurations are compared to the experimental 3C-Seq map.

information processing (Thornburg et al., 2019). The kinetic model of genetic information processes of DNA replication, transcription, and translation requires positions and length of genes along the chromosome. Spatially resolving and solving these kinetic models require cellular architecture, including spatial distributions of ribosomes and the circular chromosome. The first iteration of a bacterial cell with a synthetic genome was JCVI-syn1.0, whose synthesized genome is 1,079 kbp (Gibson et al., 2010). Syn1.0 has a doubling time of approximately 60 min and shows a spherical morphology with a radius of ~200 nm (Hutchison et al., 2016). Two additional cycles of targeted genomic reduction resulted in JCVI-syn3.0, a cell whose synthetic genome is only 531 kbp but still autonomously replicates (Hutchison et al., 2016). Syn3.0 has a slower growth rate than Syn1.0, with a doubling time of approximately 180 min (Hutchison et al., 2016), and based on optical and scanning electron microscopy (SEM), Syn3.0 exhibits a pleomorphic morphology with significant variations (Hutchison et al., 2016; Pelletier et al., 2021). The organism that is the subject of this study, Syn3A, was created from Syn3.0 through the addition of 19 genes present in Syn1.0. While this addition made a less minimal genome, it resulted in cells with a robust spherical morphology and an average doubling time of approximately 110 min (Breuer et al., 2019). Super-resolution fluorescence microscopy (STORM) imaging (private communication, Taekjip Ha) reveals that it recovers the spherical morphology of Syn1.0, with a radius of 200–250 nm.

To create chromosome geometries for our spatial models and subsequent simulations of gene expression and translation, we develop a method of generating self-avoiding circular chromosome configurations with a resolution of 11.8 bp per monomer on a 4 nm cubic lattice. To place the chromosome inside the cell volume, we use cryo-electron tomography (cryo-ET) to define the cell boundaries and ribosome distribution, which define the regions available to the chromosome. Cryo-ET data shows that the ribosomes appear to be nearly randomly distributed throughout the cell. In cryo-ET of bacteria, the position of the chromosome is typically determined by the

absence of ribosomes. For example, the cryo-ET studies of slow-growing *Escherichia coli* show ribosomes primarily localized at the poles and along the sides of the nucleoid so that the DNA can be inferred to be confined within the enclosed nucleoid region (Roberts et al., 2011). Based on the tomograms for Syn3A presented in this study, we assume the chromosome is randomly distributed among the ribosomes. We also present experimental chromosome conformation capture (3C) maps, a technique that shows the frequency at which different regions of the chromosome are in contact with each other (Dekker et al., 2002). This map shows only a single main diagonal with some small (< 4 kbp) features along it and has no other significant features.

Using our knowledge of Syn3A's proteome data and genome, along with the experimental cryo-ET and 3C-Seq maps, we have created a physics-based model of the chromosome to generate chromosome configurations and predict contact maps. A diagram of the workflow is presented in **Figure 1** that shows the process of annotating ribosome locations and membrane from the tomograms and using the ribosome locations as constraints on chromosome configurations. The configurations are also influenced by the features present in the experimental contact map. Hi-C analysis, a variant of the chromosome conformation capture (3C) method, has been used extensively to describe the structure of eukaryotic chromatin (Lieberman-Aiden et al., 2009; van Berkum et al., 2010; Belton et al., 2012). Those chromosome contact maps are used to generate chromatin structures based on topologically associated domains (TADs) observed in the contact maps (Dekker et al., 2013; Rao et al., 2014; Fudenberg et al., 2016; Di Pierro et al., 2017). While considerations of the energy functions used in the chromatin models for eukaryotic studies are helpful in designing a bacterial study, there are bacteria-specific proteins and related effects that need to be considered when constructing a bacterial chromosome model and how the effects would appear in the resulting contact map (Le et al., 2013; Marbouty et al., 2015; Verma et al., 2019). The 3D structure of the circular bacterial chromosome at both the global and local levels is determined by effects of various nucleoid associated proteins (NAPs) and is also influenced by the crowding of ribosomes. With its reduced genome, Syn3A lacks many of the NAPs that cause significant features in the chromosome structure which leads to considerable variation from the structures and Hi-C/3C maps observed in other bacteria such as *Mycoplasma pneumoniae*, *Bacillus subtilis*, *Caulobacter crescentus*, and *Pseudomonas aeruginosa*.

The global structure of the chromosome, also known as the cellular disposition of the chromosome (Lioy et al., 2018, 2020), describes how regions of the chromosome are organized within the confines of the cell. Factors affecting the global structure include possible attachment of the chromosome to the membrane, as in *M. pneumoniae*, and loading of SMC proteins near the origin by a complete parABS system, which causes alignment of the two arms of the circular chromosome (Wang et al., 2014; Wang and Rudner, 2014; Lioy et al., 2020). The parABS system includes two proteins, parA and parB, which site-specifically load SMC onto *parS* sites on the DNA (Livny et al., 2007). Both of these effects result in a secondary diagonal, orthogonal to the main diagonal, in chromosome contact maps as observed in maps for *M. pneumoniae*, *B. subtilis*, *C. crescentus*, and *P. aeruginosa* (Le et al., 2013; Marbouty et al., 2015; Tran et al., 2017; Trussart et al., 2017; Lioy et al., 2020). However, the experimental contact map presented in the results section reveals that Syn3A does not have a secondary diagonal. Syn3A does not have a complete parABS system because it lacks the parB protein (Breuer et al., 2019) and the complete signature *parS* sites, i.e. no sequences greater than a 10/16 match to the consensus sequence (Livny et al., 2007) were identified in a BLAST search of the genome. For comparison, when performing the same search on *B. subtilis*, we found matches of 14/16 and higher. Livny et al. also identified *Meso*, *Urea* and *Mycoplasmas* as members of the Firmicutes that lack complete parABS systems (Livny et al., 2007). Therefore, we do not expect to see alignment of the two arms of the chromosome *via* the parABS system and would not expect to see a secondary diagonal due to this effect. Additionally, Syn3A does not have any annotated proteins that attach the DNA to the membrane (Breuer et al., 2019), so we would not expect to see a secondary diagonal due to attachment of the DNA, unlike *M. pneumoniae* which has an attachment organelle (Trussart et al., 2017).

Factors affecting local structure include supercoiling, plectonemic loops resulting from supercoiling, small loops formed by SMC bridging distant chromosome segments, and bending and stiffening by proteins such as histone-like protein (HU), heat-stable nucleoid structuring protein (H-NS), and integration host factor (IHF) (Dame, 2005; Ohniwa et al., 2011; Dame and Tark-Dame, 2016; Dame et al., 2019; Verma et al., 2019; Birnie and Dekker, 2021). These micro level effects can strongly affect gene expression as localized crowding affects the access of the RNA polymerase (RNAP) to genes and supercoiling and plectonemes affect the RNAP's ability to transcribe a gene (Kim et al., 2019). Of the proteins HU, IHF, and H-NS, Syn3A only has one gene, JCVISYN3A_0350, which is annotated as a putative histone-like protein with a proteomics count of 28. The count of 28 for this protein is significantly lower than the counts seen in other bacteria. For example, fast growing *E. coli* contains more than 12,000 HU (Wang et al., 2015), *B. subtilis* contains almost 9,000 HU (Wang et al., 2015), and *Mesoplasma florum* contains 9,500 HU (Matteau et al., 2020). Due to its small count, we do not expect any significant contributions to the stabilizing of chromosome loops by the protein encoded by gene JCVISYN3A_0350 and do not include it in our model.

Supercoiling is formed during transcription by the RNAP, which induces positive supercoiling in the forward direction and negative supercoiling in the reverse (Chong et al., 2014; Verma et al., 2019). Supercoiling can be eliminated by topoisomerases, gyrases, and positive/negative supercoiling annihilating each other along free DNA (Chong et al., 2014; Dorman, 2019; Verma et al., 2019). The experimental contact map presented in the Results is too sparse to distinguish between short (< 10 kbp) supercoiled domains and loops and we do not see any larger interaction domains. We do not include supercoiling because of this, and given the low proteomics count of HU, we infer the DNA is in a relaxed state. As discussed above, Syn3A does not have genes that code for proteins that would attach the

DNA to the cell membrane. Since the chromosome is therefore most likely not fixed at any location, we assume that, in general, DNA is mostly free, allowing positive and negative supercoiling to annihilate each other more easily. Additionally, the genome-wide proteomics counts show that we have a total of 187 RNAP and roughly 250 DNA gyrases that can alleviate positive supercoiling, 150 type IV DNA topoisomerases that can alleviate negative supercoiling, and 175 type I topoisomerases that can alleviate either (Chong et al., 2014; Breuer et al., 2019). Other bacteria are observed to have fewer topoisomerases and gyrases than RNAP, for example, fast-growing *E. coli* has roughly 3,800 topoisomerase I, 1,200 topoisomerase IV, and 6,000 to 8,000 gyrases while having over 10,000 RNAP (Bremer and Dennis, 2008; Wang et al., 2015). Another Gram-positive bacterium, *B. subtilis*, has 3,000 RNAP while only having 1,200 gyrases, 900 topoisomerase I, and 200 topoisomerase IV (Wang et al., 2015). The more closely related *M. pneumoniae* has 5,000 topoisomerase I, 200 topoisomerase IV, and 1,800 gyrases while having around 6,000 RNAP (Kühner et al., 2009). It is then more likely in these systems where larger domains have been observed in their chromosome contact maps that the proteins removing supercoiling cannot keep up with the supercoiling induced by RNAP due to their lower relative counts. Therefore, it is our assumption that as supercoiling is formed by RNAP in Syn3A, there are sufficient gyrases, topoisomerases, and negative/positive supercoiling pair annihilations to keep the DNA in a more relaxed configuration with no significant supercoiled domains.

While Syn3A does not have a complete parAB*S* system, it does contain 202 structural maintenance of chromosomes (SMC) proteins, which can bridge distant loci *via* loop extrusion powered by ATP-hydrolysis (Ganji et al., 2018; van Ruiten and Rowland, 2018). The SMC protein is a long coiled-coil protein that dimerizes and has head and hinge domains separated by approximately 50 nm (Diebold-Durand et al., 2017). The number of SMC in Syn3A is smaller than the 448 observed in *B. subtilis* (Wang et al., 2015) and 900 observed in *M. pneumoniae* (Kühner et al., 2009), but Syn3A also has a smaller volume and shorter chromosome, which could result in a higher density of loops. *M. florum* is not much larger than Syn3A and only has 85 SMCs (Matteau et al., 2020). With a higher density of SMC in both volume and chromosome length, we assume the effects of SMC looping can be significant in the chromosome structure of Syn3A. We manually annotate any of the observed regions of contact along the main diagonal in the experimental 3C-Seq map as possible loops (< 4 kbp) and implement them as looping restraints in our chromosome model.

Finally, based on the cryo-ET images of Syn3A cells presented in the results, the chromosome in Syn3A is more constrained by ribosomes than in other bacteria. From the cryo-ET of Syn3A we infer that the ribosomes are uniformly distributed throughout the cells and that there is no clearly-defined condensed nucleoid region. The lack of a condensed nucleoid region is in contrast to the rod-shaped *E. coli* where the ribosomes are primarily located at the poles and along the sides of the nucleoid region (Nevo-Dinur et al., 2011; Roberts et al., 2011; Bakshi et al., 2012). We saw this distribution in cryo-ET data of slow-growing *E. coli* that was part of a previous Lattice Microbes (LM) simulation of the lac genetic switch (Roberts et al., 2011). We observe a ribosome number density of 12,920–19,370 ribosomes/μm^3 in Syn3A cells, which is higher than the density of 4,200 ribosomes/μm^3 in *M. pneumoniae* (Trussart et al., 2017; O'Reilly et al., 2020). The density of ribosomes in *E. coli* was previously found to be 27,000 ribosomes/μm^3 (Bakshi et al., 2012), which is greater than the density in Syn3A, but the inferred ribosome density within the nucleoid region is 2,000–8,000 ribosomes/μm^3. Given this, relative to other bacterial cells, the crowding of the ribosomes in Syn3A more strongly constrains the possible chromosome configurations.

In this paper, we first explain how the cellular architecture and ribosome distributions are obtained from three-dimensional cryo-electron tomograms. Using ensembles of constrained DNA configurations from our circular chromosome model on a lattice, we predict contact maps for individual cells at resolutions of 250 bp and greater and compare them to our experimental 3C-Seq map at 1,000 bp resolution. The DNA configurations in this study are generated with the intent of incorporating them into stochastic whole-cell models of Syn3A simulated using the reaction-diffusion master equation (RDME) as implemented in LM (Roberts et al., 2013; Hallock et al., 2014; Earnest et al., 2018). In the whole-cell simulations, the cellular space is divided into cubic subvolumes, so we chose to model the DNA as a lattice polymer. The DNA configurations, cell sizes, and ribosome locations presented here will later be directly incorporated into cell geometry in the kinetic simulations and will influence both diffusion and the locations at which genetic information reactions take place. We also identify potential complexes formed from ribosomes and DNA in our spatial model, such as polysomes and expressomes (O'Reilly et al., 2020), that would affect the reactions within a kinetic model.

2 METHODS

2.1 Reconstructing Cell Geometries From Cryo-Electron Tomograms

2.1.1 Tomogram Collection and Processing

One of the primary challenges for cellular cryo-ET is to prepare a specimen such that it is thin enough to be transparent to electrons and to vitrify thoroughly. Due to the small size of Syn3A, this can be accomplished by placing cells on a Quantifoil EM grid and blotting the majority of the liquid away followed by plunge freezing, leaving a thin layer of ice with cells embedded within it.

Initially, samples were frozen on Quantifoil Cu 200 mesh R 1/4 grids (Electron Microscopy Sciences), which have patterned holes 1 μm in diameter, with 4 μm spacing between holes (i.e., 5 μm periodicity). Our rationale was that using smaller holes would retain more Syn3A cells on the grid after blotting. However, we found that because the synthetic cells lack a cell wall, they are supple and get distorted in the direction of the flow of the medium as it is blotted away from the grid. This issue was resolved by switching to Quantifoil Cu 200 mesh R 2/1 grids, which have larger holes. Syn3A cells were grown to mid-log-phase at 37°C in SP4 medium (Williamson and Whitcomb, 1975) using KnockOut™ serum replacement (Invitrogen), to a density

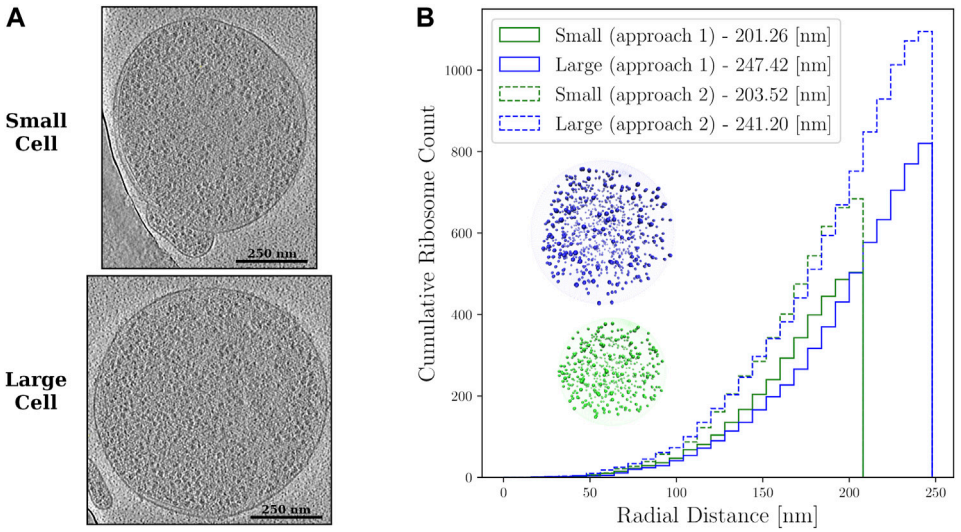


FIGURE 2 | (A)—Z-slices from the cryo-ET data of the small and large Syn3A cells. The ribosomes are the objects with higher density than the surrounding cytoplasm that are distributed throughout the cells. **(B)**—Cumulative ribosome distributions with 8 nm bins in the reconstructed spherical geometries of a small cell of radius 201.26 nm with 503 ribosomes and large cell of radius 247.42 nm with 820 ribosomes when the first approach was used, and a small cell of radius 203.52 nm with 684 ribosomes and large cell of radius 241.20 nm with 1,095 ribosomes when the second approach was used. Also shown are the reconstructed spherical geometries resulting from the first approach to template matching.

of $\sim 10^8$ cells/ml. 4 µl of sample was deposited on glow-discharged grid, blotted on the backside of the grid for 6 s using Whatman No. 1 filter paper, and plunged into a 50/50 mixture of ethane and propane (Airgas) cooled to liquid nitrogen temperatures using a manual plunger (Max Planck Institute of Biochemistry).

These *Mycoplasma* cells with synthetic genomes were more radiation sensitive than we have encountered for other bacteria. Imaging conditions were chosen to keep cumulative electron dose under 120 e/Å^2. All data were acquired using a Titan Krios (ThermoFisher Scientific, TFS) at 300 kV and a Gatan K2 camera with a GIF energy filter, using SerialEM v3.7.4 automated protocols (Mastronarde, 2005; Schorb et al., 2019). The microscopic parameters were: 1) Pixel size: 0.53 nm (FOV: 2 µm) or 0.43 nm (FOV: 1.6 µm), 2) Target defocus: 6 µm, 3) Total accumulated dose: 90–120 e/Å^2, 4) Tilt scheme: dose symmetric from 0° to ± 60° every 2°, 5) 70 µm objective aperture. Individual tilt-series frames were aligned using MotionCor2 (Zheng et al., 2017). Tomograms were reconstructed using IMOD v4.10.29 (Kremer et al., 1996; Mastronarde, 1997; Mastronarde and Held, 2017) and binned by four for downstream template matching. Additionally, non-linear anisotropic diffusion (NAD) filtering was applied in IMOD to enhance contrast for visualization.

At the pixel size and target defocus used for acquisition, the ribosome distributions are easily discerned and can be seen for the small and large cells in **Supplementary Figures S1,S2**, respectively. The small cell's dimensions and ribosome count were in good agreement with those reported previously (Hutchison et al., 2016; Breuer et al., 2019). However, the cells were flattened into ellipsoids, and sometimes further elongated. This well-known effect from blotting seems amplified in these cells due to the absence of a cell wall. The frozen cells were flattened to ~ 160 nm.

To determine the ribosome distribution inside cells, we used two different approaches based on template matching, with one of them continuing to 3-D classification. First, one has to identify all ribosomes within the tomogram. Template matching is performed by creating a 3-D template of the target structure, and comparing it to each voxel in the tomogram using a 6-D search (three spatial and three rotational degrees of freedom) to identify regions that correlate highly with the template. It is noteworthy that the contrast difference between ribosomes and their surroundings in Syn3A was greatly reduced compared to other bacteria, e.g. *E. coli*, suggesting that the mass density (molecular crowding) of Syn3A is higher.

In our first approach, we used Dynamo v1.1.509 (Castaño-Díez et al., 2012) with a bacterial ribosome structure (PDB: 5MDZ) as the initial template filtered to 20 Å resolution in UCSF Chimera (Pettersen et al., 2004), resampled to match the pixel size, and contrast scaled to match that of the target tomograms. A threshold cross correlation was selected so that it contained most ribosomes that were clearly inside the cell boundary. Final particle positions were inspected visually, and removed if they were membrane segments. Membranes were segmented using TomoSegMemTV (Martinez-Sanchez et al., 2014), and ribosomes outside of this segmented membrane were excluded. Starting with a high-correlation threshold, the first approach initially identified 547 ribosomes in the small cell and 849 ribosomes in the large cell. Fitting approximate cell boundaries in **section 2.1.2** reduced these ribosome counts to 503 and 820 for the small cell and large cell, respectively.

In our second approach, tilt-series were preprocessed using Warp v1.0.9 (Tegunov and Cramer, 2019) for sub-frame motion

TABLE 1 | Summary of the ribosome distributions and cell geometries resulting from the two template matching methods for both the small and the large cell.

	Small cell		Large cell	
	Approach 1	**Approach 2**	**Approach 1**	**Approach 2**
Ribosomes from template matching	547	718	849	1,136
Extraneous ribosomes	44	34	29	41
Remaining ribosomes	503	684	820	1,095
SA-equivalent sphere radius (nm)	201.26	203.52	247.42	241.20
Ribosome density (ribosomes/μm^3)	14,730	19,370	12,920	18,630

correction and 3D-CTF estimation. Tilt-series were aligned using IMOD and the final reconstructions were created in Warp for subsequent processing. Template matching of ribosomes within tomograms was performed in Warp, using an initial ribosome template generated from about 200 manually picked particles from Syn3A tomograms using IMOD to avoid template bias. Extracellular particles were initially discarded based on cell boundaries defined in Dynamo. Obvious false positives (e.g., membrane segments, ice particles) were manually removed. The remaining particles were used for 3D alignment and classification in Relion v3.1 (Scheres, 2012). For each cell of interest, particles were subject to successive rounds of binary classification with a large (500 Å or 83 binned pixels) mask, the class which contained particles that did not appear as ribosomes were removed from subsequent rounds. This was done until the two classes reached about equal population. A schematic of the overall process is presented in **Supplementary Figure S3**. Coordinates and orientations of the remaining particles were imported into Amira for visualization. While starting with a lower correlation threshold, this second approach resulted in 718 ribosomes in the small cell and 1,136 ribosomes in the large cell, as the quality of fit increased. An additonal round of binary classification, deemed too restrictive, gave counts similar to the first approach. Fitting approximate cell boundaries in **section 2.1.2** reduced these ribosome counts to 684 and 1,095 for the small cell and large cell, respectively.

The second approach that starts with a lower correlation threshold and includes subsequent iterative 3-D classification, is more accurate to find the final true-positive ribosomes and ribosome distributions (Lasker et al., 2021). However, it requires considerably more resources and expertise. Thus, we introduce both approaches. Even though they give slightly different distributions, both are in agreement with estimates from other experimental and computational data (Breuer et al., 2019), and notably do not significantly affect the outcome of the chromosome geometries generated, as shown in **Figure 2**. A summary of the ribosome counts for both approaches at each stage of our workflow are presented in **Table 1**.

2.1.2 Determining the Spherical Cell Size and Ribosome Distribution

Given a set of ribosome coordinates, the bounding membrane and shape of the deformed cell can be approximated using an ellipsoid. This was done by calculating an ellipsoid with minimal surface area that encloses the centers of all the ribosomes. The solution for the minimal surface area enclosing ellipsoid (MSAEE) was found using the minimize routine in the SciPy package with the sequential least-squares programming (SLSQP) method. To optimize the calculation, only the convex polytope of the ribosome coordinates was used to constrain the enclosing ellipsoid. The optimal enclosing ellipsoid is always constrained by four ribosomes that form a tetrahedral shape bounding the ellipsoid.

Some ribosomes identified by template matching are extraneous ribosomes, e.g., ribosomes that are present in the cell periphery but correspond to a nearby lysed cell. The extraneous ribosomes were iteratively removed from the set of coordinates and a series of ellipsoids were iteratively fit after each extraneous ribosome removal until the relative change in the ellipsoid surface area between iterations fell below 0.001%. At the end of each iteration, we choose the ribosome among the four bounding ribosomes having the greatest projection along the major axis of the enclosing ellipsoid as the extraneous ribosome and remove it. The number of extraneous ribosomes removed are summarized in **Table 1** for all cases. **Figure 2A** shows z-slices of the cryo-ET data for both the small and large Syn3A cells. Notably, the spatial distribution of ribosomes in these cells are largely homogeneous, but small regions of about 150 nm appear to have fewer ribosomes than the surrounding cytoplasm.

After an ellipsoid approximating the membrane surface was calculated, both the ellipsoid and the enclosed ribosome coordinates were transformed to a spherical cell with equivalent surface area. A surface-area preserving transformation was chosen as previous measurements on bilayer vesicles indicated that the membrane area can only strain by approximately 5% before lysing (Needham and Nunn, 1990), thus we assume that there is a small change in volume during the blotting procedure due to mass transport of water across the membrane. The equation of an ellipsoid centered at $\boldsymbol{c}$ is $\{\boldsymbol{x} \in \mathbb{R}^3 \big| (\boldsymbol{x}-\boldsymbol{c})^T \boldsymbol{A}^T \boldsymbol{A} (\boldsymbol{x}-\boldsymbol{c}) \leq 1\}$, where $\boldsymbol{A}$ is the matrix describing the shape of the ellipsoid, and the equation of a sphere centered at the origin is $\{\boldsymbol{x} \in \mathbb{R}^3 \big| \boldsymbol{x}^T \boldsymbol{R}^{-2} \boldsymbol{x} \leq 1\}$, where $\boldsymbol{R} = R\boldsymbol{I}$ and R is the radius of the sphere. For ribosome coordinates, $\{\boldsymbol{r}_i\}$, the transformed coordinates, $\{\boldsymbol{\rho}_i\}$, are given by transforming all of the coordinates to a unit sphere centered at the origin by translating them by the vector $\boldsymbol{c}$ and transforming them with the matrix $\boldsymbol{A}$. The coordinates in the unit sphere representation are then scaled by the matrix $\boldsymbol{R}$ to a sphere with surface area equivalent to the MSAEE. The overall transformation is given by

$$\boldsymbol{\rho}_i = \boldsymbol{R}\boldsymbol{A}(\boldsymbol{r}_i - \boldsymbol{c}). \quad (1)$$

The transformation preserves the relative distances amongst the ribosomes and the shapes of the voids between the ribosomes. In a

final step, the ribosome coordinates are expanded anisotropically along the semiaxes of the ellipsoid to ensure the ribosomes at the extremes reach the membrane. Two representative spherical geometries resulting from this transformation and the radial distribution of ribosomes are shown in **Figure 2B**.

After the transformed ribosome coordinates are determined for the spherical cell, the coordinates are projected onto the 8 nm cubic lattice used for Lattice Microbes simulations and converted to a star shape comprised of seven 8 nm cubic sites to approximate the ribosome diameter of 20 nm. The set of ribosome coordinates on the 8 nm lattice and the boundary imposed by the cell membrane then serve as constraints when generating the ensemble of chromosome configurations.

2.2 Modeling Bacterial Chromosome Configurations

The three primary objectives of creating a chromosome model for Syn3A are creating realistic spatial heterogeneities due to DNA crowding that are discernable at the 8 nm resolution used in spatially-resolved kinetic models of Syn3A, matching the cell architecture dictated by the cell boundary and ribosome distribution, and reproducing the intra-chromosomal interactions in chromosome conformation capture experiments through DNA-looping.

Computational models for chromosomes can be broadly classified into two groups, direct models and inverse models (Rosa and Zimmer, 2014). This distinction is not entirely black and white and it is discussed in the following paragraphs. Direct models use a minimal set of assumptions about the underlying physics of DNA or chromatin to create a polymer model, and the results of simulating the model can then be compared to experimental data (Rosa and Zimmer, 2014). These models range on length-scale from 1 bp per monomer models of the *E. coli* chromosome (Hacker et al., 2017) to 500–50,000 bp per monomer models of human chromosomes (Di Pierro et al., 2016). The models at the smallest length scales often use a Kratky-Porod model (Kratky and Porod, 1949) or a worm-like chain model for the polymer, where the persistence length of the DNA is explicitly incorporated. In contrast, the models at the largest length scales often use a Rouse model (Rouse, 1953) for the polymer, in which the monomers are assumed to be uncorrelated equilibrium globules of DNA. These models based on Rouse dynamics are well-suited for eukaryotic chromosomes on the order of 10^7-10^8 bp, where the DNA is organized in nucleosomes comprised of histone octamers and other higher-order structures. A comprehensive discussion of possible interactions in the direct models of DNA polymers can be found in the review by Haddad et al. (Haddad et al. 2017) and the Minimal Chromatin Model of Di Pierro et al. (Di Pierro et al. 2016). The complexity of interactions in polymer models of DNA can range from those in homopolymer models to block copolymer models, and finally heteropolymer models (Haddad et al., 2017). Additionally, direct chromosome models can include the influence of NAPs, SMC, or bridging proteins in strings and binders models (Annunziatella et al., 2018; Ryu et al., 2021), where other particles diffuse amongst the chromosome and cause multi-point intrachromosomal interactions. After a polymer model has been specified and the chromosome of interest has been mapped to the model, molecular dynamics or Monte Carlo methods are used to sample configurations of the direct models.

Inverse models are data-driven and use large sets of experimental data to create a compatible model (Rosa and Zimmer, 2014; Oluwadare et al., 2019). The most common form of experimental data used in inverse models are chromosome contact maps resulting from 3C methods. The interaction frequencies in the contact maps are inverted to produce distance-based restraints for the chromosome models (Rosa and Zimmer, 2014). In addition to these distance-based restraints, constraints that are based on the known properties of the chromosome, such as the topology and excluded-volume effects, can be incorporated into the inverse models. A single ideal chromosome configuration that simultaneously satisfies all restraints and constraints can then be determined using iterative methods (Duan et al., 2010; Lesne et al., 2014; Hua and Ma, 2019). However, in reality, no single chromosome configuration will capture all of the interactions present in the contact map, as the contact map is an average over a population of cells. Instead, methods such as simulated annealing are used to find families of optimal chromosome configurations (Rosa and Zimmer, 2014; Junier et al., 2015). The chromosome of *M. pneumoniae* (Trussart et al., 2017) and that of *C. crescentus* (Umbarger et al., 2011) were modeled in this fashion using the Integrative Modeling Platform (Russel et al., 2012). Inverse models have also been built using maximum entropy techniques (Di Pierro et al., 2017; Messelink et al., 2021).

At the start of this study there was no experimental chromosome contact data for Syn3A, so we chose to create a direct model of the chromosome and because we intend to incorporate the chromosome configurations in simulations of whole-cell models using a lattice-based methodology (Roberts et al., 2013), we decided to use a lattice polymer model. There is a rich history of proteins and other polymers being modeled using discrete lattice models (Verdier and Stockmayer, 1962; Heilmann and Rotne, 1982; Lau and Dill, 1989; Madras et al., 1990; Dill et al., 1995). Bacterial chromosome configurations have previously been directly modeled using lattice models (Buenemann and Lenz, 2010; Messelink et al., 2021) and continuous models have been constructed by interpolating between lattice models and relaxing the system (Goodsell et al., 2018). However, none of the models satisfied all three of our requirements of 1) being at the spatial resolution needed to introduce spatial heterogeneities on the 8 nm lattice, 2) self-avoidance, and 3) able to be constrained by the cell boundary and ribosomes. We investigated modifying an existing model, such as Goodsell et al.'s (Goodsell et al., 2018), but found that none were easily extensible.

2.2.1 Growing a Self-Avoiding Polygon Model of Syn3A's Chromosome

We model the circular chromosome of Syn3A as a circular lattice polymer. To account for the volume-exclusion effects, the circular lattice polymer is required to be strictly self-avoiding.

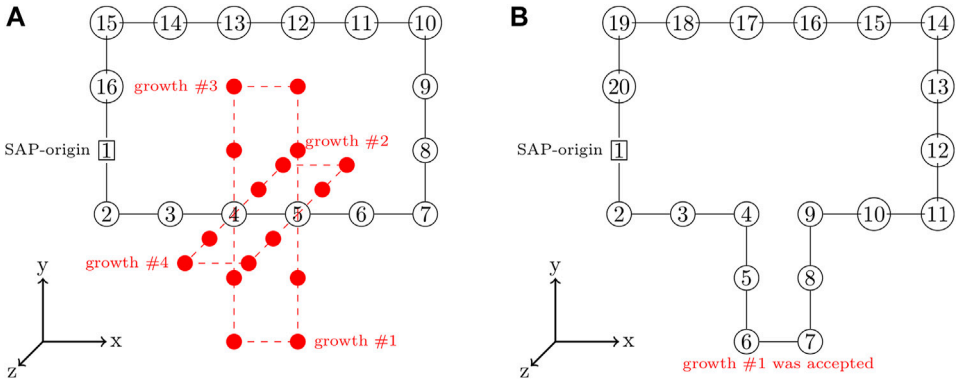


FIGURE 3 | (A)—SAP with the set of proposed growths orthogonal to branch-point at monomers 4 and 5 shown in red. **(B)**—SAP after growth #1 with a size of 4 was accepted and incorporated into the SAP, increasing the SAP size from 16 monomers to 20 monomers.

These circular and self-avoiding configurations of monomers on a lattice are known as self-avoiding polygons (SAPs) and have been previously used to model *E. coli* and *C. crescentus* chromosomes (Buenemann and Lenz, 2010). The SAP model of Syn3A's circular chromosome is defined on a 4 nm cubic lattice and each monomer is represented by a 4 nm × 4 nm × 4 nm cube. These monomers contain cylindrical segments of DNA 4 nm in length, which corresponds to approximately 11.8 bp per monomer. The 543 kbp chromosome of Syn3A is represented by 46,188 of these monomers. The total volume excluded by monomers in the chromosome is 2,956,032 nm^3. At the two extremes, in the small cell with a radius of 201.26 nm, a single chromosome occupies nearly 9% of the cytoplasmic volume, and in the large cell with a radius of 247.42 nm, a single chromosome occupies just below 4% of the cytoplasmic volume.

Mathematically, the SAP configurations within the reconstructed cell geometries are described by the set of monomer coordinates, $\{\boldsymbol{r}_i\}$, on the cubic lattice, that satisfy four different constraints, two SAP constraints, a circularity constraint (g^{circ}) and a self-avoidance constraint (g^{SA}), and two cell geometry constraints, a membrane constraint (h^{mem}) and a ribosome constraint (h^{ribo}). The circularity constraint requires that consecutive monomers are adjacent in the lattice, the self-avoidance constraint requires that no monomers share coordinates, the membrane constraint requires that the monomers remain within the cell, and the ribosome constraints require that the monomers do not intersect any ribosomes. The ribosomes in the 8 nm lattice representation are converted to a 4 nm lattice representation, where they are the same star shape, but now formed from fifty-six 4 nm cubes. These constraints are formulated mathematically using constraint functions that are equal to 1 when the constraints are satisfied and 0 when the constraints are not satisfied. All four of these constraints must be satisfied while growing and moving the SAP. While satisfying the constraints, the configurations are sampled from the canonical ensemble with a Hamiltonian that specifies intrachromosomal interactions, including looping, which will be referred to as restraints. The Hamiltonian is described in **section 2.2.3**.

A SAP with a greater number of monomers can be grown from an existing SAP by severing the bond between a pair of consecutive monomers and adding a closed branch orthogonal to the vector between that pair of monomers (Buenemann and Lenz, 2010; Goodsell et al., 2018). This is done in an unbiased fashion by randomly selecting consecutive pairs of monomers to serve as a branch-point and then randomly proposing growths in the orthogonal directions, an example of proposed growths is depicted in **Figure 3A**. Each proposed growth is only accepted if the resulting SAP satisfies all of the constraints. For example, growth #1 in **Figure 3B** may have been accepted because all of the other proposed growths violated the ribosome constraints. If a satisfactory growth can not be found, then the SAP is moved before searching for growths again. Pseudocode for the SAP growth algorithm is presented in **Algorithm 1**.

Algorithm 1 | SAP-Growth Algorithm.

```
while N < N_target do
    accepted←False
    max_growth_size← min (max_size, N_target − N)
    iter← 0
    while accepted=False and iter<max_iter do
        propose branch-point on SAP
        calculate set of growths orthogonal to branch-point with maximum size of max_growth_size
        growth_size←max_growth_size
        while accepted=False and growth_size > 0 do
            for i=1,2×(dim−1) do
                proposed_growth←growths(i) with size growth_size
                if proposed_growth satisfies constraints then
                    accepted←True
                    incorporate proposed_growth in SAP
                    break
                else
                    iter←iter+1
                end
            end
            growth_size←growth_size−2
        end
    end
    if accepted=False then
        move SAP
    end
end
```

2.2.2 Circularity-Preserving Moves and Proof of Ergodicity

If we start with a valid SAP configuration and then only change the configuration using moves that result in a polymer configuration still satisfying the circularity constraint, then, provided that the moves are ergodic and the new configurations are self-avoiding, we can sample SAP

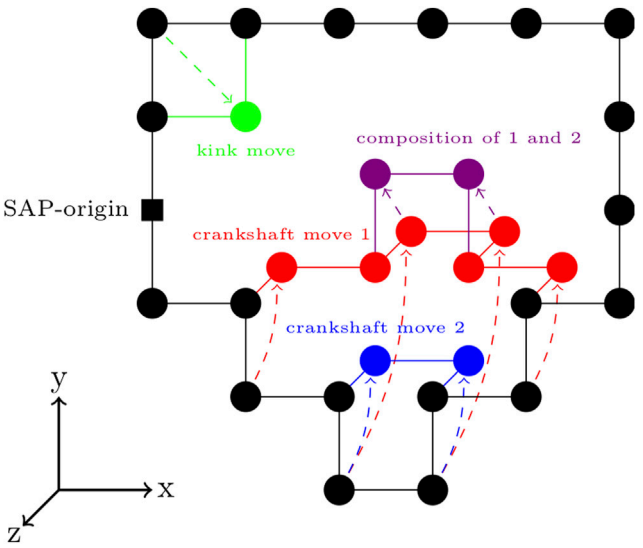


FIGURE 4 | Circularity-preserving moves on a cubic lattice—An example kink move is shown in green. Two example crankshaft moves are shown in red and blue. Following a single enumeration of the set of possible crankshaft moves, multiple crankshaft moves can be made, provided that they are compatible with the crankshaft moves previously sampled from that set of possible crankshaft moves. An example of this is shown by the composition of crankshaft moves 1 and 2 in the purple.

configurations using random sequences of the circularity-preserving moves. The proof of ergodicity follows the proof outlined in (Messelink et al., 2021).

A SAP on a lattice may be represented as a series of displacements along the cubic lattice from a starting location (Messelink et al., 2021). Displacements in the positive and negative Cartesian directions are denoted by $\mathbb{X}^+$, $\mathbb{Y}^+$, and $\mathbb{Z}^+$and $\mathbb{X}^-$, $\mathbb{Y}^-$, and $\mathbb{Z}^-$, respectively. To ensure circularity, the number of positive and negative displacements should be equal for every direction on the lattice, or symbolically, $N_{\mathbb{A}^+} = N_{\mathbb{A}^-}$, where $\mathbb{A} = \mathbb{X}, \mathbb{Y}$, or $\mathbb{Z}$ (Messelink et al., 2021). Traveling counter-clockwise from the origin, the SAP in **Figure 4** is described by the sequence - $(origin)\mathbb{Y}^-\mathbb{X}^+\mathbb{Y}^-\mathbb{X}^+\mathbb{Y}^-\mathbb{X}^+\mathbb{Y}^+\mathbb{X}^+\mathbb{Y}^+\mathbb{X}^+\mathbb{Y}^+\mathbb{Y}^+\mathbb{X}^-\mathbb{X}^-\mathbb{X}^-\mathbb{X}^-\mathbb{Y}^-\mathbb{Y}^- \rightarrow (origin)$.

There are a variety of circularity-preserving moves that can transform the sequence while maintaining the circularity. For our program, we chose an extension of the Verdier-Stockmeyer moveset (Verdier and Stockmayer, 1962; Sokal, 1995) with kink moves and 2 to $N/2$ monomer crankshaft moves. A kink move is the interchange of two symbols in a subsequence $\mathbb{AB} \rightarrow \mathbb{BA}$ (Messelink et al., 2021). The move labeled kink move in **Figure 4** is equivalent to $\mathbb{X}^-\mathbb{Y}^- \rightarrow \mathbb{Y}^-\mathbb{X}^-$. A crankshaft move alters a motif of a specific type. The motif is a subsequence where the monomers at the start and end of the subsequence share two Cartesian coordinates (Messelink et al., 2021). Symbolically, within such a subsequence, $N_{\mathbb{A}^+} \neq N_{\mathbb{A}^-}$, while $N_{\mathbb{B}^+} = N_{\mathbb{B}^-}$ and $N_{\mathbb{C}^+} = N_{\mathbb{C}^-}$. The crankshaft move is then a rotation of magnitude $\pi/2$, π, or $3\pi/2$ about the vector separating the monomers at the start and end of the subsequence, applied to all of the monomers between those two. Generally, the transformation of symbols within the subsequence undergoing a crankshaft move will be $\mathbb{A}^{\pm} \rightarrow \mathbb{A}^{\pm}$, while $\mathbb{B}^{\pm} \rightarrow (\mathbb{C}^{\pm}, \mathbb{B}^{\mp}, \mathbb{C}^{\mp})$ and $\mathbb{C}^{\pm} \rightarrow (\mathbb{B}^{\pm}, \mathbb{C}^{\mp}, \mathbb{B}^{\mp})$. The move labeled crankshaft move 1 in **Figure 4** is equivalent to $\mathbb{Y}^-\mathbb{X}^+\mathbb{Y}^-\mathbb{X}^+\mathbb{Y}^+\mathbb{X}^+\mathbb{Y}^+ \rightarrow \mathbb{Z}^-\mathbb{X}^+\mathbb{Z}^-\mathbb{X}^+\mathbb{Z}^+\mathbb{X}^+\mathbb{Z}^+$.

Starting from a sequence of at least two symbol types satisifying the condition $N_{\mathbb{A}^+} = N_{\mathbb{A}^-}$, where $\mathbb{A} = \mathbb{X}, \mathbb{Y}$, or $\mathbb{Z}$, combining the kink and crankshaft moves can produce any sequence of symbols that also satisfies the condition (Messelink et al., 2021). This result allows for ergodic sampling of sequences, which is equivalent to ergodic sampling of polymer configurations satisfying the circularity constraint. However, the Verdier-Stockmeyer moveset is known to be non-ergodic for self-avoiding walks (SAWs) and SAPs due to the presence of knotted configurations (Madras and Sokal, 1987; Madras et al., 1990) and there is the additional challenge of confinement imposed by the ribosome and the cell boundary constraints. We attempted to mitigate these issues by incorporating the extended crankshaft moves and growing the SAPs to sample configurations that would otherwise be inaccessible by a single SAP being dynamically sampled using a Markov chain Monte Carlo method.

The relative frequencies of the kink moves and crankshaft moves have significant impact on the overall speed of the algorithm and are linked to the ergodicity (Sokal, 1995). The speed of the algorithm can be improved by performing multiple kink or crankshaft moves from a single enumeration of all possible kink or crankshaft moves in the current configuration, respectively. However, following the single enumeration, in addition to satisfying the SAP and spatial constraints, all kink or crankshaft moves performed must be compatible.

The list of possible kink moves are stored as an array of three element vectors of monomer indices, $(i-1, i, i+1)$, where the i-th monomer in the middle will be moved by interchanging two of its coordinates that match with the coordinates of the $i-1$-th and $i+1$-th monomers. After at least one kink move is proposed and accepted, all following kink moves may not have their $i-1$-th or $i+1$-th monomers be one of the middle monomers that was moved in the previously accepted kink moves. Proposed kink moves are then rejected based on this condition. The list of possible crankshaft moves are stored as an array of two element vectors (i, j) of monomer indices, where $i < j$ and i and j are the monomers defining the ends of the subset of the SAP which will be transformed by the crankshaft moves, and an array of two element vectors (d, ω), describing the length of the SAP subset, d, and the direction around the SAP in which the SAP subset is defined, ω. After at least one crankshaft move has been accepted, all following crankshaft moves must have their (i', j') either both belonging to the SAP subset that was moved by the crankshaft move or both not belonging to the SAP subset that was moved. Proposed crankshaft moves are then rejected based on this condition.

Crankshaft moves are the most computationally expensive to both enumerate and sample; however, they cause the fastest change in the configuration. The naive solution to this problem was to assign a frequency at which crankshaft moves were performed, $\eta_{\text{crankshaft}}$, and multiplicities for the number of

Algorithm 2 | SAP-Move Algorithm.

```
for i = 1, n_iter do
    move_type ← U(0, 1)
    if move_type < η_crankshaft^{-1} then
        enumerate crankshaft moves
        n_crankshaft ← u(0, 2g_crankshaft)
        attempt to perform n_crankshaft compatible crankshaft moves
    else
        enumerate kink moves
        n_kink ← u(0, 2g_kink)
        attempt to perform n_kink compatible kink moves
    end
end
```

kink and crankshaft moves that were performed after a single enumeration of kink or reflect moves, g_{kink} and $g_{\text{crankshaft}}$, respectively. These parameters describing the sampling were then manually adjusted. Using this methodology prevents ergodic sampling from ever occurring. This can be illustrated by considering the fact that as long as crankshaft moves are sampled in batches of $g_{\text{crankshaft}}$ every $\eta_{\text{crankshaft}}$ iterations, then unless only a single crankshaft move is possible, there will never be an instance in which a kink move is sampled immediately after a crankshaft move. The inverse case is also true. An alternative is to randomly select the iterations at which crankshaft moves will be enumerated and performed, where the probability is given by $p_{\text{crankshaft}} = \eta^{-1}_{\text{crankshaft}}$. Once the move type is determined using this criteria, randomly sample the number of moves to be performed from a distribution whose mean is equal to the multiplicity of the respective move type. For example, in the case of discrete uniform distributions $n_{\text{kink}} = \mathsf{u}(0, 2g_{\text{kink}})$ and $n_{\text{crankshaft}} = \mathsf{u}(0, 2g_{\text{crankshaft}})$. Now there exists the possibility that any sequence of kink and reflect moves may be sampled. Pseudocode for the SAP movement algorithm is presented in **Algorithm 2**.

2.2.3 Energy Functions and Metropolis-Hastings Sampling

The Hamiltonian for the SAP model of the chromosome has three contributions, a bending energy related to the stiffness of DNA, a nearest-neighbor interaction, and a harmonic interaction acting as a restraint to recreate the effect of DNA looping.

$$\mathcal{H} = \mathcal{H}^{\text{bend}} + \mathcal{H}^{\text{n.n.}} + \mathcal{H}^{\text{loops}} \tag{2}$$

The contribution to the Hamiltonian due to the bending stiffness of linear DNA is

$$\mathcal{H}^{\text{bend}}(\{\boldsymbol{r}_i\}) = -\kappa \sum_{i=2}^{N-1} (\boldsymbol{r}_{i+1} - \boldsymbol{r}_i) \cdot (\boldsymbol{r}_i - \boldsymbol{r}_{i-1}) \tag{3}$$

and is parameterized by the bending energy per unit length squared, κ. This Hamiltonian incurs an energy penalty for every bend in the lattice polymer and can be used to model the stiffness of a polymer, a quantity often characterized by the persistence length. One interpretation of the persistence length, l_p, is the constant describing the exponential rate at which the polymer orientations become decorrelated (Brinkers et al., 2009; Hsu and Binder, 2012; Zhang et al., 2019)

$$\frac{\langle (\boldsymbol{r}_{i+s+1} - \boldsymbol{r}_{i+s}) \cdot (\boldsymbol{r}_i - \boldsymbol{r}_{i-1}) \rangle_{\text{mono.}}}{l^2} = \exp\left(-sl/l_p\right), \tag{4}$$

where $\langle f(\boldsymbol{r}_i) \rangle_{\text{mono}}$ is the average over the N monomers in the configuration and l is the lattice size. Consider the case of a SAW on a cubic lattice, in which the lattice polymer can become immediately decorrelated, thus consider the case when $s = 1$

$$\begin{aligned} \langle (\boldsymbol{r}_{i+1} - \boldsymbol{r}_i) \cdot (\boldsymbol{r}_i - \boldsymbol{r}_{i-1}) \rangle_{\text{mono.}} &= \frac{1}{N-2} \left[\sum_{i=2}^{N-1} (\boldsymbol{r}_{i+1} - \boldsymbol{r}_i) \cdot (\boldsymbol{r}_i - \boldsymbol{r}_{i-1}) \right] \\ &= -\frac{\mathcal{H}^{\text{bend}}(\{\boldsymbol{r}_i\})}{(N-2)\kappa} \end{aligned} \tag{5}$$

leading to an equation with the bending Hamiltonian parameterized by κ. Assuming the lattice polymer is in thermal equilibrium at inverse temperature $\beta = 1/k_B T$, we can take a thermal average of this equation

$$-\frac{\langle \mathcal{H}^{\text{bend}}(\{\boldsymbol{r}_i\}) \rangle}{(N-2)\kappa l^2} = \exp\left(-l/l_p\right) \tag{6}$$

and κ can be calculated by solving this root-finding problem through Monte Carlo sampling of SAW configurations using Wang-Landau sampling (Wang and Landau, 2001). In this study, the value of κl^2 ($3.872 k_B T$) was estimated using the exact solution for a non-reversal random walk and the consensus persistence length for DNA of 50 nm (Vologodskii et al., 1992; Manning, 2006; Brinkers et al., 2009; Geggier et al., 2010; Mantelli et al., 2011).

$$\kappa = -\frac{1}{\beta l^2} \log\left[\frac{e^{l/l_p} - 1}{4} \right] \tag{7}$$

The contribution to the Hamiltonian due to pairwise nearest-neighbor interactions is

$$\mathcal{H}^{\text{n.n.}} = \epsilon \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \delta^K\left(l - \left|\mathbf{r}_i - \mathbf{r}_j\right|\right) \tag{8}$$

and was used to tune the excluded-volume effects of DNA ($\epsilon = k_B T$). Lastly, the contribution to the Hamiltonian when looping restraints are imposed is

$$\mathcal{H}^{\text{loops}}(\{\boldsymbol{r}_i\}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} k_{ij} \left|\boldsymbol{r}_i - \boldsymbol{r}_j\right|^2 \tag{9}$$

These pairwise harmonic interactions were used to create looping between portions of chromosome bound by SMC proteins ($k_{ij} l^2 = 10{,}000 k_B T$).

A Markov chain Monte Carlo algorithm (Metropolis et al., 1953; Hastings, 1970) was used to sample configurations governed by this Hamiltonian from the canonical ensemble. We use the Metropolis criterion, $A(\{\boldsymbol{r}_i'\}, \{\boldsymbol{r}_i\}) = \min(1, P(\{\boldsymbol{r}_i'\})/P(\{\boldsymbol{r}_i\}))$, (Metropolis et al., 1953), for the acceptance probability of moving from the current configuration, $\{\boldsymbol{r}_i\}$, to the proposed configuration, $\{\boldsymbol{r}_i'\}$. The probability of a configuration satisfying the SAP constraints (g^{circ} and g^{SA}) and geometric constraints (h^{mem} and h^{ribo}) is

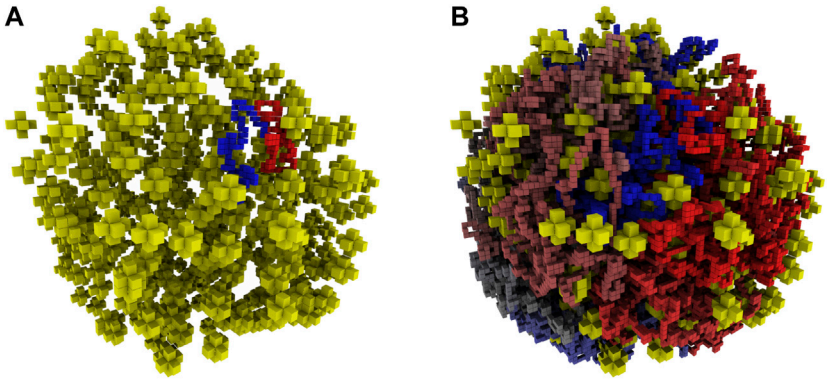


FIGURE 5 | (A)—Reconstructed ribosome distribution in the small cell. The 100 monomers on either side of the origin are shown in red and blue. Ribosomes are depicted as yellow stars in the 8 nm lattice representation. **(B)**—Complete chromosome configuration generated on the 4 nm lattice within the reconstructed architecture of the small cell. The circular chromosome is colored starting at the origin as red to grey to blue, before returning to the origin where blue and red meet.

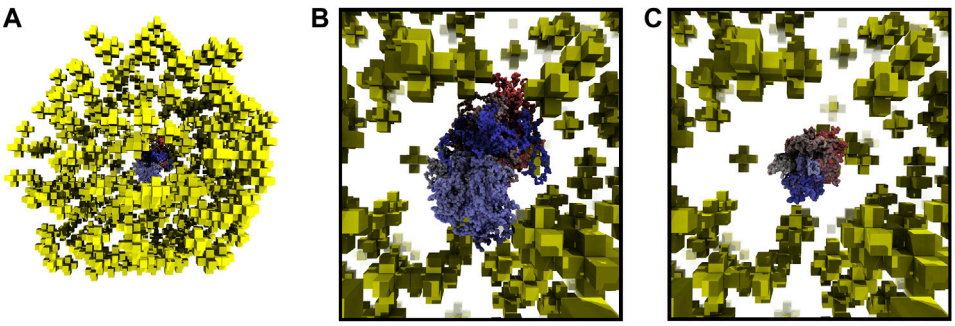


FIGURE 6 | (A)—Centroid of 30 chromosome configurations is shown within the ribosome distribution. The same color scheme for centroid is used as for the chromosome in **Figure 5**. **(B)**—Magnified view of centroid in 6A calculated from 30 configurations, radius of gyration is 24.93 nm. **(C)**—Magnified view of centroid calculated from 90 configurations, radius of gyration is 11.99 nm.

$$P(\{\boldsymbol{r}_i\}) = \frac{1}{\mathcal{Z}}\left[g^{\mathrm{circ}}(\{\boldsymbol{r}_i\})g^{\mathrm{SA}}(\{\boldsymbol{r}_i\})\right] \times \left[h^{\mathrm{mem}}(\{\boldsymbol{r}_i\}, R)h^{\mathrm{ribo}}(\{\boldsymbol{r}_i\}, \{\boldsymbol{\rho}_i\})\right] \times \exp\left[-\beta\mathcal{H}(\{\boldsymbol{r}_i\})\right], \tag{10}$$

where $\mathcal{Z}$ is the canonical partition function of the system found by summing over all possible configurations of N monomers on a cubic lattice. Assuming the current configuration, $\{\boldsymbol{r}_i\}$, and the proposed configuration, $\{\boldsymbol{r}_i'\}$, always satisfy the circularity constraint because they are generated from sequences of circularity-preserving moves, then the ratio of probabilities is

$$\frac{P(\{\boldsymbol{r}_i'\})}{P(\{\boldsymbol{r}_i\})} = \frac{g^{\mathrm{SA}}(\{\boldsymbol{r}_i'\})h^{\mathrm{mem}}(\{\boldsymbol{r}_i'\}, R)h^{\mathrm{ribo}}(\{\boldsymbol{r}_i'\}, \{\boldsymbol{\rho}_i\})}{g^{\mathrm{SA}}(\{\boldsymbol{r}_i\})h^{\mathrm{mem}}(\{\boldsymbol{r}_i\}, R)h^{\mathrm{ribo}}(\{\boldsymbol{r}_i\}, \{\boldsymbol{\rho}_i\})} \times \exp\left(-\beta\left[\mathcal{H}(\{\boldsymbol{r}_i'\}) - \mathcal{H}(\{\boldsymbol{r}_i\})\right]\right). \tag{11}$$

Additionally, if the proposed configuration satisfies the self-avoidance and geometric constraints, which can be determined without evaluating energy changes, then the acceptance probability given by the Metropolis criterion, $A(\{\boldsymbol{r}_i'\}, \{\boldsymbol{r}_i\}) = \min(1, e^{-\beta\Delta E})$, is simply a function of the energy difference, $\Delta E = \mathcal{H}(\{\boldsymbol{r}_i'\}) - \mathcal{H}(\{\boldsymbol{r}_i\})$, and the sampling favors low-energy configurations that better agree with the stiffness of DNA, the excluded-volume effects, and the DNA-looping restraints.

2.2.4 Summary of Complete Algorithm for Generating Chromosome Configurations

The final algorithm generated chromosome configurations by alternating cycles of growing and moving the SAP configurations to relax the newly grown portion. Pseudocode for the final algorithm is presented in **Algorithm 3**. In an early implementation, a single relaxation occurred after the growth was completed, but it was found that the alternating cycles of growth and relaxation were required because the combined effects of confinement and the exponentially increasing attrition rate due to violations of the self-avoidance constraint became overwhelming as the SAP grew larger. The relative frequencies and durations of these alternating growth and relaxation cycles were chosen empirically to maximize the speed of generating relaxed configurations of the complete chromosome. An example of how the alternating growth and

Algorithm 3 | Complete Algorithm For Generating Chromosome Configurations.

while $N < N_{final}$ **do**
  $N_{\text{target}} \leftarrow \tau(N, N_{\text{final}}, \boldsymbol{\alpha})$
  grow N-monomer SAP until it reaches N_{target} monomers - *see SAP-growth algorithm* 1
  $n_{\text{move}} \leftarrow \sigma(N, N_{\text{final}}, \boldsymbol{\gamma})$
  move N-monomer SAP for n_{move} iterations - *see SAP-move algorithm* 2
end

relaxation cycles affect the total energy over the course of a simulation for a test case with 5,000 monomers is presented in **Supplementary Figure S4**. While this procedure more rapidly relaxes the system, the exponentially increasing attrition rate of rejected moves prevents us from definitively stating that we reach equilibrium in the system with 46,188 monomers. The relative frequencies and durations are described by functions, τ and σ, both are dependent on the current number of monomers, *N*, and separately depend on empirical parameter vectors, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, respectively. The algorithm was implemented in Fortran 90 and a single chromosome configuration of 46,188 monomers can be generated in approximately 12–14 h on a single CPU core at 3.5 GHz. The algorithm is embarrassingly parallel and the program uses OpenMP to generate multiple configurations simultaneously. An example reconstructed cell architecture is shown in **Figure 5A** and the resulting constrained chromosome configuration is shown in **Figure 5B**.

Starting with fixed ribosome positions and cell orientation from the cryo-ET, we initialize the configurations by randomly placing a circular fragment of the chromosome and then independently generate hundreds of chromosomes within an otherwise identical cell. To test if the monomers along the chromosome are identically distributed within the cell, we calculate the centroid of the ensemble of chromosome configurations. The monomer coordinates of the centroid are the ensemble averages of the monomer coordinates in the chromosome configurations. The center of mass of a sphere is at its center, thus we expect the centroid of the ensemble of chromosome configurations to be approximately located at the center of the spherical cell. We find the centroid of 30 configurations to be located in the center of the cell, as shown in **Figure 6A**. Furthermore, if the number of identically distributed chromosome configurations is increased, we expect the centroid to collapse to the center, which we quantify with its radius of gyration. The centroid of 30 configurations in **Figure 6B** has a radius of gyration of 24.93 nm and it is reduced to 11.99 nm when the centroid is calculated from 90 configurations, as shown in **Figure 6C**.

Other bacteria that are not genetically-minimal have additional regulatory systems used to control their chromosome organization, such as attachment organelles and parABS systems. Due to these regulatory systems, their chromosomes show consistent configurations that correlate the genomic position with the internal structure of the cell (Umbarger et al., 2011; Marbouty et al., 2015; Trussart et al., 2017) and this is reflected in their centroids. For example, in a model of *M. pneumoniae*'s chromosome, Trussart et al. saw a consistent alignment and interweaving of the two chromosome arms of the centroid (Trussart et al., 2017). As a comparison, we tested fixing the origin of our chromosome at the membrane and found that the centroid had a consistent alignment of the two chromosome arms at the fixed origin and monomers near the origin were found near the membrane (data not shown). Since there are no interactions correlating the genomic position and the internal structure of the cell, we compared the average radius of gyration for chromosome configurations generated in the small cell with and without ribosomes present to test the excluded volume effect of ribosomes. The average radius of gyration without ribosomes was 145.40 nm and was 133.59 nm when ribosomes were present. We also tested the effect of further increasing the number of ribosomes by randomly placing 497 ribosomes in addition to the 503 from the tomogram in the small cell and found that the average radius of gyration further decreased to 124.29 nm. We attribute this reduction in the average radius of gyration to the additional confinement caused by the volume exclusion of the ribosomes.

2.3 3C-Seq Library Preparation

JCVI-syn3A chromosome contact maps were prepared with 3C-Seq (Lioy and Boccard, 2018), a chromosome conformation capture technique reminiscent of Hi-C (Crémazy et al., 2018). The protocols differ in that following the restriction digestion of the fixed chromosome, restriction fragment ends are not filled-in with biotin-labelled nucleotides in 3C-Seq (Crémazy et al., 2018; Lioy and Boccard, 2018). The modification reduces the cost of chromosome conformation capture in prokaryotes since the requirement for biotin-labelled nucleotides is alleviated. 3C-Seq increases the diversity of restriction enzyme options available for library preparation from only enzymes that generate 5′-overhangs that can be filled-in by the Klenow fragment, to include enzymes that generate 3′-overhangs, and blunt-ends. Furthermore, sticky ends generated by restriction digestion are not "blunted" in 3C-Seq, increasing ligation efficiency since sticky-end ligation occurs more efficiently than blunt-end ligation. In addition, the absence of biotin at restriction fragment ends eliminates the requirement of removing biotin-labels from unligated ends, DNA purification following biotin removal, and enrichment of biotin-labelled ligation junctions, effectively, reducing the library preparation time by at least 30%.

Syn3A was cultured to stationary phase in 25 ml of SP4-KO medium in a 50 ml conical tube at 37°C. The cells were fixed with a final concentration of 1% formaldehyde (Sigma-Aldrich) at 25°C for 30 min and 4°C for a further 30 min. The reaction was quenched with 0.125 M glycine (Sigma-Aldrich) for 15 min at 4°C. The fixed cells were collected by centrifugation and washed twice with 1X HE pH 8.0 [10 mM HEPES (Sigma-Aldrich), 1 mM EDTA (Sigma-Aldrich)]. The cell pellet was flash-frozen with liquid nitrogen in a 1.5 ml low-binding microfuge tube and stored at −80°C until use. Fixed Syn3A cells were resuspended in 100 µl of 1X HE pH 8.0 and mechanically sheared with 0.5 mm glass beads (Sigma-Aldrich) using a vortex mixer. Membranous structures in the lysate were solubilised with 0.5% SDS (Sigma-Aldrich) for 15 min at 37°C in a Thermomixer® (Eppendorf) with shaking at 1,000 rpm. SDS was quenched with 1% Triton X-100 (Sigma-Aldrich) in 1X CutSmart buffer (NEB) for 15 min at 37°C in a Thermomixer® (Eppendorf) with shaking at 1,000 rpm. The extracted chromatin was digested with 100 U of NlaIII (NEB) for 3 h at 37°C. The reaction was

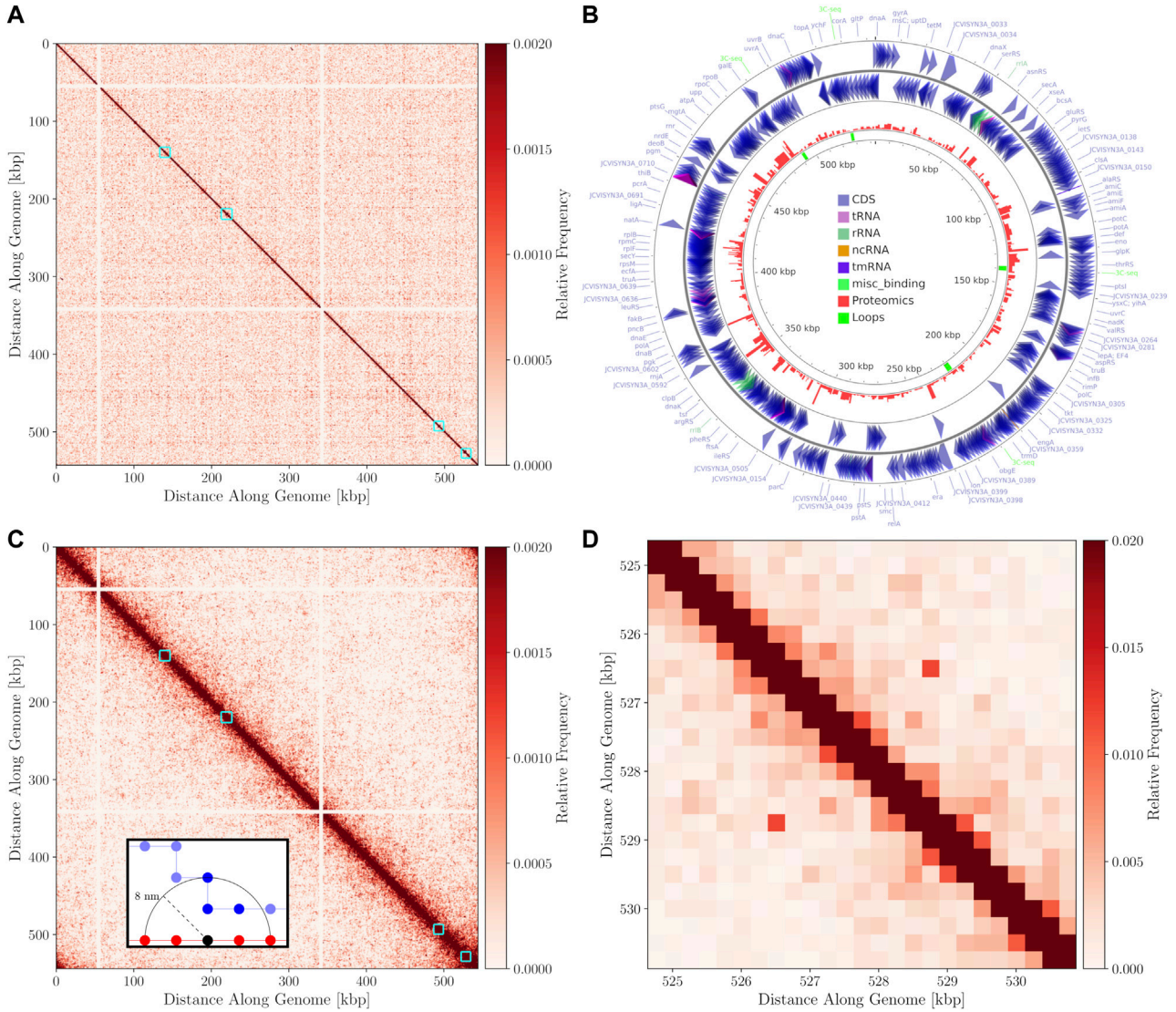


FIGURE 7 | (A)—3C-Seq contact map at 1,000 bp resolution with the color-scale adjusted to make weak secondary features along the diagonal more apparent. The four manually annotated loops listed in **Table 2** are indicated with cyan boxes. **(B)**—Circular chromosome of Syn3A with features shown as arrows and arcs around the perimeter - constructed using CGview (Petkau et al., 2010). The proteomics of a 400 nm Syn3A cell (Breuer et al., 2019) are plotted in red around the middle ring and the innermost ring contains the annotated loops in green. **(C)**—*In silico* contact map resulting from 150 configurations with looping interactions added at the positions of the manually annotated loops in **Table 2**. A locus size of 1,000 bp was used to match the 3C-Seq map. The interactions at the ribosomal RNA operons have been removed from the map to enhance visual clarity. Cyan squares are again used to indicate regions containing the loops. **(D)**—Magnified view of the region within the cyan square containing the fourth loop in the *in silico* contact map. The map was recalculated at a resolution of 250 bp and the maximum of the color-scale was increased to better resolve the characteristic signature of a loop.

terminated with 0.5% SDS (Sigma-Aldrich) for 20 min at 37°C. The digested chromatin was centrifuged at 20,000 xg for 1 h at 4°C. The supernatant was removed and the gel-like pellet was dissolved in 200 µl of nuclease-free water (ThermoFisher Scientific). The DNA concentration of the dissolved chromatin was determined using the Qubit® HS dsDNA assay kit (ThermoFisher Scientific) and the Qubit® fluorometer (ThermoFisher Scientific). 3 µg of DNA was used for ligation in 1X T4 DNA ligase buffer (NEB) supplemented with 100 µg/ml BSA (NEB) in a final volume of 1,000 µl. The reaction was carried out with 4000 CEU of T4 DNA ligase (NEB) at 16°C for 16 h and 25°C for 1 h. Ligation was terminated with 10 mM EDTA pH 8.0 (usb Corporation). Ligated DNA (the 3C library) was extracted twice with 25:24:1 phenol:chloroform:isoamyl alcohol (Sigma-Aldrich) and once with chloroform (Sigma-Aldrich). The library was precipitated with 0.1 × 1.0 M NaOAc (Sigma-Aldrich) pH 8.0, 0.025 × 5 mg/ml glycogen (Invitrogen), and 2.5 × 100% ethanol (Sigma-Aldrich) at -20°

TABLE 2 | Loops inferred from 3C-Seq library of Syn3A. The gene annotations and locus tags are those in the NCBI entry for Syn3A's genome (https://www.ncbi.nlm.nih.gov/nuccore/CP016816.2) and the locus tags are abbreviated to only the 4-digit number.

Start (bp)	Stop (bp)	Length (bp)	% *in silico*	Genes encompassed	Proteomics
138,324	141,557	3,233	8.7	pdhC	182
—				lpdA	182
217,523	221,263	3,740	12.7	ywjA (0371)	175
—				ywjA (0372)	65
491,413	493,784	2,371	16.7	lgt	55
—				trx	100
526,669	528,859	2,190	20.7	0877	9
—				0878	7

C overnight. Precipitated DNA was pelleted by centrifugation and the pellet washed twice with 70% ethanol (Sigma-Aldrich). The pellet was air-dried and dissolved in 50.0 µl of 10 mM Tris (Sigma-Aldrich) pH 8.0. The 3C library was purified with 3X KAPA HyperPure beads (KAPA Biosystems) and eluted in 20.0 µl of 10 mM Tris (Sigma-Aldrich) pH 8.0.3C-Seq libraries for next-generation sequencing were prepared using the KAPA HyperPlus Kit (KAPA Biosystems) according to the manufacturer's protocol. 3C-Seq libraries were sequenced on an Illumina® platform.

3 RESULTS

3.1 3C-Seq and *in Silico* Contact Maps

The 3C-Seq library prepared using the restriction enzyme NlaIII had a total of 1,819,715 reads that were mapped at a resolution of 1,000 bp. A histogram of restriction digestion fragment sizes and distribution NlaIII cut sites in Syn3A's chromosome are presented in **Supplementary Figures S5,S6**, respectively. The contact map was normalized to be a doubly-stochastic matrix using the matrix-balancing procedure of Knight and Ruiz (Knight and Ruiz, 2012; Rao et al., 2014) and is shown in **Figure 7A**. The chromosome contact map shows a primary diagonal of high interaction frequency that reflects the physical proximity of loci that lie close to each other along the primary sequence of the DNA polymer. A secondary diagonal cannot be detected implying the absence of inter-arm interactions along the chromosome. The absence of a secondary diagonal is in contrast to the chromosome contact maps of *M. pneumoniae* (Trussart et al., 2017), *B. subtilis* (Marbouty et al., 2015), and *C. crescentus* (Le et al., 2013; Tran et al., 2017). Notably, there are two regions of the chromosome that are devoid of interactions, these regions correspond to the two identical ribosomal RNA operons in Syn3A and can be seen in **Figure 7B**. No interactions were assigned to these regions as sequencing reads arising from either copy could not be distinguished. There are smaller secondary features along the diagonal that we interpret to be regions of high interaction due to looping. However, as this is a preliminary map with a low read depth and signal-to-noise ratio, chromosome architecture cannot be reliably interpreted and standard loop and chromosome interaction domain (CID) annotation software (Durand et al., 2016b) was unable to reliably process the map. Upon visual inspection at a resolution of 250 bp, the map shows four interactions, with distinct signatures reminiscent of loops (Fudenberg et al., 2016). Snapshots of the four interactions at 250 bp resolution in Juicebox (Durand et al., 2016a) are shown in **Supplementary Figure S7**. We infer the end points of these loops to be such that they fully encompass genes in the corresponding regions of the chromosome. The positions of these manually annotated loops, the genes they encompass, and the corresponding proteomics are presented in **Table 2**, and the loop locations within the contact map can be seen in **Figure 7A**.

By comparing the annotated loops to the proteomics (Breuer et al., 2019), as shown in **Figure 7B**, we can investigate correlations between the relative expression levels and the locations of the loops. For reference, the average proteomics count in Syn3A is approximately 180 (Breuer et al., 2019). We will refer to the loops according to their order along the genome. The first loop encompasses the genes *pdhC* and *lpdA*, which respectively code for the E2 and E3 subunits of the PDH complex. These genes have identical proteomics counts and lower expression levels than the genes surrounding them. Upstream are genes coding for enzymes in the main pathway of the central metabolism in Syn3A (Breuer et al., 2019) and downstream are genes coding for components of the PTS system, another essential part of the central metabolism. The second loop encompasses the genes *ywjA* (0371) and *ywjA* (0372), which code for the two subunits of the flippase. This is the longest loop and the two genes within it have the greatest disparity in expression levels. The third loop encompasses the genes *lgt* and *trx*, which code for lipoprotein diacylglyceryl transferase and thioredoxin reductase, respectively. The proteomics counts of both proteins coded by these genes are lower than average, as are those of the genes immediately downstream. However, less than 10 kbp upstream is an operon for ribosomal proteins, which contains some of the most highly-expressed genes in Syn3A's genome (Breuer et al., 2019). The fourth loop encompasses genes that code for two uncharacterized proteins, JCVISYN3A_0877 and JCVISYN3A_0878, both of which have very low proteomics counts. The expression levels of the nearby genes are similarly low. Our most consistent findings are twofold. First, the loops are all between 2 and 4 kbp in length. Second, the loops often contain genes with common expression levels.

The chromosome configurations on the 4 nm lattice, with 11.8 bp monomers, enable the calculation for contact matrices at any resolution greater than 11.8 bp per locus. Equally-sized contiguous regions of the chromosome can be classified as loci and the pairwise interactions between the loci counted according to the relative pairwise distances between monomers belonging to the loci. In the foreground of **Figure 7C** is a representative example of the interaction counting. When counting the total number of interactions between the red and blue loci using an arbitrary threshold distance indicated by the dashed line, the black monomer in the red loci contributes three interactions to the total interaction count between the loci. Due to a relative scarcity of chromosome models at a similar resolution in terms of bp per monomer and uncertainty about what proteins are involved in protein-DNA formaldehyde cross-linking (Dekker et al., 2002; van Berkum et al., 2010), the distance for assessing interactions can be chosen from a minimum of 4 nm corresponding to lattice spacing to a maximum of 50 nm. The maximum distance corresponds to the length of SMC proteins, which is the maximal distance spanned by a nucleoid-associated protein in Syn3A (Diebold-Durand et al., 2017; Marko et al., 2019; Ryu et al., 2021). We selected a contact radius of 8 nm because it is an integer multiple of our lattice spacing and the resulting maps show the best agreement with 3C-Seq map. This distance metric can be used alongside any locus size converted to units of monomers to generate contact matrices for ensembles of computationally generated chromosome configurations.

Contact maps were calculated for chromosome configurations within the small cell. A locus size of 1,000 bp was chosen to match the resolution of the 3C-Seq chromosome contact map. As was the case for the experimental contact map, the contact map was normalized to be a doubly-stochastic matrix using a matrix-balancing procedure. Unfortunately, the precision of the *in silico* contact maps is limited by the number of chromosome configurations used to calculate the ensemble-averaged interaction frequencies, a number many orders of magnitude lower than the number of cells in typical 3C-Seq experiments.

In the absence of a sequence-specific system, such as the parABS system, dictating the global structure of the chromosome and promoting inter-arm interactions, we decided to explore a test case of introducing looping interactions at the positions of the manually annotated loops to test the efficacy of our model. We consider a loop to be successfully formed if the monomers at the endpoints are separated by less than 16 nm, the percentage of configurations with successful loop formation are shown for each loop in **Table 2**. One or more loops were formed in 75 of the 150 configurations. As expected, decreasing the length of the loop increased the probability that it was successfully formed. The *in silico* contact map generated from 150 configurations in the small cell is presented in **Figure 7C**. The contact map shows a single diagonal in **Figure 7C**, which is consistent with the 3C-Seq contact map and indicates that the majority of interactions are self-interactions within loci or interactions between neighboring loci. The strongest signal characteristic of a loop was observed for the fourth loop, which is the shortest, and **Figure 7D** shows a magnified view of the surrounding region in the contact map.

Using the 3C-Seq map, we plotted the interaction frequency as a function of genomic distance and observed a plateau after the initial decrease in interaction frequency. We fit a power law of the form $P(x) \propto x^s$ to two regimes within the strictly-decreasing region before the plateau, i.e. the region extending from self-interactions along the diagonal to interactions with loci at distances less than or equal to 10 kbp away, and found a range of exponents ($s = -0.519$ to $s = -2.210$). We repeated this calculation for the *in silico* map and found a narrower range of exponents than in the 3C-Seq case ($s = -0.720$ to $s = -1.132$).

Plots of the two datasets and their contact laws are presented in **Supplementary Figure S8**. In both cases, the steepest rate of change and largest exponent was found in the region whose lower limit corresponded to interactions of loci separated by 1 kbp. For the *in silico* case, the calculated values of s are in closer agreement to the value expected when confined homopolymers are organized as fractal globules ($s = -1$), with clearly defined territories caused by topological constraints, rather than equilibrium globules ($s = -1.5$) (Lua et al., 2004; Lieberman-Aiden et al., 2009; Mirny, 2011; Rosa and Zimmer, 2014; Sanborn et al., 2015). The organization of the chromosome into territories can be observed for the *in silico* case in **Figure 5B** as the separation into distinct colored regions.

The plateau is more pronounced in the 3C-Seq case, than the *in silico* case, and all interactions in the 3C-Seq dataset are nearly equally probable at genomic distances greater than 100 kbp. While the plateau in the 3C-Seq dataset is a characteristic of equilibrium globules (Lieberman-Aiden et al., 2009; Mirny, 2011) and some mathematically-predicted fractal globules, such as the Sierpinski triangle and inside-out Hilbert curves (Sanborn et al., 2015), we are unable to infer a topological state of the Syn3A chromosomes sampled using 3C-Seq because of the significant variations in the exponents of the power law and the sensitivity to the regime chosen for fitting. It is possible that these variations and the steep drop off are a consequence of the low coverage in the preliminary 3C-Seq map or reflect biologically relevant levels of organization.

3.2 Spatial Model of JCVI-syn3A

Computational modeling of spatially-resolved kinetics in Syn3A is done by simulating the reaction-diffusion master equation (RDME) in Lattice Microbes (LM) (Roberts et al., 2013; Hallock et al., 2014; Earnest et al., 2017, 2018; Bianchi et al., 2018) using a stochastic simulation algorithm. When using the RDME, physical space is discretized into a cubic lattice representation. The size of the cubic lattice dictates both the resolution of the spatial modeling and the maximum allowable timestep when modeling the kinetics, smaller lattice sizes reduce the maximum allowable timestep. A lattice size of 8 nm was chosen as an acceptable compromise between creating a high-resolution spatial model of Syn3A, while permitting simulations over biologically-relevant time scales. Each of these 8 nm lattice sites can contain a maximum of sixteen particles. Previous work on combining LM simulations and tomogram data, directly reconstructed cell architectures in LM (Earnest et al., 2017). Unfortunately, as discussed earlier, the tomograms do not

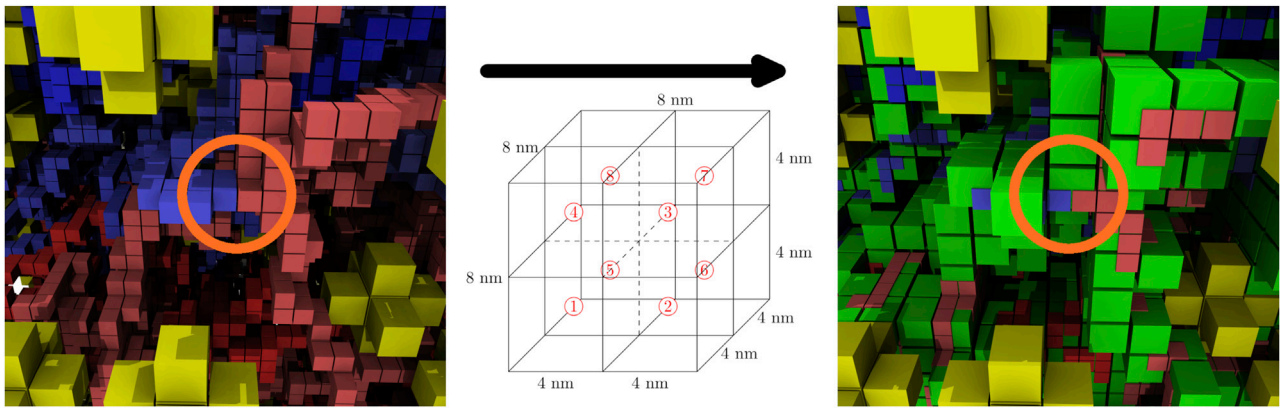


FIGURE 8 | Coarse-graining, before and after—The coarse graining procedure localizes up to eight effective monomers in the 4 nm chromosome configurations (same color scheme as **Figure 5**) within the 8 nm chromosome lattice sites (green). Circled in orange there is an example of genomically-distant regions being localized within the same chromosome site.

show well-defined DNA strands. Instead, chromosome configurations consistent with the ribosome distributions observed in the tomograms are generated using our method, and those are used for the spatial models.

We have also used the spherical cell architecture reconstructed from the tomograms to predict the number of ribosomes involved in polysomes, the number of ribosomes at or near the membrane, and the number of ribosomes close enough to DNA to form an expressome. To predict the number of ribosomes involved in possible polysomes, we calculate the pairwise distances between all ribosome pairs in the spherical cell. Annotating any pair within a center-to-center distance of 22 nm to be in a possible polysome, as was experimentally measured in *E. coli* (Brandt et al., 2009), we calculate 194, approximately 39%, of the 503 ribosomes from the first template matching method (approach 1) are involved in possible polysomes. In the second template matching method with 3D classification (approach 2), this number increases to 373, approximately 55%, of the 684 ribosomes in possible polysomes. If we instead use a center-to-center distance of 18 nm such that the ribosomes are almost in contact, we find that 125, approximately 25%, of the 503 ribosomes from approach 1 are involved in possible polysomes. In approach 2, we find that 274, approximately 40%, of the 684 ribosomes are in possible polysomes using the 18 nm distance. In cryo-ET of a closely related organism *M. pneumoniae*, sub-tomogram averaging over many cells predicted that an average of 16.4% of ribosomes are involved in polysomes (O'Reilly et al., 2020). The discrepancies arise from several factors: First, *M. pneumoniae* only has 300 ribosomes per cell (Seybert et al., 2006; Yus et al., 2009; O'Reilly et al., 2020) and has a larger volume than Syn3A (Kühner et al., 2009), so it is not unreasonable that the higher ribosome density in Syn3A results in a larger fraction ribosomes involved in polysomes. Second, the method to define polysomes in (O'Reilly et al., 2020) exclusively looks at ribosomes that are in very close proximity and oriented so that the mRNA exit channel of one ribosome aligns with the mRNA entry of the next. In cells, polysomes are likely to be more relaxed than this configuration. Our predicted fractions of ribosomes involved in polysomes are all lower than the 70% observed in fast-growing *E. coli* using absorption spectroscopy (Phillips et al., 1969; Forchhammer and Lindahl, 1971).

To estimate the number of ribosomes on or near the membrane, we calculate the number of ribosomes within a cytoplasmic shell directly inside the membrane. We annotate a ribosome as being within the cytoplasmic shell if its center is within the shell. Using a 10 nm thick shell, the approximate radius of a ribosome, we find that 53, approximately 10%, of the 503 ribosomes from approach 1 are near the membrane. We find a similar number of ribosomes within the same distance in approach 2, 60 (9%) of the 684 ribosomes. If we extend the shell to 20 nm thick, we find that 122, approximately 24%, of the 503 ribosomes in approach 1 are near the membrane. In approach 2, we found 136, or 20%, of the 684 ribosomes are within 20 nm of the membrane. The range of our calculated fractions agrees with the observed 15% of ribosomes being membrane-bound in cryo-ET of *S. melliferum* (Ortiz et al., 2006).

Expressomes are macromolecular complexes of RNA polymerases (RNAPs) and ribosomes that couple transcription and translation, they were first identified in *E. coli* (Kohler et al., 2017). In *M. pneumoniae* a maximum of 19% of ribosomes have been identified to be in an expressome complex in which NusA and NusG help to connect or direct the mRNA from production by RNAP to the ribosome (O'Reilly et al., 2020). Given the proteomics counts of RNAP, NusA, and NusG within a 400 nm cell of 187, 238, and 464 respectively (Breuer et al., 2019), the possibility of expressome complexes emerging in the whole-cell model is certainly possible. Using the 4 nm lattice representation, we searched for possible expressomes by counting chromosome monomers directly adjacent to the star-shaped ribosomes. We find that on average there are 106 of the 503 ribosomes in approach 1 with a DNA monomer directly adjacent, a fraction of approximately 21%. From ribosomes identified in approach 2, we found that 127 of the 684 ribosomes, roughly 19%, were directly adjacent to the DNA on

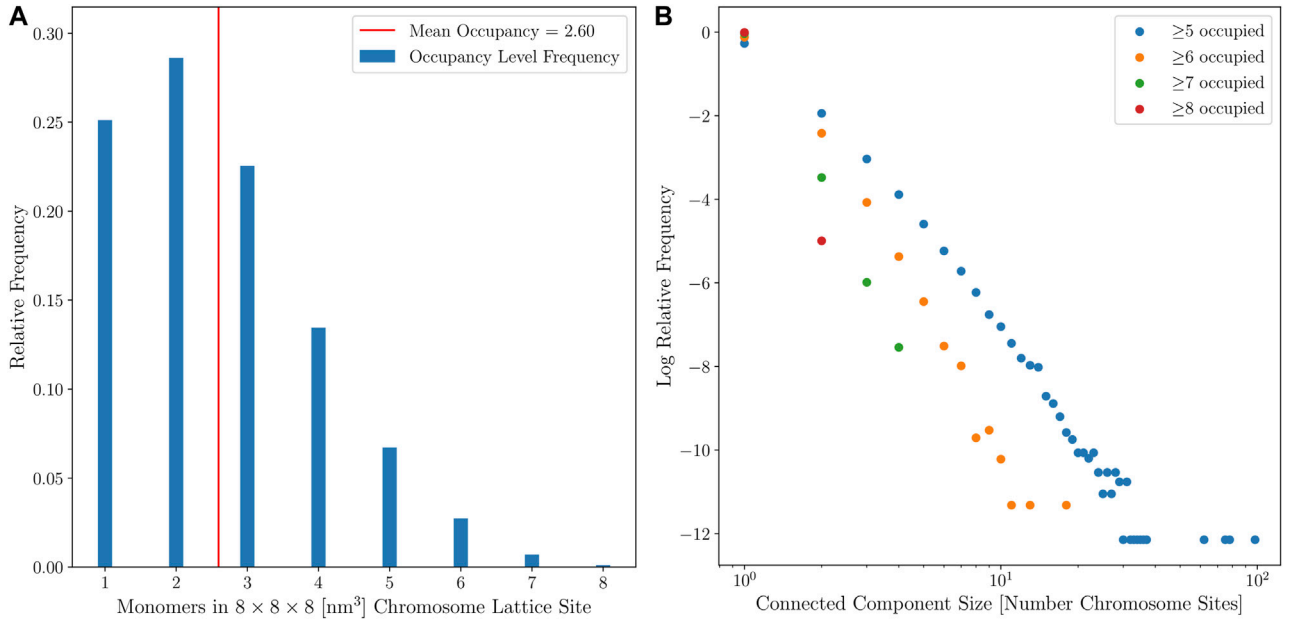


FIGURE 9 | (A)—Relative frequencies of chromosome site monomer occupancies in the small cell from 150 configurations with no looping interactions. **(B)**—Distribution of connected-component sizes for different monomer occupancy thresholds for the small cell from 150 configurations with no looping interactions.

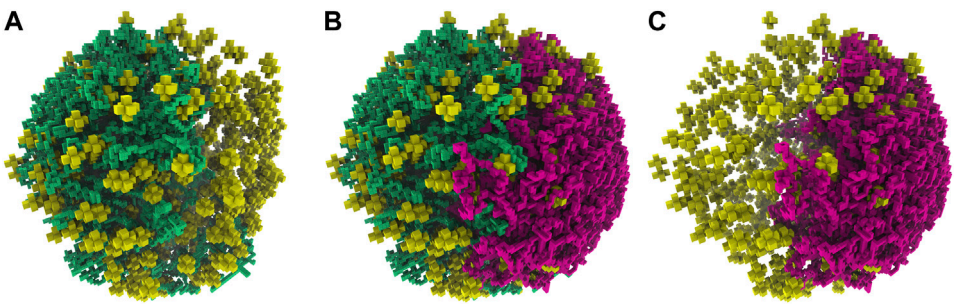


FIGURE 10 | (A)—First chromosome (green) generated within the large cell architecture of radius 247.42 nm and containing 820 ribosomes (yellow). **(B)**—An additional second chromosome (magenta) was generated within the large cell architecture while avoiding the first chromosome (green). **(C)**—The isolated second chromosome (magenta) in the large cell architecture after the removal of the first chromosome.

average. This is in good agreement with the fraction of 2.8–19% of ribosomes found to be in expressomes in *M. pneumoniae* by O'Reilly et al. (O'Reilly et al., 2020).

The computationally-generated chromosome configurations on the 4 nm lattice are converted to the 8 nm lattice before being used in RDME simulations. The conversion is done using a coarse-graining procedure where the 4 nm effective monomers are localized within the 8 nm lattice site containing them. These 8 nm lattice sites are then identified as chromosome sites. Due to the self-avoiding nature of the chromosome model, each 8 nm chromosome site can contain up to a maximum of eight monomers, where each monomer contains 11.8 bp of DNA. The number of monomers within the 8 nm lattice site are directly converted to up to eight of the maximum of sixteen particles within a lattice site. This coarse-graining procedure preserves the overall volume exclusion of the chromosome and the spatial heterogeneities caused by varying chromosome densities throughout the cell, and allows for genomically distant pieces of the chromosome to be spatially localized within the 8 nm chromosome sites. **Figure 8** shows the coarse-graining of a chromosome configuration in the small cell. The kinetic model of genetic information processing in Syn3A (Thornburg et al., 2019) can then be extended to include the effects of RNA polymerases diffusing between the spatial locations of genes within the chromosome (Weng and Xiao, 2014).

Two means were used to quantify the diffusivity of the coarse-grained chromosome configurations. First, the average monomer occupancy of the coarse-grained chromosome sites was calculated, with the target average occupancy being under 3

monomers per coarse-grained chromosome site, as shown in **Figure 9A**. Second, connected-component labeling was used to identify contiguous regions of high monomer occupancy, where diffusing particles may become trapped for extended periods of simulation time, as shown in **Figure 9B**. Maintaining an acceptable number of particles per lattice site has a significant impact on the efficiency of the multi-particle diffusion used in the GPU-accelerated LM (Hallock et al., 2014). The results for the small cell when the second approach for template-matching was applied, with the high ribosome packing density, showed that all 150 computationally-generated chromosome configurations were diffuse enough to be used for RDME simulations of Syn3A. The configurations satisfy the first criterion and there were no instances in which connected-components with an occupancy greater than 8 monomers per coarse-grained chromosome site could form closed shapes on the 8 nm lattice. The large cell is presumed to be near the end of the cell cycle, after DNA replication has been completed, and a second chromosome can be placed within the cell architecture, as shown in **Figure 10**. We assume the small cell and large cell are representative examples of cells at the start and end of the cell cycle, respectively, and the combination of the small and large cell architectures enable whole-cell simulations of Syn3A at the start and end of the cell cycle.

4 DISCUSSION

We developed a procedure to reconstruct single-cell geometries of Syn3A cells from cryo-electron tomograms. The procedure has two parts, the determination of the cell size and subsequent transformation of the ribosome distribution to a cell with spherical geometry, and the generation of circular chromosome configurations constrained by the spherical cell boundary and the ribosome distribution, and restrained by a small number of DNA loops observed in the experimental 3C-Seq map. Cell geometries were reconstructed for the small cell assumed to be at the start of the cell cycle and a large cell considered to be near the end of the cycle, when two chromosomes would be present.

The 3C-Seq chromosome contact map at a resolution of 1,000 bp has no secondary diagonal and confirms our assumption that Syn3A has no factors affecting the global structure of the chromosome. We based this assumption on our knowledge of the genome-scale gene essentiality and proteomics data, which indicated Syn3A lacked a parABS system or attachment organelle. Our computational model of the chromosome reproduced this behavior while constrained by the reconstructed cell geometry. Furthermore, we generated the DNA configurations under the assumption that the DNA was in a relaxed state with limited supercoiling. This was justified due to the high abundance of proteins that modify the supercoiling state, topoisomerases and gyrases, relative to the number of RNAP, and the relatively low abundance of proteins that form topological constraints and stabilize supercoiled loops, such as HU. We were able to model local structures, whose signatures were observed in the 3C-Seq map at a resolution of 250 bp. Currently, SMC is the only annotated protein in Syn3A that can form unsupercoiled loops, so it is possible the observed loops are formed by SMC. Recent studies show that SMC functions through active loop extrusion rather than static loop stabilization (van Ruiten and Rowland, 2018), so there are potentially additional unannotated effects and/or proteins causing the experimentally observed loops or causing localization of the actively-extruding loops through preferential SMC binding at the annotated locations. Our chromosome model does not include active loop extrusion and is only capable of reproducing the results of active loop extrusion in an ensemble average sense. At this time, we wish to avoid making further definitive statements about the nature of the local structure of Syn3A's chromosome until deeper sequencing is completed. Future experiments with additional restriction enzymes that cut the DNA at complementary positions and greater depth of the reads will help to improve our analysis.

We can speculate that the significant differences in the global chromosome organization of bacterial cells with natural genomes, such as *B. subtilis*, *C. crescentus*, and *M. pneumoniae*, to Syn3A with its synthetic genome are a result of genome minimization, both natural and targeted. The parent organism from which all variants (Syn1.0, Syn3.0, and Syn3A) are descended is *M. mycoides*, a choice that was made because *Mycoplasma* cells have small genomes that have been naturally reduced over evolutionarily-long time scales. This reduction likely occurred because they are parasitic organisms that can rely on a stable environment provided by their host. *Mycoplasmas* have dispensed of the genes that code for complex regulatory systems, such as the parABS system, and the remaining genes largely code for environment-independent functions essential to all life (Hutchison et al., 2016). Chromosome organization at the local level is dictated by NAPs and the supercoiling state of the DNA. Notably, while there is a significant disparity in the relative proteomics counts of NAPs in Syn3A and naturally-occurring bacteria, with the majority of NAPs being wholly absent from Syn3A's genome, there is no such disparity in the counts of proteins that modify the supercoiling state of the DNA. These proteins are essential to the function of Syn3A, which is not surprising due to the relationship between supercoiling and the universal process of transcription (Chong et al., 2014; Dorman, 2019).

From the reconstructed cell geometries, we estimated fractions of ribosomes that could be attached to the membrane or are complexed in possible polysomes and expressomes. We simply used distances between ribosomes and the membrane, other ribosomes, and the DNA, respectively, to predict these numbers. To confirm these estimates of the polysomes we could use the orientations of the ribosomes' entry and exit channels such that the ribosomes can pass mRNA between each other (O'Reilly et al., 2020). The membrane-bound ribosomes can be further characterized by determining which of those ribosomes have their 50S subunit facing the membrane (Ortiz et al., 2006). Further analysis of expressomes would require a template involving the RNAP and the essential transcription factor NusA that was found

to attach the RNAP to the ribosome in cryo-ET of *M. pneumoniae* (O'Reilly et al., 2020). In the same *M. pneumoniae* study, subtomogram averaging was used to more confidently assign expressome structures along with orientation of the mRNA entry site to help identify ribosomes complexed with RNAP.

The effects of ribosomes attached to the membrane or complexed in polysomes and expressomes can all be included in future whole-cell, spatially-resolved kinetic models. The configurations resulting from the SAP model of the bacterial chromosome are directly transferrable to the 8 nm lattice representation used for LM simulations of whole Syn3A cells through a coarse-graining procedure. The coarse-grained chromosome configurations specify the spatial heterogeneities caused by DNA-crowding in whole-cell kinetic models of Syn3A and define the spatial locations of genes to investigate spatial and temporal correlations in gene expression (Weng and Xiao, 2014; Thornburg et al., 2019). Future work will focus on assigning chromosomal interactions based on improved experimental 3C-Seq libraries, improving the model to include dynamic formation and relaxation of supercoiling and plectonemic loops, and incorporating dynamic representations of the chromosome (Miermans and Broedersz, 2020) within the LM simulations, which will include DNA diffusion and chromosome replication. The compactness and degree local structure of the DNA determines the accessibility of its genes to RNAP which is an important consideration in the whole-cell simulations of all the cellular networks being developed for the minimal cell JCVI-syn3A.

DATA AVAILABILITY STATEMENT

The tilt-series in this study have been deposited in the Electron Microscopy Public Image Archive (https://www.ebi.ac.uk/pdbe/emdb/empiar/) under EMPIAR entry numbers EMPIAR-10685 (large cell) and EMPIAR-10686 (small cell). The reconstructed tomograms in this study have been deposited in the Electron Microscopy Data Bank (https://www.ebi.ac.uk/pdbe/emdb/index.html/) under EMDB entry numbers EMD-23660 (large cell) and EMD-23661 (small cell). 3C-Seq libraries of Syn3A are available from the 4TU repository (https://data.4tu.nl/) under DOI: https://doi.org/10.4121/14333618. The software used in this study can be found at https://github.com/brg4/SAP_chromosome.

AUTHOR CONTRIBUTIONS

BG: development of method for reconstructing cell architecture from tomograms, development of chromosome model, development of coarse-graining procedure, contact map analysis, cell architecture analysis, and writing original draft.

ZT: development of method for reconstructing cell architecture from tomograms, development of chromosome model, development of coarse-graining procedure, cell architecture analysis, assisted in programming of Jupyter notebooks, data-curation, and writing original draft.

ZL-S: development of method for reconstructing cell architecture from tomograms, development of chromosome model, development of coarse-graining procedure, contact map analysis, cell architecture analysis, and writing original draft.

VL: tomogram collection and processing, development of method for reconstructing cell architecture from tomograms, data-curation, and writing original draft.

EV: development of method for reconstructing cell architecture from tomograms, data-curation, and writing original draft.

F-ZR: 3C-Seq library, experimental 3C map, and editing manuscript.

RD: 3C-Seq library, experimental 3C map, and editing manuscript.

JG: establishing the network of collaborators, development of the minimal cell, and editing manuscript.

FUNDING

BG, ZT, and ZL-S: Partial support from NSF MCB 1818344 and 1840320, The Center for the Physics of Living Cells NSF PHY 1430124, and The Physics of Living Systems Student Research Network NSF PHY 1505008. The cell figures in the workflow diagram and all lattice representations of ribosomes and DNA were prepared using Visual Molecular Dynamics (VMD), developed by the NIH Center for Macromolecular Modeling and Bioinformatics in the Beckman Institute at UIUC, with support from NIH P41-GM104601-28.

VL and EV: This work was supported by an NIH Director's New Innovator Award 1DP2GM123494-01 (to EV) and NIH 5T32GM7240-40 (to VL). VL is also supported in part by NIH R35GM118290 awarded to Susan S. Golden. This work on Syn3A by VL and EV is also supported in part by NSF MCB 1818344. This work was supported by the National Science Foundation MRI grant (NSF DBI 1920374). We acknowledge the use of the UCSD Cryo-Electron Microscopy Facility which is supported by NIH grants to Dr Timothy S. Baker and a gift from the Agouron Institute to UCSD. Molecular graphics and analyses performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311.

F-ZR and RD: This research was supported by a VICI grant (VICI 016.160.613) and an ENW Groot grant (OCENW.GROOT. 2019.012) from the Netherlands Organization for Scientific Research (RD).

JG: Partial support from NSF MCB 1818344, 1840301 and 1840320.

ACKNOWLEDGMENTS

BG, ZT, and ZL-S: We thank John Stone at the Beckman Institute at UIUC for his assistance in preparing scientific visualizations

using VMD. F-ZR and RD: We thank Utrecht Sequencing Facility for providing sequencing service and data. Utrecht Sequencing Facility is subsidized by the University Medical Center Utrecht, Hubrecht Institute, Utrecht University and The Netherlands X-omics Initiative (NWO project 184.034.019). We thank Wouter de Laat and Amin Allahyar (Hubrecht Institute, The Netherlands) for discussions and assistance with data analysis.
We thank Kim Wise at the J. Craig Venter Institute (JCVI) for providing JCVI-Syn3A cells and growth medium to the laboratories of EV and RD.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.644133/full#supplementary-material

REFERENCES

Annunziatella, C., Chiariello, A. M., Esposito, A., Bianco, S., Fiorillo, L., and Nicodemi, M. (2018). Molecular Dynamics Simulations of the Strings and Binders Switch Model of Chromatin. *Methods*. 142, 81–88. doi:10.1016/j.ymeth.2018.02.024

Bakshi, S., Siryaporn, A., Goulian, M., and Weisshaar, J. C. (2012). Superresolution Imaging of Ribosomes and Rna Polymerase in Live escherichia Coli Cells. *Mol. Microbiol.* 85, 21–38. doi:10.1111/j.1365-2958.2012.08081.x

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: A Comprehensive Technique to Capture the Conformation of genomes3D Chromatin Architecture. *Methods*. 58, 268–276. doi:10.1016/j.ymeth.2012.05.001

Bianchi, D. M., Peterson, J. R., Earnest, T. M., Hallock, M. J., and Luthey-Schulten, Z. (2018). Hybrid CME-ODE Method for Efficient Simulation of the Galactose Switch in Yeast. *IET Syst. Biol.* 12, 170–176. doi:10.1049/iet-syb.2017.0070

Birnie, A., and Dekker, C. (2021). Genome-in-a-box: Building a Chromosome from the Bottom up. *ACS Nano*. 15, 111–124. doi:10.1021/acsnano.0c07397

Brandt, F., Etchells, S. A., Ortiz, J. O., Elcock, A. H., Hartl, F. U., and Baumeister, W. (2009). The Native 3d Organization of Bacterial Polysomes. *Cell*. 136, 261–271. doi:10.1016/j.cell.2008.11.016

Bremer, H., and Dennis, P. P. (2008). Modulation of Chemical Composition and Other Parameters of the Cell at Different Exponential Growth Rates. *EcoSal Plus*. 3, 8. doi:10.1128/ecosal.5.2.3

Breuer, M., Earnest, T. M., Merryman, C., Wise, K. S., Sun, L., Lynott, M. R., et al. (2019). Essential Metabolism for a Minimal Cell. *eLife*. 8, e36842. doi:10.7554/eLife.36842

Brinkers, S., Dietrich, H. R. C., de Groote, F. H., Young, I. T., and Rieger, B. (2009). The Persistence Length of Double Stranded Dna Determined Using Dark Field Tethered Particle Motion. *J. Chem. Phys.* 130, 215105. doi:10.1063/1.3142699

Buenemann, M., and Lenz, P. (2010). A Geometrical Model for Dna Organization in Bacteria. *PLOS ONE*. 5, e13806–13. doi:10.1371/journal.pone.0013806

Castaño-Díez, D., Kudryashev, M., Arheit, M., and Stahlberg, H. (2012). Dynamo: A Flexible, User-Friendly Development Tool for Subtomogram Averaging of Cryo-EM Data in High-Performance Computing Environments. *J. Struct. Biol.* 178, 139–151. doi:10.1016/j.jsb.2011.12.017

Chong, S., Chen, C., Ge, H., and Xie, X. S. (2014). Mechanism of Transcriptional Bursting in Bacteria. *Cell*. 158, 314–326. doi:10.1016/j.cell.2014.05.038

Crémazy, F. G., Rashid, F. M., Haycocks, J. R., Lamberte, L. E., Grainger, D. C., and Dame, R. T. (2018). Determination of the 3D Genome Organization of Bacteria Using Hi-C. *Methods Mol Biol.* 3, 3–18. doi:10.1007/978-1-4939-8675-0_1

Dame, R. T., Rashid, F.-Z. M., and Grainger, D. C. (2019). Chromosome Organization in Bacteria: Mechanistic Insights into Genome Structure and Function. *Nat. Rev. Genet.* 21, 227–242. doi:10.1038/s41576-019-0185-4

Dame, R. T., and Tark-Dame, M. (2016). Bacterial Chromatin: Converging Views at Different Scales. *Curr. Opin. Cel Biol.* 40, 60–65. doi:10.1016/j.ceb.2016.02.015

Dame, R. T. (2005). The Role of Nucleoid-Associated Proteins in the Organization and Compaction of Bacterial Chromatin. *Mol. Microbiol.* 56, 858–870. doi:10.1111/j.1365-2958.2005.04598.x

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the Three-Dimensional Organization of Genomes: Interpreting Chromatin Interaction Data. *Nat. Rev. Genet.* 14, 390–403. doi:10.1038/nrg3454

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing Chromosome Conformation. *Science*. 295, 1306–1311. doi:10.1126/science.1067799

Di Pierro, M., Cheng, R. R., Lieberman Aiden, E., Wolynes, P. G., and Onuchic, J. N. (2017). De Novo prediction of Human Chromosome Structures: Epigenetic Marking Patterns Encode Genome Architecture. *Proc. Natl. Acad. Sci. USA*. 114, 12126–12131. doi:10.1073/pnas.1714980114

Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G., and Onuchic, J. N. (2016). Transferable Model for Chromosome Architecture. *Proc. Natl. Acad. Sci. USA*. 113, 12168–12173. doi:10.1073/pnas.1613607113

Diebold-Durand, M.-L., Lee, H., Ruiz Avila, L. B., Noh, H., Shin, H.-C., Im, H., et al. (2017). Structure of Full-Length Smc and Rearrangements Required for Chromosome Organization. *Mol. Cel.* 67, 334–347.e5. doi:10.1016/j.molcel.2017.06.010

Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D., et al. (1995). Principles of Protein Folding-Aa Perspective from Simple Exact Models, *Protein Sci.* 4, 561–602. doi:10.1002/pro.5560040401

Dorman, C. J. (2019). Dna Supercoiling and Transcription in Bacteria: a Two-Way Street. *BMC Mol. Cel Biol.* 20, 26. doi:10.1186/s12860-019-0211-6

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., et al. (2010). A Three-Dimensional Model of the Yeast Genome. *Nature*. 465, 363–367. doi:10.1038/nature08973

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., et al. (2016a). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cel Syst.* 3, 99–101. doi:10.1016/j.cels.2015.07.012

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., et al. (2016b). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cel Syst.* 3, 95–98. doi:10.1016/j.cels.2016.07.002

Earnest, T. M., Cole, J. A., and Luthey-Schulten, Z. (2018). Simulating Biological Processes: Stochastic Physics from Whole Cells to Colonies. *Rep. Prog. Phys.* 81, 052601. doi:10.1088/1361-6633/aaae2c

Earnest, T. M., Watanabe, R., Stone, J. E., Mahamid, J., Baumeister, W., Villa, E., et al. (2017). Challenges of Integrating Stochastic Dynamics and Cryo-Electron Tomograms in Whole-Cell Simulations. *J. Phys. Chem. B*. 121, 3871–3881. doi:10.1021/acs.jpcb.7b00672

Forchhammer, J., and Lindahl, L. (1971). Growth Rate of Polypeptide Chains as a Function of the Cell Growth Rate in a Mutant of escherichia Coli 15. *J. Mol. Biol.* 55, 563–568. doi:10.1016/0022-2836(71)90337-8

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L. A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cel Rep.* 15, 2038–2049. doi:10.1016/j.celrep.2016.04.085

Ganji, M., Shaltiel, I. A., Bisht, S., Kim, E., Kalichava, A., Haering, C. H., et al. (2018). Real-time Imaging of Dna Loop Extrusion by Condensin. *Science*. 360, 102–105. doi:10.1126/science.aar7831

Geggier, S., Kotlyar, A., and Vologodskii, A. (2010). Temperature Dependence of DNA Persistence Length. *Nucleic Acids Res.* 39, 1419–1426. doi:10.1093/nar/gkq932

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R.-Y., Algire, M. A., et al. (2010). Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science*. 329, 52–56. doi:10.1126/science.1190719

Goodsell, D. S., Autin, L., and Olson, A. J. (2018). Lattice Models of Bacterial Nucleoids. *J. Phys. Chem. B*. 122, 5441–5447. doi:10.1021/acs.jpcb.7b11770

Hacker, W. C., Li, S., and Elcock, A. H. (2017). Features of Genomic Organization in a Nucleotide-Resolution Molecular Model of the escherichia Coli Chromosome. *Nucleic Acids Res.* 45, 7541–7554. doi:10.1093/nar/gkx541

Haddad, N., Jost, D., and Vaillant, C. (2017). Perspectives: Using Polymer Modeling to Understand the Formation and Function of Nuclear Compartments. *Chromosome Res.* 25, 35–50. doi:10.1007/s10577-016-9548-2

Hallock, M. J., Stone, J. E., Roberts, E., Fry, C., and Luthey-Schulten, Z. (2014). Simulation of Reaction Diffusion Processes over Biologically Relevant Size and Time Scales Using Multi-Gpu Workstations. *Parallel Comput.* 40, 86–99. doi:10.1016/j.parco.2014.03.009

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*. 57, 97–109. doi:10.1093/biomet/57.1.97

Heilmann, O. J., and Rotne, J. (1982). Exact and Monte Carlo Computations on a Lattice Model for Change of Conformation of a Polymer. *J. Stat. Phys.* 27, 19–35. doi:10.1007/BF01011737

Hsu, H.-P., and Binder, K. (2012). Stretching Semiflexible Polymer Chains: Evidence for the Importance of Excluded Volume Effects from Monte Carlo Simulation. *J. Chem. Phys.* 136, 024901. doi:10.1063/1.3674303

Hua, K.-J., and Ma, B.-G. (2019). Evr: Reconstruction of Bacterial Chromosome 3d Structure Models Using Error-Vector Resultant Algorithm. *BMC Genomics*. 20, 738. doi:10.1186/s12864-019-6096-0

Hutchison, C. A., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., et al. (2016). Design and Synthesis of a Minimal Bacterial Genome. *Science*. 351, aad6253. doi:10.1126/science.aad6253

Junier, I., Spill, Y. G., Marti-Renom, M. A., Beato, M., and le Dily, F. (2015). On the Demultiplexing of Chromosome Capture Conformation Data. *FEBS Lett.* 589, 3005–3013. doi:10.1016/j.febslet.2015.05.049

Kim, S., Beltran, B., Irnov, I., and Jacobs-Wagner, C. (2019). Long-distance Cooperative and Antagonistic Rna Polymerase Dynamics via Dna Supercoiling. *Cell.* 179, 106–119. doi:10.1016/j.cell.2019.08.033

Knight, P. A., and Ruiz, D. (2012). A Fast Algorithm for Matrix Balancing. *IMA J. Numer. Anal.* 33, 1029–1047. doi:10.1093/imanum/drs019

Kohler, R., Mooney, R. A., Mills, D. J., Landick, R., and Cramer, P. (2017). Architecture of a Transcribing-Translating Expressome. *Science*. 356, 194–197. doi:10.1126/science.aal3059

Kratky, O., and Porod, G. (1949). Röntgenuntersuchung Gelöster Fadenmoleküle. *Recl. Trav. Chim. Pays-bas*. 68, 1106–1122. doi:10.1002/recl.19490681203

Kremer, J. R., Mastronarde, D. N., and McIntosh, J. R. (1996). Computer Visualization of Three-Dimensional Image Data Using IMOD. *J. Struct. Biol.* 116, 71–76. doi:10.1006/jsbi.1996.0013

Kühner, S., van Noort, V., Betts, M. J., Leo-Macias, A., Batisse, C., Rode, M., et al. (2009). Proteome Organization in a Genome-Reduced Bacterium. *Science*. 326, 1235–1240. doi:10.1126/science.1176343

Lasker, K., Boeynaems, S., Lam, V., Stainton, E., Jacquemyn, M., Daelemans, D., et al. (2021). A Modular Platform for Engineering Function of Natural and Synthetic Biomolecular Condensates. *bioRxiv*. 45, 11–19. doi:10.1101/2021.02.03.429226

Lau, K. F., and Dill, K. A. (1989). A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins. *Macromolecules*. 22, 3986–3997. doi:10.1021/ma00200a030

Le, T. B. K., Imakaev, M. V., Mirny, L. A., and Laub, M. T. (2013). High-resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science*. 342, 731–734. doi:10.1126/science.1242059

Lesne, A., Riposo, J., Roger, P., Cournac, A., and Mozziconacci, J. (2014). 3d Genome Reconstruction from Chromosomal Contacts. *Nat. Methods*. 11, 1141–1143. doi:10.1038/nmeth.3104

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*. 326, 289–293. doi:10.1126/science.1181369

Lioy, V. S., and Boccard, F. (2018). Conformational Studies of Bacterial Chromosomes by High-Throughput Sequencing Methods. In *High-Density Sequencing Applications in Microbial Molecular Genetics*, ed. A. J. Carpousis (Cambridge, MA: Academic Press), Vol. 612 of Methods in Enzymology. 25–45. doi:10.1016/bs.mie.2018.07.007

Lioy, V. S., Cournac, A., Marbouty, M., Duigou, S., Mozziconacci, J., Espéli, O., et al. (2018). Multiscale Structuring of the *E. coli* Chromosome by Nucleoid-Associated and Condensin Proteins. *Cell.* 172, 771–783.e18. doi:10.1016/j.cell.2017.12.027

Lioy, V. S., Junier, I., Lagage, V., Vallet, I., and Boccard, F. (2020). Distinct Activities of Bacterial Condensins for Chromosome Management in pseudomonas Aeruginosa. *Cel Rep.* 33, 108344. doi:10.1016/j.celrep.2020.108344

Livny, J., Yamaichi, Y., and Waldor, M. K. (2007). Distribution of Centromere-like Pars Sites in Bacteria: Insights from Comparative Genomics. *J Bacteriol.* 189, 8693–8703. doi:10.1128/JB.01239-07

Lua, R., Borovinskiy, A. L., and Grosberg, A. Y. (2004). Fractal and Statistical Properties of Large Compact Polymers: a Computational Study. *Polymer* 45, 717–731. doi:10.1016/j.polymer.2003.10.073

Madras, N., Orlitsky, A., and Shepp, L. A. (1990). Monte Carlo Generation of Self-Avoiding Walks with Fixed Endpoints and Fixed Length. *J. Stat. Phys.* 58, 159–183. doi:10.1007/BF01020290

Madras, N., and Sokal, A. D. (1987). Nonergodicity of Local, Length-Conserving Monte Carlo Algorithms for the Self-Avoiding Walk. *J. Stat. Phys.* 47, 573–595. doi:10.1007/BF01007527

Manning, G. S. (2006). The Persistence Length of Dna Is Reached from the Persistence Length of its Null Isomer through an Internal Electrostatic Stretching Force. *Biophysical J.* 91, 3607–3616. doi:10.1529/biophysj.106.089029

Mantelli, S., Muller, P., Harlepp, S., and Maaloum, M. (2011). Conformational Analysis and Estimation of the Persistence Length of Dna Using Atomic Force Microscopy in Solution. *Soft Matter*. 7, 3412–3416. doi:10.1039/C0SM01160F

Marbouty, M., LeGall, A., Cattoni, D. I., Cournac, A., Koh, A., Fiche, J.-B., et al. (2015). Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. *Mol. Cel.* 59, 588–602. doi:10.1016/j.molcel.2015.07.020

Marko, J. F., De Los Rios, P., Barducci, A., and Gruber, S. (2019). DNA-segment-capture Model for Loop Extrusion by Structural Maintenance of Chromosome (SMC) Protein Complexes. *Nucleic Acids Res.* 47, 6956–6972. doi:10.1093/nar/gkz497

Martinez-Sanchez, A., Garcia, I., Asano, S., Lucic, V., and Fernandez, J.-J. (2014). Robust Membrane Detection Based on Tensor Voting for Electron Tomography. *J. Struct. Biol.* 186, 49–61. doi:10.1016/j.jsb.2014.02.015

Mastronarde, D. N. (1997). Dual-Axis Tomography: An Approach with Alignment Methods that Preserve Resolution. *J. Struct. Biol.* 120, 343–352. doi:10.1006/jsbi.1997.3919

Mastronarde, D. N. (2005). Automated Electron Microscope Tomography Using Robust Prediction of Specimen Movements. *J. Struct. Biol.* 152, 36–51. doi:10.1016/j.jsb.2005.07.007

Mastronarde, D. N., and Held, S. R. (2017). Automated Tilt Series Alignment and Tomographic Reconstruction in IMOD. *J. Struct. Biol.* 197, 102–113. doi:10.1016/j.jsb.2016.07.011

Matteau, D., Lachance, J. C., Grenier, F., Gauthier, S., Daubenspeck, J. M., Dybvig, K., et al. (2020). Integrative Characterization of the Near-minimal Bacterium Mesoplasma Florum. *Mol. Syst. Biol.* 16, e9844. doi:10.15252/msb.20209844

Messelink, J. J. B., van Teeseling, M. C. F., Janssen, J., Thanbichler, M., and Broedersz, C. P. (2021). Learning the distribution of single-cell chromosome conformations in bacteria reveals emergent order across genomic scales. *Nat. Commun.* 12, 1963. doi:10.1038/s41467-021-22189-x

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* 21, 1087–1092. doi:10.1063/1.1699114

Miermans, C. A., and Broedersz, C. P. (2020). A Lattice Kinetic Monte-Carlo Method for Simulating Chromosomal Dynamics and Other (Non-)equilibrium Bio-Assemblies. *Soft Matter*. 16, 544–556. doi:10.1039/C9SM01835B

Mirny, L. A. (2011). The Fractal Globule as a Model of Chromatin Architecture in the Cell. *Chromosome Res.* 19, 37–51. doi:10.1007/s10577-010-9177-0

Needham, D., and Nunn, R. S. (1990). Elastic Deformation and Failure of Lipid Bilayer Membranes Containing Cholesterol. *Biophysical J.* 58, 997–1009. doi:10.1016/S0006-3495(90)82444-9

Nevo-Dinur, K., Nussbaum-Shochat, A., Ben-Yehuda, S., and Amster-Choder, O. (2011). Translation-independent localization of mrna in e. coli. *Science*. 331, 1081–1084. doi:10.1126/science.1195691

Ohniwa, R. L., Ushijima, Y., Saito, S., and Morikawa, K. (2011). Proteomic Analyses of Nucleoid-Associated Proteins in escherichia Coli, pseudomonas Aeruginosa, bacillus Subtilis, and staphylococcus Aureus. *PLOS ONE*. 6, e19172–18. doi:10.1371/journal.pone.0019172

Oluwadare, O., Highsmith, M., and Cheng, J. (2019). An Overview of Methods for Reconstructing 3-d Chromosome and Genome Structures from Hi-C Data. *Biol. Proced. Online* 21, 7. doi:10.1186/s12575-019-0094-0

O'Reilly, F. J., Xue, L., Graziadei, A., Sinn, L., Lenz, S., Tegunov, D., et al. (2020). In-cell Architecture of an Actively Transcribing-Translating Expressome. *Science*. 369, 554–557. doi:10.1126/science.abb3758

Ortiz, J. O., Förster, F., Kürner, J., Linaroudis, A. A., and Baumeister, W. (2006). Mapping 70s Ribosomes in Intact Cells by Cryoelectron Tomography and Pattern Recognition. *J. Struct. Biol.* 156, 334–341. doi:10.1016/j.jsb.2006.04.014

Pelletier, J. F., Sun, L., Wise, K. S., Assad-Garcia, N., Karas, B. J., Deerinck, T. J., et al. (2021). Genetic Requirements for Cell Division in a Genomically Minimal Cell. *Cell*. 183, 1–11. doi:10.1016/j.cell2021.03.008

Petkau, A., Stuart-Edwards, M., Stothard, P., and Van Domselaar, G. (2010). Interactive Microbial Genome Visualization with GView. *Bioinformatics*. 26, 3125–3126. doi:10.1093/bioinformatics/btq588

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

Phillips, L. A., Hotham-Iglewski, B., and Franklin, R. M. (1969). Polyribosomes of *Escherichia coli*. *J. Mol. Biol.* 40, 279–288. doi:10.1016/0022-2836(69)90475-6

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3d Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*. 159, 1665–1680. doi:10.1016/j.cell.2014.11.021

Roberts, E., Magis, A., Ortiz, J. O., Baumeister, W., and Luthey-Schulten, Z. (2011). Noise Contributions in an Inducible Genetic Switch: A Whole-Cell Simulation Study. *Plos Comput. Biol.* 7, e1002010–21. doi:10.1371/journal.pcbi.1002010

Roberts, E., Stone, J. E., and Luthey-Schulten, Z. (2013). Lattice Microbes: High-Performance Stochastic Simulation Method for the Reaction-Diffusion Master Equation. *J. Comput. Chem.* 34, 245–255. doi:10.1002/jcc.23130

Rosa, A., and Zimmer, C. (2014). Computational Models of Large-Scale Genome Architecture. *Int. Rev. Cel Mol Biol.* 307, 275–349. doi:10.1016/B978-0-12-800046-5.00009-6

Rouse, P. E. (1953). A Theory of the Linear Viscoelastic Properties of Dilute Solutions of Coiling Polymers. *J. Chem. Phys.* 21, 1272–1280. doi:10.1063/1.1699180

Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., et al. (2012). Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *Plos Biol.* 10, e1001244–5. doi:10.1371/journal.pbio.1001244

Ryu, J.-K., Bouchoux, C., Liu, H. W., Kim, E., Minamino, M., de Groot, R., et al. (2021). Bridging-induced phase separation induced by cohesin SMC protein complexes. *Sci. Adv.* 7, eabe5905. doi:10.1126/sciadv.abe5905

Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., et al. (2015). Chromatin Extrusion Explains Key Features of Loop and Domain Formation in Wild-type and Engineered Genomes. *Proc. Natl. Acad. Sci. USA*. 112, E6456–E6465. doi:10.1073/pnas.1518552112

Scheres, S. H. W. (2012). Relion: Implementation of a Bayesian Approach to Cryo-Em Structure Determination. *J. Struct. Biol.* 180, 519–530. doi:10.1016/j.jsb.2012.09.006

Schorb, M., Haberbosch, I., Hagen, W. J. H., Schwab, Y., and Mastronarde, D. N. (2019). Software Tools for Automated Transmission Electron Microscopy. *Nat. Methods*. 16, 471–477. doi:10.1038/s41592-019-0396-9

Seybert, A., Herrmann, R., and Frangakis, A. S. (2006). Structural Analysis of Mycoplasma Pneumoniae by Cryo-Electron Tomography. *J. Struct. Biol.* 156, 342–354. doi:10.1016/j.jsb.2006.04.010

Sokal, A. D. (1995). "Monte Carlo Methods for the Self-Avoiding Walk," in *Monte Carlo and Molecular Dynamics Simulations Polymer*. Editor K. Binder (USA: Oxford University Press, Inc)chap. 2. 45–124.

Tegunov, D., and Cramer, P. (2019). Real-time Cryo-Electron Microscopy Data Preprocessing with Warp. *Nat. Methods*. 16, 1146–1152. doi:10.1038/s41592-019-0580-y

Thornburg, Z. R., Melo, M. C. R., Bianchi, D., Brier, T. A., Crotty, C., Breuer, M., et al. (2019). Kinetic Modeling of the Genetic Information Processes in a Minimal Cell. *Front. Mol. Biosci.* 6, 130. doi:10.3389/fmolb.2019.00130

Tran, N. T., Laub, M. T., and Le, T. B. K. (2017). Smc Progressively Aligns Chromosomal Arms in caulobacter Crescentus but Is Antagonized by Convergent Transcription. *Cel Rep.* 20, 2057–2071. doi:10.1016/j.celrep.2017.08.026

Trussart, M., Yus, E., Martinez, S., Baù, D., Tahara, Y. O., Pengo, T., et al. (2017). Defined Chromosome Structure in the Genome-Reduced Bacterium Mycoplasma Pneumoniae. *Nat. Commun.* 8, 14665. doi:10.1038/ncomms14665

Umbarger, M. A., Toro, E., Wright, M. A., Porreca, G. J., Baù, D., Hong, S.-H., et al. (2011). The Three-Dimensional Architecture of a Bacterial Genome and its Alteration by Genetic Perturbation. *Mol. Cel.* 44, 252–264. doi:10.1016/j.molcel.2011.09.010

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., et al. (2010). Hi-c: A Method to Study the Three-Dimensional Architecture of Genomes. *JoVE*. 6(39): e1869. doi:10.3791/1869

van Ruiten, M. S., and Rowland, B. D. (2018). Smc Complexes: Universal Dna Looping Machines with Distinct Regulators. *Trends Genet.* 34, 477–487. doi:10.1016/j.tig.2018.03.003

Verdier, P. H., and Stockmayer, W. H. (1962). Monte Carlo Calculations on the Dynamics of Polymers in Dilute Solution. *J. Chem. Phys.* 36, 227–235. doi:10.1063/1.1732301

Verma, S. C., Qian, Z., and Adhya, S. L. (2019). Architecture of the escherichia Coli Nucleoid. *Plos Genet.* 15, e1008456–35. doi:10.1371/journal.pgen.1008456

Vologodskii, A. V., Levene, S. D., Klenin, K. V., Frank-Kamenetskii, M., and Cozzarelli, N. R. (1992). Conformational and Thermodynamic Properties of Supercoiled Dna. *J. Mol. Biol.* 227, 1224–1243. doi:10.1016/0022-2836(92)90533-P

Wang, F., and Landau, D. P. (2001). Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Phys. Rev. Lett.* 86, 2050–2053. doi:10.1103/PhysRevLett.86.2050

Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D., and Mering, C. (2015). Version 4.0 of PaxDb: Protein Abundance Data, Integrated across Model Organisms, Tissues, and Cell-lines. *PROTEOMICS*. 15, 3163–3168. doi:10.1002/pmic.201400441

Wang, X., and Rudner, D. Z. (2014). Spatial Organization of Bacterial Chromosomes. Curr Opin Microbiol. 22, 66–72. doi:10.1016/j.mib.2014.09.016

Wang, X., Tang, O. W., Riley, E. P., and Rudner, D. Z. (2014). The Smc Condensin Complex Is Required for Origin Segregation in bacillus Subtilis. *Curr. Biol.* 24, 287–292. doi:10.1016/j.cub.2013.11.050

Weng, X., and Xiao, J. (2014). Spatial Organization of Transcription in Bacterial Cells. *Trends Genet.* 30, 287–297. doi:10.1016/j.tig.2014.04.008

Williamson, D. L., and Whitcomb, R. F. (1975). Plant Mycoplasmas: A Cultivable Spiroplasma Causes Corn Stunt Disease. *Science*. 188, 1018–1020. doi:10.1126/science.188.4192.1018

Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V., Yamada, T., Chen, W.-H., et al. (2009). Impact of Genome Reduction on Bacterial Metabolism and its Regulation. *science*. 326, 1263–1268. doi:10.1126/science.1177263

Zhang, J.-Z., Peng, X.-Y., Liu, S., Jiang, B.-P., Ji, S.-C., and Shen, X.-C. (2019). The Persistence Length of Semiflexible Polymers in Lattice Monte Carlo Simulations. *Polymers*. 11, 295. doi:10.3390/polym11020295

Zheng, S. Q., Palovcak, E., Armache, J.-P., Verba, K. A., Cheng, Y., and Agard, D. A. (2017). MotionCor2: Anisotropic Correction of Beam-Induced Motion for Improved Cryo-Electron Microscopy. *Nat. Methods*. 14, 331–332. doi:10.1038/nmeth.4193

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 © 2021 Gilbert, Thornburg, Lam, Rashid, Glass, Villa, Dame and Luthey-Schulten. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 26 July 2021
doi: 10.3389/fmolb.2021.686086

Investigating Crosstalk Among PTMs Provides Novel Insight Into the Structural Basis Underlying the Differential Effects of Nt17 PTMs on Mutant Httex1 Aggregation

Anass Chiki[1†], Zhidian Zhang[1,2†], Kolla Rajasekhar[1], Luciano A. Abriata[2], Iman Rostami[1,3], Lucien F. Krapp[2], Driss Boudeffa[1], Matteo Dal Peraro[2] and Hilal A. Lashuel[1]**

[1]Laboratory of Molecular and Chemical Biology of Neurodegeneration, School of Life Sciences, Brain Mind Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, [2]Laboratory for Biomolecular Modeling, School of Life Sciences, Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, [3]Laboratory of Biomolecular Research, Department of Biology and Chemistry, Paul Scherrer Institute, Villigen, Switzerland

Post-translational modifications (PTMs) within the first 17 amino acids (Nt17) of the Huntingtin protein (Htt) have been shown to inhibit the aggregation and attenuate the toxicity of mutant Htt proteins *in vitro* and in various models of Huntington's disease. Here, we expand on these studies by investigating the effect of methionine eight oxidation (oxM8) and its crosstalk with lysine 6 acetylation (AcK6) or threonine 3 phosphorylation (pT3) on the aggregation of mutant Httex1 (mHttex1). We show that M8 oxidation delays but does not inhibit the aggregation and has no effect on the final morphologies of mHttex1aggregates. The presence of both oxM8 and AcK6 resulted in dramatic inhibition of Httex1 fibrillization. Circular dichroism spectroscopy and molecular dynamics simulation studies show that PTMs that lower the mHttex1 aggregation rate (oxM8, AcK6/oxM8, pT3, pT3/oxM8, and pS13) result in increased population of a short N-terminal helix (first eight residues) in Nt17 or decreased abundance of other helical forms, including long helix and short C-terminal helix. PTMs that did not alter the aggregation rate (AcK6) of mHttex1 exhibit a similar distribution of helical conformation as the unmodified peptides. These results show that the relative abundance of N- vs. C-terminal helical conformations and long helices, rather than the overall helicity of Nt17, better explains the effect of different Nt17 PTMs on mHttex1; thus, explaining the lack of correlation between the effect of PTMs on the overall helicity of Nt17 and mHttex1 aggregation *in vitro*. Taken together, our results provide novel structural insight into the differential effects of single PTMs and crosstalk between different PTMs in regulating mHttex1 aggregation.

Keywords: post-translational modifications, huntingtin, molecular dynamics simulation, chemical semi-synthesis, atomic force microscope

OPEN ACCESS

Edited by:
Joanna Trylska, University of Warsaw, Poland

Reviewed by:
Marek Cieplak, Institute of Physics (PAN), Poland
Beata Wielgus-Kutrowska, University of Warsaw, Poland
Birgit Strodel, Helmholtz-Verband Deutscher Forschungszentren (HZ), Germany

***Correspondence:**
Hilal A. Lashuel
hilal.lashuel@epfl.ch
Matteo Dal Peraro
matteo.dalperaro@epfl.ch

[†]These authors contributed equally to this work

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *26 March 2021*
Accepted: *06 July 2021*
Published: *26 July 2021*

Citation:
Chiki A, Zhang Z, Rajasekhar K, Abriata LA, Rostami I, Krapp LF, Boudeffa D, Dal Peraro M and Lashuel HA (2021) Investigating Crosstalk Among PTMs Provides Novel Insight Into the Structural Basis Underlying the Differential Effects of Nt17 PTMs on Mutant Httex1 Aggregation. Front. Mol. Biosci. 8:686086. doi: 10.3389/fmolb.2021.686086

INTRODUCTION

Huntington's disease (HD) is a fatal, autosomal neurodegenerative disease characterized by motor (Huntington, 1872; Ross and Tabrizi, 2011) and cognitive declines (Marder et al., 2000) as well as psychiatric symptoms (Paulsen et al., 2001). HD is caused by a mutation in the huntingtin gene (*HTT*), resulting in an expansion in the CAG repeat within its first exon (Gusella et al., 1983; MacDonald, 1993), which is then translated into an extended polyglutamine (polyQ) repeat in the huntingtin protein (Htt) (Kremer et al., 1994). HD occurs when the length of the polyQ repeat is higher than the critical threshold of ≥36 (Snell et al., 1993). At the neuropathological level, HD is characterized by neuronal degeneration in the striatum and the cortex (Reiner et al., 1988; Rosas et al., 2003) and the formation and accumulation of nuclear inclusions composed of mutant Htt (polyQ repeat ≥36) aggregates and other proteins (DiFiglia et al., 1997; Hodgson et al., 1999). Several studies have shown that these inclusions are composed of fibrillar and potentially oligomeric species derived from N-terminal fragments containing expanded polyQ repeats (Mangiarini et al., 1996; DiFiglia et al., 1997; Lunkes et al., 2002; Landles et al., 2010; Sathasivam et al., 2013). One of the major N-terminal fragments found in these inclusions represents an N-terminal fragment that corresponds to exon1 of the Htt protein (Httex1) (Landles et al., 2010; Sathasivam et al., 2013). Overexpression of mutant Httex1 alone with polyQ length ranging from 80 to 175 reproduces many aspects of HD pathology in various animal and cellular models, including the formation of huntingtin inclusions (Mangiarini et al., 1996; Martindale et al., 1998; Scherzinger et al., 1999; Barbaro et al., 2015).

Although increasing evidence suggests that mutant Htt aggregation and toxicity play central roles in HD's pathogenesis, the molecular events responsible for triggering mutant Htt aggregation, the nature of the toxic species, and the mechanisms by which they cause neurodegeneration remain unknown. Initial efforts focused on disentangling the relationship between Htt aggregation and toxicity and HD have focused on identifying modifiers of Htt aggregation based on targeting the polyQ repeat domain, which has been shown to be the primary sequence responsible for initiating Htt aggregation. However, recent studies suggest that post-translational modifications (PTMs) in close proximity or far from the polyQ domain have the potential to modify not only mutant Htt levels and functions, but also its aggregation and toxicity (Thompson et al., 2009; Aiken et al., 2009; Steffan et al., 2004; Gauthier et al., 2004; Warby et al., 2005; Yanai et al., 2006; Jeong et al., 2009; O'Rourke et al., 2013). Interestingly, several of these PTMs occur within the first N-terminal 17 amino acids of Htt, directly flanked by the polyQ domain (**Figure 1A**). These modifications include phosphorylation at multiple serine and threonine residues (T3, S13, and S16), acetylation, ubiquitination, and SUMOylation at selected lysine residues (K6, K9, K15) (**Figure 1A**). Mutating both S13 and S16 to aspartate to mimic the phosphorylation was shown to reverse the pathology of mutant Htt in an HD mouse model (Gu et al., 2009) and modulate Htt aggregation in different cellular models (Mangiarini et al., 1996; Martindale et al., 1998; Scherzinger et al., 1999; Barbaro et al., 2015). Recently, we showed that phosphorylation at T3, S13, and/or S16 inhibited the aggregation of WT and mutant Httex1 *in vitro* (Chiki et al., 2017; DeGuire et al., 2018).

Furthermore, TBK1 (TANK-binding kinase 1) mediated phosphorylation of mutant Httex1 at S13 and S16 lowers its levels and results in a significant reduction in Httex1 aggregation and inclusion formation in different cellular models and a *C. elegans* model of HD (Hegde et al., 2020). These findings, combined with our observation that the levels of phosphorylated Htt at T3 are decreased in pathological conditions (Cariulo et al., 2017), suggest that phosphorylation within the Nt17 domain protects against mutant Htt aggregation. Similarly, SUMOylation at multiple N-terminal lysine residues inhibits Httex1 aggregation *in vitro* (Sedighi et al., 2020). Interestingly, although lysine acetylation at K6, K9, or K15 does not significantly alter the aggregation profile of Httex1 (Chiki et al., 2017), acetylation at K6 was shown to significantly inverse the inhibitory effects of phosphorylation at T3 on aggregation (Chiki et al., 2017). Similarly, phosphorylation of S13 and S16 was shown to regulate Httex1 acetylation at K9 (Thompson et al., 2009), and ubiquitination/SUMOylation at K6 and K9 (Thompson et al., 2009; O'Rourke et al., 2013). Finally, competing modifications, such as ubiquitination and SUMOylation, were shown to exert different effects on Htt levels and degradation (Steffan et al., 2004). Taken together, these findings, combined with the fact that multiple reversible PTMs cluster within eight amino acids in Nt17, suggest that many of these modifications act in concert rather than individually and that the Htt PTM code involves crosstalk between different PTMs, in particular, those that exist close to each other.

One PTM that remains unstudied and its effect on mutant Htt aggregation remains unknown is the oxidation of methionine 8 (oxM8). This residue is highly conserved (DiGiovanni et al., 2016), and the levels of M8 oxidation were shown to be higher in the R6/2 HD mouse model (Mitomi et al., 2012). However, the absence of immunochemical methods to detect M8 oxidation has precluded studies aimed at understanding its relevance and potential roles in HD neuropathology. Despite this, methionine oxidation is one of the most common modifications that occur under oxidative stress (Hoshi and Heinemann, 2001), and has been linked to the pathogenesis of HD (Kumar and Ratan, 2016) and other neurodegenerative diseases. Elevated levels of oxidative stress markers, including reactive oxygen species (ROS), were found in blood (Chen et al., 2007) and postmortem brains of HD patients (Polidori et al., 1999). Although there is no direct evidence establishing that M8 is oxidized, several studies have suggested that this could occur under oxidative stress conditions and that M8 oxidation could act as a sensor of ROS to regulate Htt phosphorylation and localization (DiGiovanni et al., 2016; Son et al., 2019). Mitomi and colleagues (Mitomi et al., 2012) showed, using hydrogen peroxide (H_2O_2) mediated oxidizing conditions *in vitro*, that oxidation at M8 occurs only post-aggregation on non-soluble forms of mutant Httex1. Recent studies based on quantitative

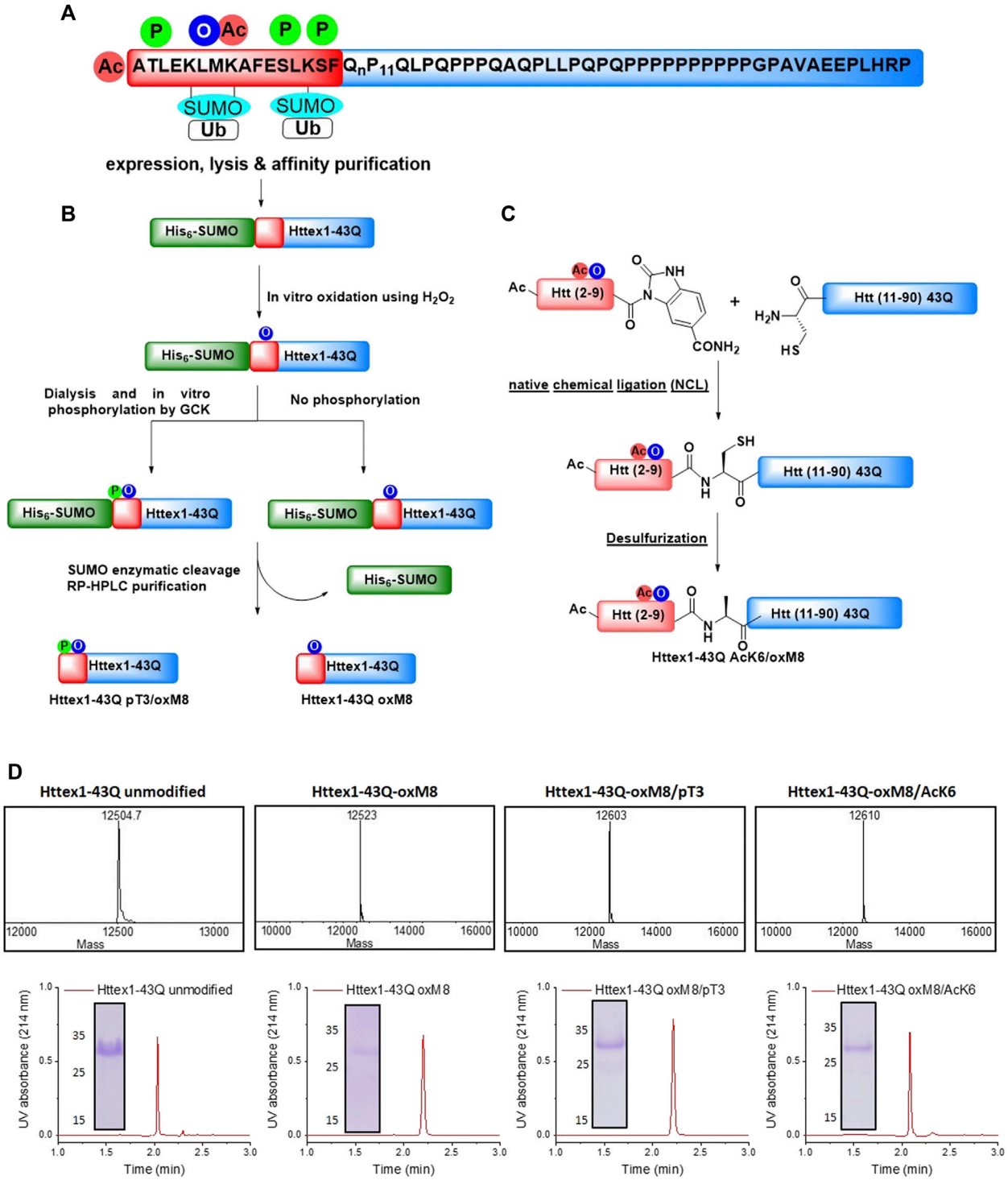


FIGURE 1 | Chemical and enzymatic methods for the generation of Httex1-43Q-oxM8, Httex1-43Q-oxM8/pT3, and Httex1-43Q-oM8/AcK6. **(A)** Schematic presentation of the Httex1 sequence, highlighting the cluster of PTMs in the Nt17 domain. **(B)** Schematic presentation of the SUMO-based strategy used to produce Httex1-43Q-oxM8 and Httex1-43Q-oxM8/pT3. The SUMO-Httex1-43Q was produced and purified by Nickel IMAC purification and subsequently oxidized or both oxidized and phosphorylated by GCK. Next, the SUMO tag was removed by ULP1, and the desired protein was purified by RP-HPLC. **(C)** Schematic representation for the semisynthetic strategy used for the generation of mHttex1-oxM8/AcK6 (adapted from Chiki et al. (2017)). **(D)** Characterization of Httex1-43Q-oxM8, Httex1-43Q-oxM8/pT3, Httex1-43Q-oM8/AcK6, and unmodified Httex1-43Q by ESI/MS, UPLC, and SDS-PAGE.

NMR showed that TiO_2 nanoparticles induced M8 oxidation and abolished the aggregation of a model peptide of Httex1, and reduced the model peptide binding to lipids micelles (Ceccon et al., 2018). However, most of these studies were carried out either using Nt17 peptides, GST-tagged Httex1 protein (Mitomi et al., 2012), or a model peptide consisting of only the Nt17 linked to short polyQ tracts of only 7 or 10 glutamines, lacking the proline-rich domain and most the C-terminal domain of Httex1 (Ceccon et al., 2018). Additionally, in some studies, the H_2O_2 was not removed from the solution prior to the aggregation experiments.

Here, we used an integrative approach combining biophysical and computational studies to gain insight into the effect of M8 oxidation and potential crosstalk between M8 oxidation with neighboring PTMs in regulating the structure and aggregation of mutant Httex1. Toward this goal, we used chemoenzymatic approaches to produce site-specific modified Httex1 proteins bearing different PTMs, including: 1) oxM8, 2) oxM8 and phosphorylation at T3, and 3) oxM8 and acetylation at K6. With these proteins in hand, we performed systematic studies to understand the effect of M8 oxidation and its crosstalk with pT3 and AcK6 in modulating the structure, aggregation, and fibrillar morphology of Httex1. To further understand the structural basis underlying the effects of these modifications on mutant Httex1 aggregation, we performed circular dichroism (CD) analysis and atomistic molecular simulations to investigate how these different combinations of PTMs influence Nt17 structure and helicity. Our findings provide new insight into the PTM code of Nt17, and the differential effects of PTM-induced conformational changes on mutant Httex1 (mHttex1) aggregation.

RESULTS

Production of Oxidized Mutant Httex1 Proteins

Httex1 contains two methionine residues within its Nt17 domain. The N-terminal methionine is cleaved by aminopeptidase resulting in N-terminally acetylated alanine as the first residue (Arnesen et al., 2010). This means that once expressed, Httex1 contains only a single methionine residue at position 8 (M8). Investigating the biological role of M8 oxidation is more challenging than other Nt17 PTMs, such as phosphorylation, because there are no natural amino acids that mimic methionine oxidation and factors that are known to induce protein oxidation, such as oxidative stress, are not always chemoselective. Therefore, to investigate the effect of M8 oxidation on Httex1 aggregation, we sought to prepare mutant Httex1 proteins that are homogeneously oxidized at M8. Several protocols have been used to oxidize methionine residues in proteins *in vitro*, including treatment with H_2O_2, transition metal ions, 2,2-azobis (2-amidinopropane) dihydrochloride (AAPH), tert-butyl hydroperoxide (t-BHP), and UV exposure (Keck, 1996; Wasylaschuk et al., 2007; Ji et al., 2009; Luo et al., 2011).

In the context of Httex1, previous studies have shown that treatment with H_2O_2 allows for efficient oxidation of M8 within a GST-Httex1 fusion protein (Mitomi et al., 2012), Nt17 peptides (DiGiovanni et al., 2016), or Httex1 model peptides consisting of Nt17 domain with ten additional glutamine residues (Nt17Q10) (Ceccon et al., 2018). Some of the limitations of these studies are: 1) none of the proteins used represent the native sequence of Httex1; 2) the extent of oxM8 was not always confirmed by mass spectrometry or other methods; or 3) in some studies, the structural or aggregation properties of the protein and peptide models was assessed in the presence of the oxidizing agent, H_2O_2. It is known that incubation of proteins under such harsh conditions for an extended period might modify other residues within the sequence, such as histidine or phenylalanine (Li et al., 1995), which could alter the sequence and biophysical properties of Httex1.

To overcome these limitations, we developed a strategy for producing native mutant Httex1 specifically oxidized at M8 (oxM8). We took advantage of recent advances from our lab that allow for the generation of milligram quantities of highly pure mutant Httex1 fused to the SUMO protein (Reif et al., 2018) (**Figure 1B**). The desired PTMs (oxidation or phosphorylation) are then introduced, using enzymatic or chemical approaches, into Nt17 of the fusion protein, followed by removal of the SUMO protein and purification of the modified mutant Httex1 by RP-HPLC (Chiki et al., 2020).

Mutant Httex1, with 43 glutamine residues (mHttex1) fused to a SUMO tag protein at its N-terminal (SUMO-mHttex1), was expressed and purified as previously described (**Supplementary Figure 1A,B**) (Reif et al., 2018). Next, the fusion protein was subjected to oxidation using 400 mM of H_2O_2 in 50 mM Tris, 500 mM NaCl, 500 mM Imidazole, pH 7.5. The oxidation reaction was monitored over time by ESI/MS (**Supplementary Figure 1C**). For each time point, an analytical scale cleavage reaction of the SUMO tag using the ULP1 enzyme was performed, and the resulting cleaved product was analyzed by ESI/MS. As shown in **Supplementary Figure 1C**, the M8 residue was completely oxidized after 2 h of incubation with H_2O_2. The oxidized SUMO-mHttex1 was then subjected to ULP1 cleavage. Upon verifying the complete removal of the SUMO tag, *m*-Httex1-oxM8 was immediately purified by reverse phase HPLC (RP-HPLC) (**Supplementary Figure 1E**). The fractions containing the protein of interest were then pooled together and lyophilized. The final purity of the mHttex1-oxM8 was verified by ESI/MS, UPLC, and SDS-PAGE (**Figures 1D**).

To investigate the effect of potential crosstalk between oxM8 and other Nt17 PTMs on mutant Httex1 aggregation, we generated mutant Httex1 proteins oxidized at M8 and phosphorylated at T3 (pT3) or acetylated at K6 (AcK6). To produce mHttex1 oxidized at M8 and phosphorylated at T3 (mHttex1-oxM8/pT3), we first generated SUMO-mHttex1-oxM8 as described above. After overnight dialysis to remove the excess of H_2O_2, the fusion protein was co-incubated overnight with GCK kinase, a kinase that we recently reported to efficiently phosphorylate Httex1 specifically at T3 (Chiki et al., 2020). The extent of phosphorylation was monitored by ESI/MS. As shown in **Supplementary Figure 1D**, the ESI/MS spectrum of the SUMO-mHttex1-oxM8 showed an additional +80 Da, indicating the addition of a single phosphate group. The

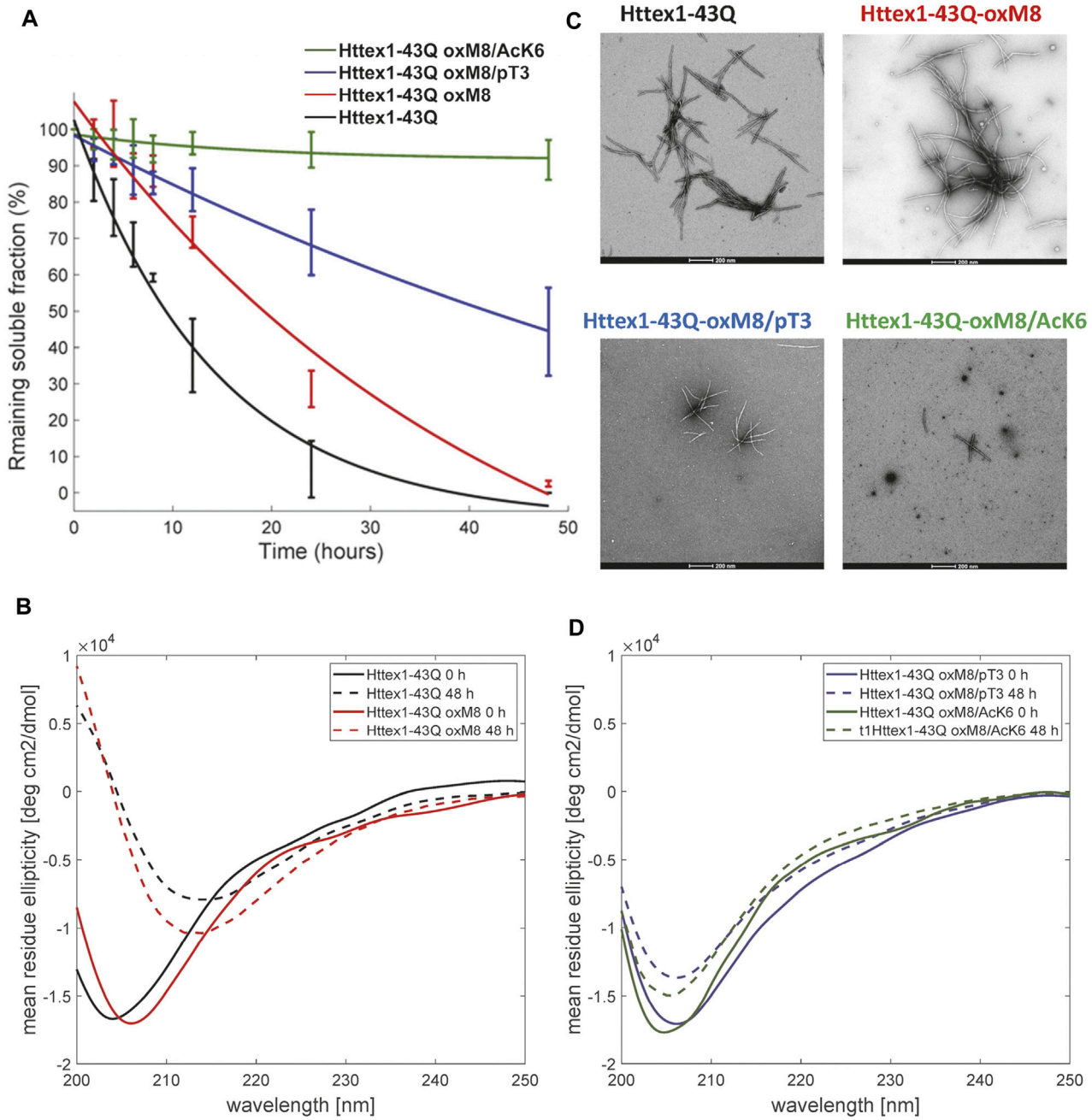


FIGURE 2 | Biophysical studies to elucidate the effect of pT3, OxM8, and AcK6 on mutant Httex1 aggregation. **(A)** Httex1 aggregation studies (at 5 μM) of Httex1-43Q-oxM8, Httex1-43Q-oxM8/pT3, Httex1-43Q-oM8/AcK6, and unmodified mHttex1 monitored by UPLC sedimentation assay to determine the proportion of remaining monomer (error bars, S.D. derived from $n = 3$ independent experiments fitted with exponential decay). **(B)** Secondary structure analysis of unmodified Httex1-43Q and Httex1-43Q-oxM8 by CD at 0 and 48 h of aggregation. **(C)** Electron microscopy after 48 h of aggregation (scale bars = 200 nm). **(D)** Secondary structure analysis of Httex1-43Q-oxM8 and Httex1-43Q-oxM8 by CD at 0 and 48 h post-aggregation initiation.

SUMO-mHttex1-oxM8/pT3 was then subjected to UPL1 cleavage followed by RP-HPLC purification of mHttex1-oxM8/pT3 (**Supplementary Figure 1F**). The protein's purity was verified by ESI/MS, UPLC, and SDS-PAGE (**Figure 1D**).

To produce mutant Httex1 that is both oxidized at M8 and acetylated at lysine 6, we employed a semisynthetic protein strategy (**Figure 1C**) that we previously used to introduce single or multiple PTMs within Nt17 of mHttex1 (Chiki et al., 2020). mHttex1-AcK6/oxM8 was produced using native chemical ligation between the Htt A10C-90 43Q and Ac-2-9-Nbz AcK6/oxM8 peptides (characterization by ESI-MS and UPLC is shown in **Supplementary Figure 2B**), followed by desulfurization to

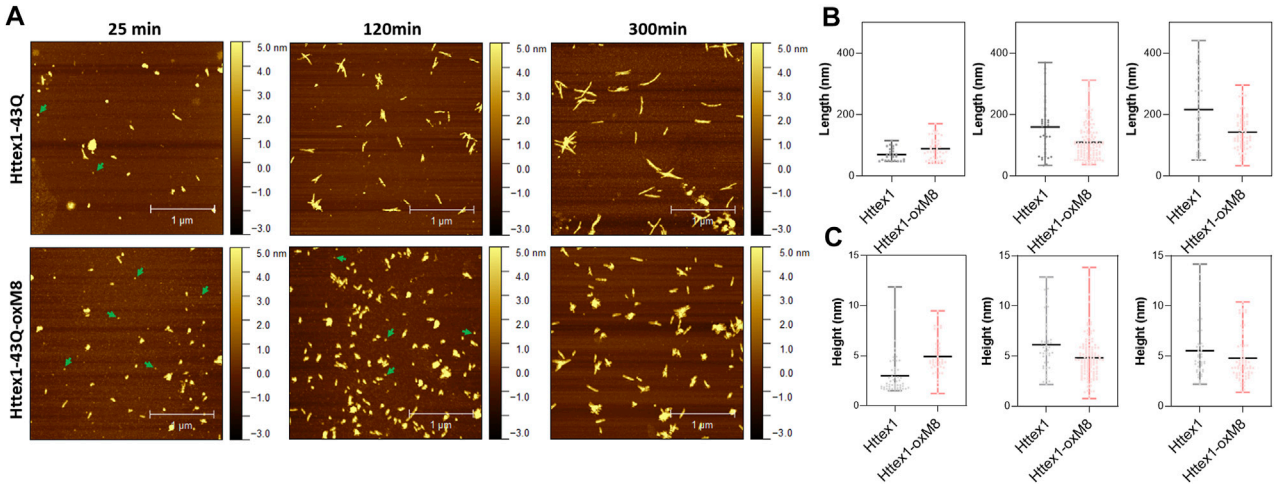


FIGURE 3 | AFM analysis comparison between unmodified and Httex1-43Q-OxM8. **(A)** Time-dependent AFM images of aggregates formed by Httex1-43Q and Httex1-43Q-oxM8 (nm, nanometer), the oligomers are indicated with green arrows. Quantitative analysis of the length **(B)** and height **(C)** of fibrils measured from AFM images. All the plots for the length and height of the Httex1-43Q-oxM8 fibrils differ from the Httex1-43Q with a statistical significance of at least $p < 0.05$ (scattered plot represents the mean and range).

convert cysteine 10 back to the native alanine (**Supplementary Figure 2C**). The purity of mHttex1 AcK6/oxM8 was determined by SDS-PAGE, UPLC, and ESI/MS (**Figure 1D**).

Oxidation at M8 Delays the Aggregation of Mutant Httex1

To determine the effect of oxM8 on mutant Httex1 aggregation, mHttex1-oxM8, and mHttex1 were subjected to a disaggregation protocol, as previously reported (Reif et al., 2018). Any remaining aggregates were removed by filtration through a 100 kDa filter before initiating the aggregation. The extent of and kinetics of aggregation were monitored by quantifying the amount of soluble proteins at different time points using a previously described UPLC-based sedimentation assay (Vieweg et al., 2016; O'Nuallain et al., 2006a; Jayaraman et al., 2011). Also, the effect of oxidation on the secondary structure and morphology of the fibrils was also assessed by circular dichroism spectroscopy (CD) and electron microscopy (EM). **Figure 2A** shows the percentage of the remaining soluble protein over time. As expected, unmodified mHttex1 exhibited almost complete depletion of the soluble monomer (**Figure 2A**) and full conversion into aggregates after 48 h, which was confirmed by a shift in the CD spectra from random coil to a *ß*-sheet structure (**Figure 2B**) and the presence of mature long fibrils (**Figure 2C**). In contrast, mHttex1-oxM8 showed a delay in aggregation compared to the unmodified mHttex1 (**Figure 2A**). The aggregation of mHttex1-oxM8 showed 68 and 39% remaining monomer after 12 and 24 h, respectively, compared to 40 and 13% for the unmodified mHttex1 (**Figure 2A**). However, after 48 h, both proteins exhibited complete aggregation as discerned by the complete disappearance of soluble Httex1 (**Figure 2A**) and the CD spectra of both proteins, which showed a signal that corresponds to predominantly *ß*-sheet-rich structures (**Figure 2B**). The fibrils formed by mHttex1 oxM8 after 48 h were similar to those formed by the unmodified mutant Httex1, suggesting that oxM8 influences aggregation's kinetics but not the final structure of the fibrils (**Figure 2C**).

Interestingly, we consistently observed the presence of a population of oligomers for the oxidized Httex1 at early time points (**Figure 3** and **Supplementary Figure 3**). These oligomer populations were observed in a higher amount after 4 h of aggregation by EM, at which unmodified mHttex1 already showed fibrils formation (**Supplementary Figure 3**). After 6 h, mHttex1-oxM8 showed a mixture of oligomers and short fibrils (**Supplementary Figure 3**), whereas only fibrils were observed in the case of the unmodified mHttex1. To further validate our observations and quantitatively assess differences in dimensions and morphological properties of the aggregates formed by the two proteins, we performed time-dependent atomic force microscopy (AFM) studies. Consistent with our EM analysis, unmodified mHttex1 and mHttex1-oxM8 showed the presence of both fibrils and oligomers at 25 min, as shown in **Figure 3A**. Interestingly, at 2 h, mHttex1 predominantly formed longer fibrils with a mean length of 167 nm compared to mHttex1 oxM8, which formed a mixture of oligomers and short fibrils (111 nm). Similarly, after 6 h of incubation, unmodified mHttex1 formed fibrils with a mean length of 215 nm and a height of 5.5 nm (**Figure 3**), whereas mHttex1 oxM8 samples showed a mixture of shorter fibrils and protofibril-like structures with average lengths and heights of 140 and 4.8 nm, respectively (**Figure 3**). These observations are consistent with the slower aggregation kinetics observed for the mHttex1 oxM8 in the sedimentation assay. Taken together, these results suggest that the delay in the aggregation of the oxidized mutant Httex1 could be associated with the formation and accumulation of oligomers (on- or off-pathway), that eventually convert into fibrils at later stages.

Crosstalk Between M8 Oxidation and Phosphorylation at T3 or Acetylation at K6

To determine the effect of PTM crosstalk on aggregation within the Nt17 domain, we investigated and compared the aggregation kinetics and properties of Httex1-oxM8/pT3 and -oxM8/AcK6. Previously, we showed that >60% of mHttex1-pT3 remains soluble after 48 h of aggregation, indicating an inhibitory effect of T3 phosphorylation on mutant Httex1 *in vitro* (Chiki et al., 2017). As shown in **Figure 2**, the oxidation of M8 does not seem to influence the inhibitory effect of phosphorylation at T3. Even after 48 h of incubation, ~50% of the protein remained soluble, whereas more than 98% of the unmodified proteins have been converted to fibrillar aggregates (**Figure 2A**). The CD spectrum showed that the remaining protein retains a random coil signature, suggesting that the remaining soluble protein represents monomers or disordered soluble oligomers (**Figure 2D**). Analysis by EM showed the accumulation of oligomers and short fibrils similar to those we previously reported for mHttex1-pT3 (Chiki et al., 2017). On the other hand, oxidation of M8 in the context of the K6 acetylated mutant Httex1 dramatically altered the aggregation properties of the protein. Previously, we have shown that acetylation of K6 or K9 does not significantly influence the aggregation of mutant Httex1. However, as shown in **Figure 2A**, the presence of both acetylation at K6 and methionine oxidation at M8 (mHttex1-oxM8/AcK6) results in significant inhibition of mutant Httex1 aggregation. Only 8% of mHttex1 oxM8/AcK6 aggregated after 48 h (**Figure 2A**) as determined by the sedimentation assay and confirmed by CD and EM, which show predominantly disordered conformation and the accumulation of only oligomers and small fibrils (**Figure 2C,D**). These results suggest that the addition of oxM8 reversed the aggregation of mHttex1-AcK6 and illustrates how the combination of different Nt17 PTMs could differentially influence the aggregation properties of the protein.

Oxidation at M8 Decreases the Helicity of the Nt17 Peptides

To gain insight into the mechanisms by which oxM8 and its combination with other PTMs influence the aggregation of mutant Httex1, we investigated how the individual Nt17 PTMs and different combinations of PTMs influence on the structural properties of the Nt17 domain. We confirmed that Nt17-AcK6 exhibited increased helical content (Pastore et al., 2019) compared to the Nt17-WT peptide (**Supplementary Figure 4**) (Pastore et al., 2019). Additionally, by CD, we observed that Nt17-pT3 has a higher propensity to adopt helical conformation compared to other Nt17 peptides (**Supplementary Figure 4**, **Table 1**). When oxM8 was added to Nt17-AcK6 and Nt17-pT3, we observed a reduced CD signal at 222 nm for the Nt17-oxM8/pT3 and Nt17-oxM8/AcK6 peptides (**Supplementary Figure 4**), indicating a decrease in the helical content which was confirmed by determining their secondary structure content from the CD spectra (See **Supplementary Table 1**).

Furthermore, to investigate if the decrease in the helicity induced by oxM8 persists even at a high concentration of the peptides, we conducted CD studies (**Supplementary Figure 5A**) at different Nt17 peptide concentrations (60, 90, 120, 150, and 200 μM). As shown in **Supplementary Figure S5**, all of the peptides showed a small increase in the helical content with increasing concentrations (**Supplementary Figure 5B**). oxM8 slightly decreased the helical content of Nt17-pT3 peptides over different concentrations. The effect of oxM8 was stronger when it was added to the Nt17-AcK6 peptide; it significantly decreased its helical content over the whole range of tested concentrations (**Supplementary Figure 5B**). Taken together, these results suggest that the oxidation of M8 decreases the helicity of the Nt17 in the presence of PTMs, and this can play a role in modulating the aggregation of mutant Httex1.

To better understand how oxM8 and other PTMs influence the conformational properties of Nt17, we conducted atomistic molecular dynamics (MD) studies on Nt17. For these studies, we decided to include the first two glutamine residue as preliminary studies suggested that they could be involved in stabilizing Nt17 conformations. Previously, it has been shown that there is no difference in the overall structure between the Nt17 and the Nt19 (Nt17 + QQ) (Chiki et al., 2017). We simulated the Nt19 with single and double PTMs to analyze the crosstalk between oxidation at M8, and phosphorylation at T3 or S13, or acetylation at K6. The simulations were run for a total of 13 μs using the CHARMM36m force field and a modified TIP3P water model (Jorgensen et al., 1983), which have been shown to be effective for the simulation of intrinsically disordered proteins. The phosphate group of phosphorylated threonine was in a dianionic form. The parameters for phosphorylated threonine and acetylated lysine were obtained from the CHARMM36 parameters for modified residues (see Methods).

Both CD and MD showed that oxidation at M8 decreased the overall helicity of Nt19 Nt19-pT3 and Nt19-AcK6 (**Table 1**). The CD experimental data showed that for Nt19-pT3, oxidation at M8 resulted in a decrease of overall helicity by 13%, while a more significant effect was observed for Nt19-AcK6, whose helicity decreased by 69% (**Table 1**). Likewise, the MD results also showed that oxidation at M8 decreased the overall helicity of Nt19, with a more significant decrease in helicity for Nt19-AcK6 (**Table 1**). The residue-wise helicity calculated from MD showed that the core region of Nt19 (residue 4–13) had a more significant decrease, while the N-term residues (1–3) were less affected (**Figure 4**). Overall, oxM8 has a higher impact on Nt19-AcK6 and on the core region of Nt19 in general. These observations are consistent with our aggregation studies and CD analysis data.

The Formation of a Short N-Terminal Helix in the Simulations Correlates With Suppressed Aggregation *In Vitro*.

The MD results were further analyzed to explore the structure and dynamics of PTM crosstalk and gain insight into the structural basis underlying the differential effects of Nt17 PTMs on the aggregation of mutant Httex1. We first performed a principal components analysis of the backbone dihedral angles (dPCA) (Mu et al., 2005) to distinguish among the various conformations adopted by the peptides with different PTMs. The first two components that depict a large

TABLE 1 | Helicity of Nt19. CD and MD both showed decrease in helicity when M8 was oxidized.

Nt19	WT *vs.* oxM8	pT3 *vs.* oxM8/pT3	AcK6 *vs.* oxM8/AcK6
MD % change in helicity	−20	−21	−60
CD % change in helicity	NA	−13	−69

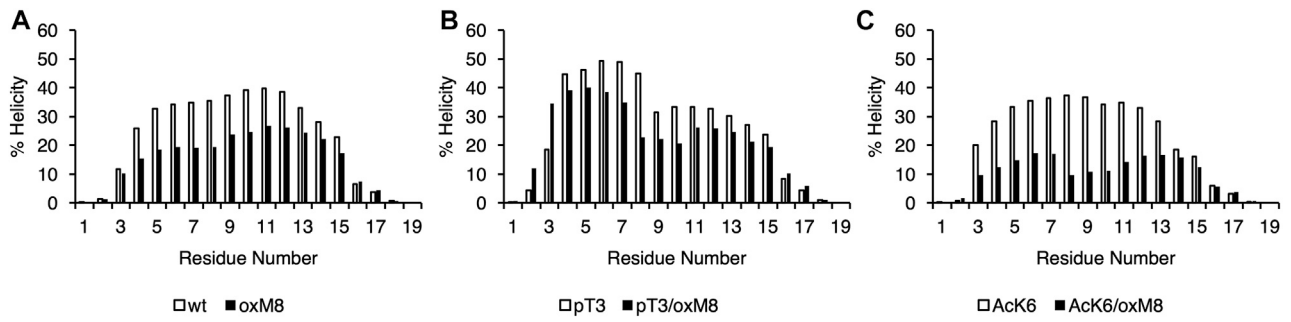


FIGURE 4 | Helical propensity of Nt19. The secondary structure was obtained with STRIDE from the MD trajectories of **(A)** WT and oxM8 **(B)** pT3 and pT3/oxM8 **(C)** AcK6 and AcK6/oxM8. The helicity for each residue was calculated based on the sum of the total number of frames that are in *a*-helix or 3_{10} helical conformations divided by the total number of frames.

part of the differences in the conformations were used for generating an energy map (**Figure 5**). dPCA showed well-separated minima corresponding to various helical forms and revealed a very broad and shallow minimum corresponding to a wide-range of unfolded conformations connected by very low barriers.

To investigate the connection between the major conformations and the aggregation rates of mHttex1 with different PTMs, we classified the PTMs into three groups based on their effects on the aggregation of mHttex1 and the helicity of Nt17. In Group 1, we grouped PTMs that showed higher helicity and had a lower aggregation rate. In Group 2, PTMs that showed lower helicity but also lower aggregation rate, while in Group 3, PTMs that showed an aggregation rate similar to unmodified Httex1. dPCA showed that for all PTMs in Group 1, there were several local minima corresponding to conformations where the first 8 N-terminal residues adopted a helical conformation while the rest remained disordered (**Figure 5**). When T3 was phosphorylated, and when both T3 and oxM8 modifications were present, this short N-terminal helix was 7 times more populated than for the unmodified peptide. Meanwhile, the N-terminal helix was 4 times more populated in oxM8/AcK6 compared to the unmodified peptide, yet the abundance of all other helical forms was reduced by half compared to the unmodified peptide (**Figure 6**). For Group 2, the abundance of the N-terminal helix of pS13 and oxM8 was similar to the unmodified peptide, but a 94 and 30% decrease was observed for other types of helices in pS13 and oxM8. For Group 3, the abundance of either short N terminal helix or other types of helices in AcK6 was similar to that of the unmodified peptide. By comparing the helicity data with aggregation data, it then appears that the higher abundance of a short N-terminal helix and lower abundance of other Nt17 helical conformations, like a long helix or short C-terminal helix in the MD simulations, correlates with lower aggregation rates of mHttex1.

Oxidation of M8 Increased Abundance of the Short N-Terminal Helix

To analyze how the PTMs modulate the stability of the short N-terminal helix, we searched for the key differences in structural dynamics in simulations of peptides with oxidized and reduced methionine. For this, we computed the distance matrix between Cα atoms for all frames of each MD simulation (**Supplementary Figure 6**), compared the distributions of Cα pairwise distances through Kullback-Leibler divergences of two distance matrices, as described in Methods, and plotted the divergences into heat maps. Heat maps comparing pT3 with WT and oxM8/pT3 with WT (**Supplementary Figure 6**) show a significant change in the distance between the first eight residues and residue 1 or 2. Heat maps comparing WT and oxM8/AcK6 show that significant changes occur at a distance between residue 4 and residue 9/10.

The distance distributions (**Figure 7A**) are similar for pT3 and pT3/oxM8 but were different for WT. The peak for oxM8/pT3 and pT3 corresponds to a conformation where the phosphate group of pT3 forms a salt bridge with the NH_3^+ group of M1 (**Figure 7C**), which is not observed in WT, where M1 and T3 remained apart. A previous study by Yalinca et al. has shown that phosphorylation stabilizes helical conformation between A2 and L7, because of the increased negative charge that counterbalances the helix dipole (Yalinca et al., 2019). Also, the distribution of distances between Cα of L4 and K9 in WT, AcK6, and oxM8/AcK6 (**Figure 7B**) is similar for WT and AcK6 but different for AcK6/oxM8. The peak for WT and AcK6 corresponded to the distance between L4 and K9 when they were in a helix conformation. The peak for AcK6/oxM8 corresponds to a conformation where there was a turn at M8, so the distance between L4 and K9 decreased.

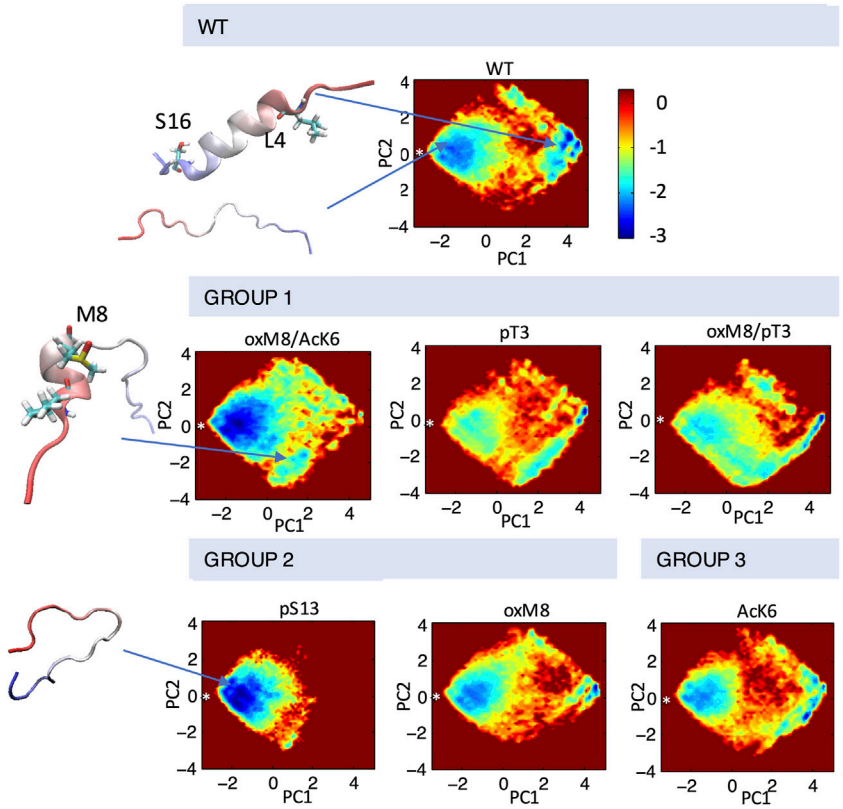


FIGURE 5 | Conformational ensembles of Nt19 with different PTMs. dPCA was performed on dihedral angles calculated for the Nt19 backbone, and a free energy map was generated with the first two components. The white asterisk indicated the starting structure (in extended conformation). The major structures that correspond to local minima include fully disordered conformation, short N-term helix, and long helix. The peptides are classified based on the effect of each PTM on helicity and aggregation rate: oxM8/Ack6, pT3, and oxM8/pT3 increased Nt19 helicity and decreased aggregation rate (GROUP 1), pS13 and oxM8 decreased Nt19 helicity and decreased aggregation rate (GROUP 2), AcK6 had no significant effect on aggregation rate (GROUP 3).

Another aspect we examined is the distance between the sulfur atom of M8 and the center of the aromatic ring of F11 because a previous study (Valley et al., 2012) has shown that when this distance is below 6 Å, an interaction between sulfur of M8 and the aromatic ring could stabilize the structure. For WT, pT3, and AcK6, a peak below 6 Å (**Figure 7D**) was observed for the distance between M8's sulfur atom and F11's aromatic ring. When M8 is oxidized, the distance distribution shifts toward larger values, suggesting the disruption of the stabilization effects brought by these interactions.

The residue pair distance analysis for WT, pT3, AcK6, oxM8/pT3, and oxM8/AcK6 shows that the change in abundance of short N-terminal helix and other helix resulted from 1) stabilization of N-terminal helix by phosphate group of T3 (pT3, oxM8/pT3) and 2) breaking the long helix between residue 8 and 9 with oxidation at M8 (oxM8/AcK6) and destabilization of the long helix due to disruption of the interaction between residue 8 and 11.

DISCUSSION

Among all Htt Nt17 PTMs, methionine oxidation remains the least well understood and studied. To address this knowledge gap, we first developed an efficient method to produce milligram quantities of highly purified untagged mutant Httex1 that is site-specifically oxidized at M8. To investigate the crosstalk between M8 oxidation and neighboring PTMs, we used protein semisynthetic and chemoenzymatic strategies developed by our group (Chiki et al., 2017) to generate mutant Httex1 proteins bearing oxidized M8 and acetylated K6 (mHttex1-AcK6/oxM8) or phosphorylated at T3 (mHttex1-AcK6/pT3). These advances enabled us to investigate, for the first time, the role of crosstalk between these different PTMs in regulating Nt17 conformation and mutant Httex1 aggregation *in vitro*. We choose K6 and T3 because of their close proximity to M8 and the fact these are the only two PTMs that have been shown to significantly influence the helicity of Nt17 and modify the effect of neighboring PTMs on Nt17 structure and Httex1 aggregation (Chiki et al., 2017).

Our results demonstrate that oxidation at M8 delayed the aggregation of mHttex1 but did not alter the morphology of the structure of the Httex1 fibrils. We hypothesized that this delay is caused by methionine oxidation-induced formation of oligomers at the early stages of the aggregation, which undergo relatively slower conversion to fibrils. As shown in **Figure 3** and highlighted

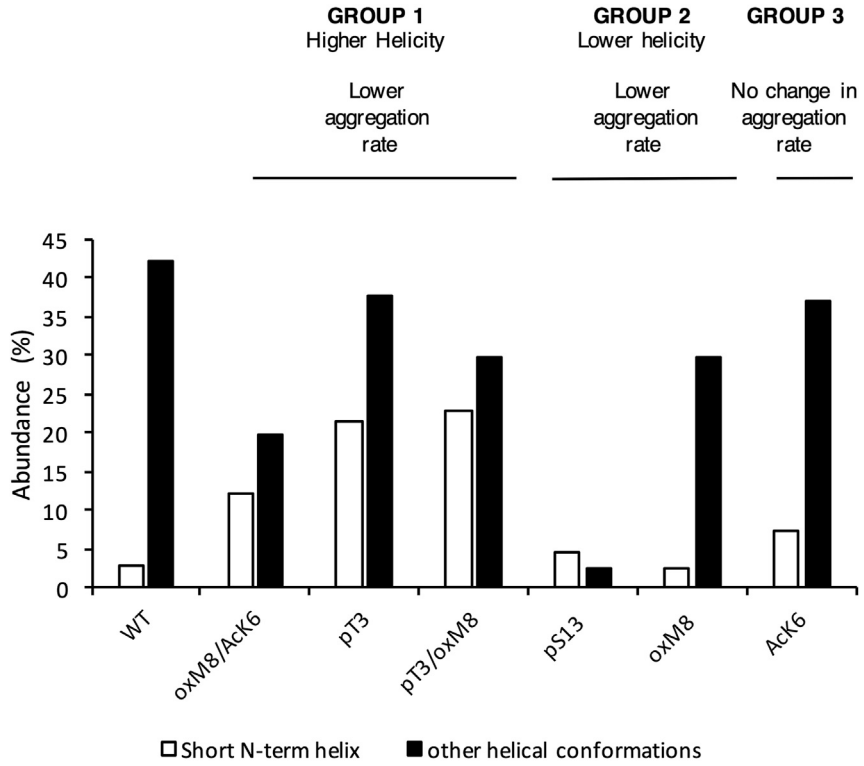


FIGURE 6 | The abundance of N-term helix was high for PTMs that delayed aggregation. The abundance of short N-terminal helix and other helical forms was calculated for each peptide. Peptides with a lower aggregation rate showed a higher abundance of N-terminal helix and a lower abundance of other helical conformations.

in previous studies (Vieweg et al., 2016; Chiki et al., 2017; DeGuire et al., 2018), oligomers are rarely observed during the fibrillization of the unmodified mutant Httex1 due to its high propensity to misfold and fibrillize. In contrast, our imaging studies of oxM8 aggregation consistently revealed the transient accumulation of oligomers at the early stages of the aggregation process (**Supplementary Figure 3**).

Although acetylation at K6 was previously shown not to influence the aggregation of mutant Httex1, when this modification is combined with oxidation at M8, we observed a strong inhibitory effect on mutant Httex1 aggregation and the accumulation of mainly oligomeric and short fibrillar structures. In contrast, the combination of methionine oxidation and phosphorylation at T3 did not modify the aggregation inhibitory effect induced by pT3. Our findings are consistent with previous observations on the effect of methionine oxidation on the aggregation of other amyloid-forming proteins. For example, oxidation at M35 attenuates the aggregation amyloid-β(1–40) (Friedemann et al., 2015; Gu and Viles, 2016), Aβ1-42, and the highly amyloidogenic Arctic Aβ1-40 variant (Johansson et al., 2007). Similarly, oxidation of methionine 109 and 112 in the prion peptide PrP106-126 inhibits its fibrillization *in vitro* (Bergström et al., 2007). Furthermore, methionine oxidation was shown to inhibits the aggregation of alpha-synuclein protein and promote the formation of stable oligomers (Zhou et al., 2010). In the case of Httex1, previous studies have shown that oxidation at M8 abolished aggregation of a model peptide consisting of Nt17 plus ten glutamine residues (Nt17Q10) (Ceccon et al., 2018).

To understand the structural basis underlying the effect of methionine oxidation and its crosstalk with pT3 and AcK6 on mHttex1, we performed CD and MD analyses of the conformational properties of the unmodified Nt17 peptide and Nt17 peptides bearing pT3, or AcK6 in the presence or absence of the oxM8. We observed that methionine oxidation reduced the helical content of Nt17, independent of co-occurring PTMs and peptide concentration. Consistent with our previous studies (Chiki et al., 2017; DeGuire et al., 2018), we observed that the effect of PTMs on the overall helicity of the Nt17 peptide did not correlate with their impact on the aggregation propensity of mutant Httex1 *in vitro*. However, it was noticed that all PTMs that are associated with a slower aggregation rates have a higher abundance of short helices at the first 8 N-term residues and a lower abundance of other helical conformations. According to the simulations, the higher abundance of short N-term helices in Nt17-pT3 and oxM8/pT3 could derive from direct stabilization via a salt bridge between the phosphate group at T3 and the NH_3^+ group of M1, while their higher abundance in Nt17-oxM8/AcK6 could result from breaking long helices with a turn between residue 9 and 10, as well as destabilization of the long helices due to disruption of the interaction between residue 8 and 11 of oxM8/AcK6. The simulation results provide an interesting perspective, i.e., that the overall helicity only provides ensemble-

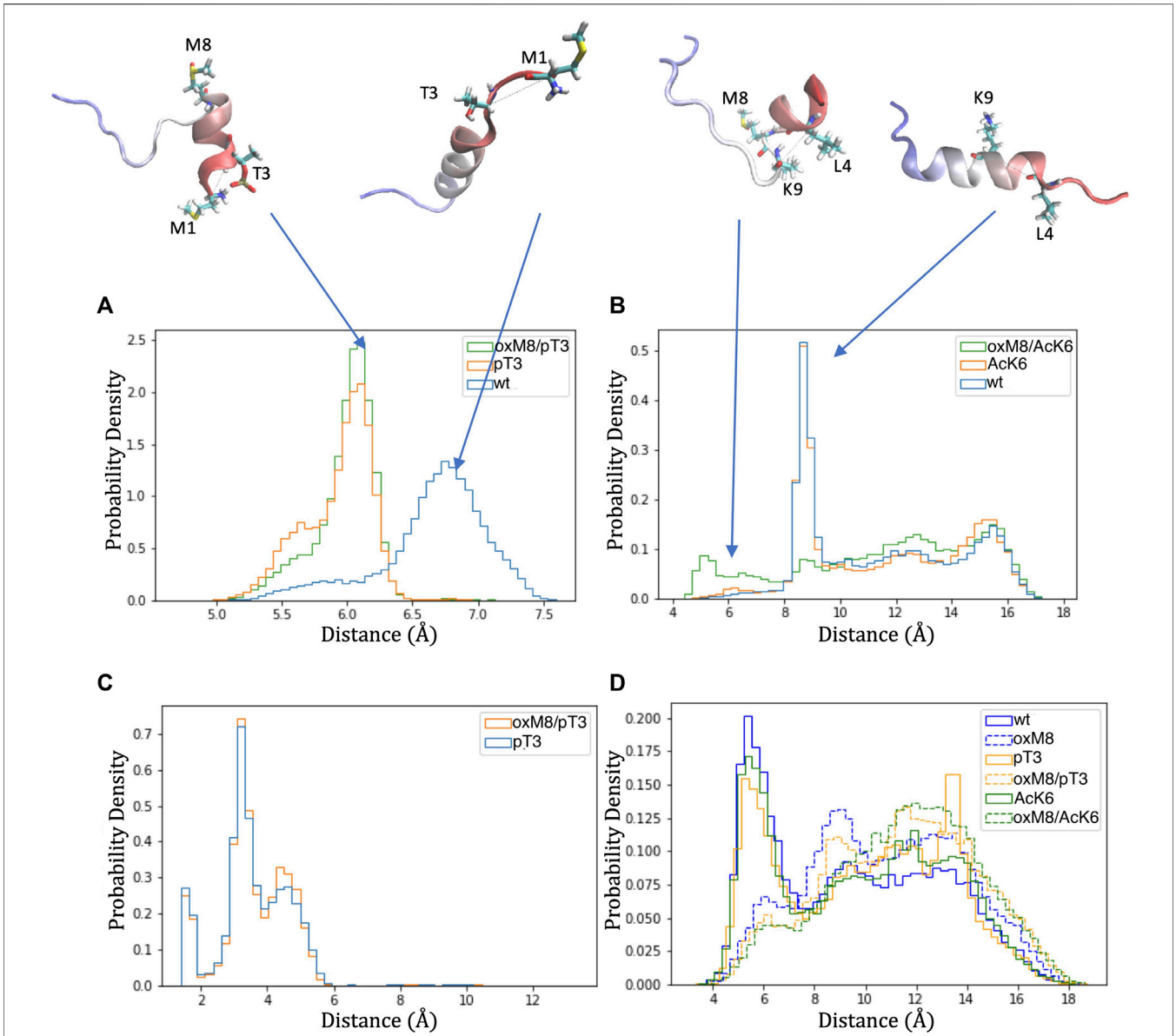


FIGURE 7 | Distance distribution of atom pairs in Nt19 with different PTMs. **(A)** Distance between Cα of M1 and Cα of T3 was lower in pT3 and oxM8/pT3 than in WT. **(B)** The distance between Cα of L4 and K9 was lowered in oxM8/AcK6 compared to WT and AcK6 due to the bend at M8. **(C)** A peak below 4 Å was shown in the distribution of the distance between hydrogens of NH_3^+ and oxygens of the pT3 phosphate group, indicating pT3 might promote ionic interaction at this N-terminal end. **(D)** The distance between the sulfur atom of M8 and the center of the F11 aromatic ring increased in peptides with M8 oxidation, indicating the stabilization by the interaction of sulfur and the aromatic ring was weakened.

averaged structure information; with the same overall helicity as measured by CD, some modified peptides contain a higher abundance of short N-term helix, and some contain a higher abundance of the long helix or short C-term helix forms. Previous studies suggested that the helical conformation of Nt17 is a major driver of the initiation of mutant Htt oligomerization and fibrillization. Our findings suggest Nt17 exists as an ensemble of different helical conformations, among which some might be vital for aggregation, while others, like short N-term helix, have no impact or even disfavor the aggregation process.

Our findings on the effect of the crosstalk between the different Nt17 PTMs highlight the complexity of the Nt17 PTM code and suggest that Htt's normal function and aggregation are likely regulated by a complex interplay between different PTMs (**Figure 8**). A major challenge in addressing this complexity is the large number of possible PTM combinations, even within short stretches of sequences in proteins, such as Nt17. This, combined with the challenges of introducing multiple PTMs into proteins, has led to either abandoning efforts to investigate this complexity or resorting to studying protein fragments. Our findings here show that while working with peptide fragments provides useful

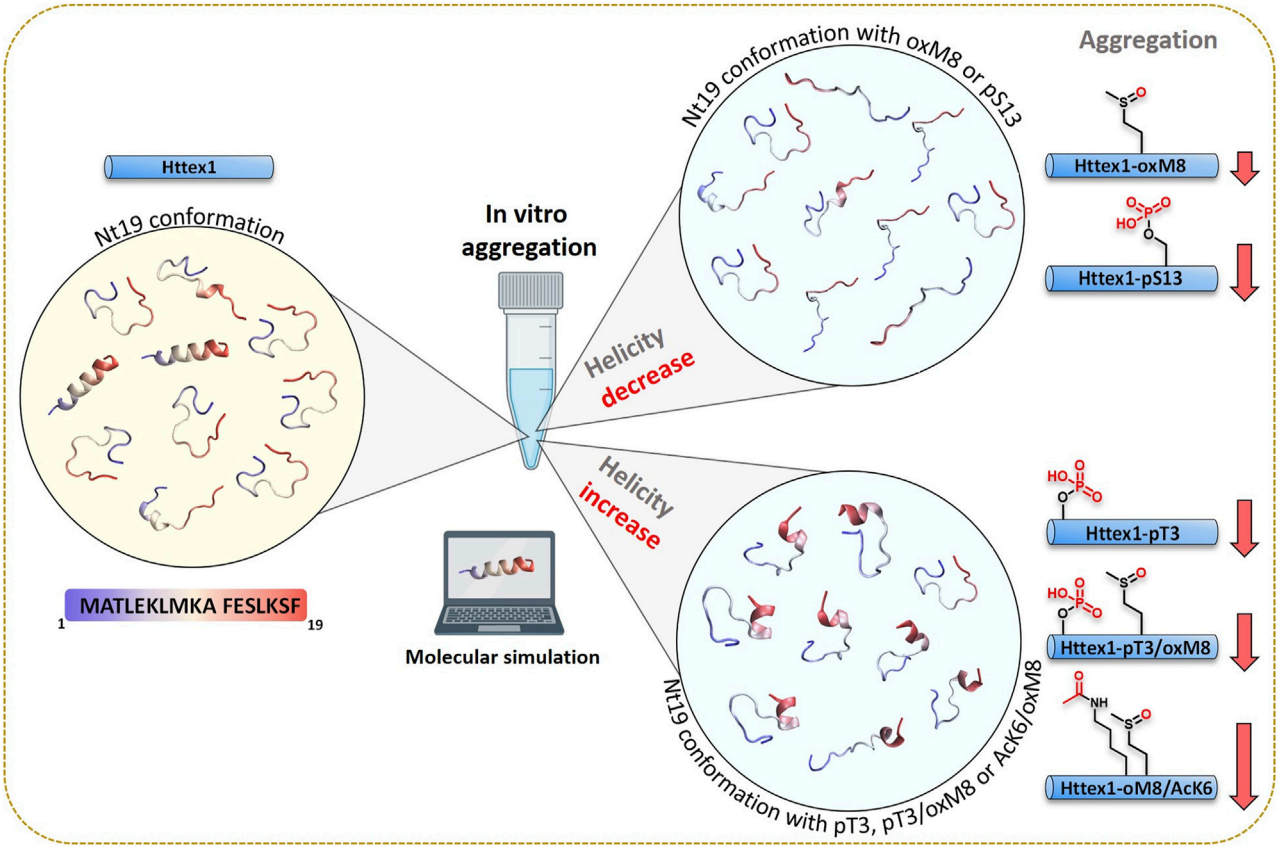


FIGURE 8 | A combined experimental and computational approach unravels the structural basis of PTM crosstalk in regulating aggregation of Huntingtin exon 1.

insights, extrapolations of findings from these fragments to predict their properties in the context of full-length proteins are not straightforward. The work presented here represents our initial efforts in this direction and aims at exploring the extent to which this is possible using peptide systems in which the effect of PTMs can be experimentally and computationally measured and predicted. Also, our work demonstrates the promising potential of integrating experimental and computational approaches for characterizing molecular mechanisms.

MATERIALS AND METHODS

Materials

The pTWIN1 vector containing human Httex1 fused to His6-SUMO was ordered from GeneArt Gene Synthesis (Life Technologies); *E. coli* B ER2566 from NEB; ampicillin, DTT, isopropyl *ß*-D-1-thiogalactopyranoside (IPTG), and hydrogen peroxide solution 30%, imidazole, complete Protease Inhibitor Cocktail, magnesium chloride ($MgCl_2$), magnesium sulfate ($MgSO_4$), and trifluoroacetic acid from Sigma; PMSF from AppliChem; EGTA solution from Boston Bioproducts; Mg-ATP from Cayman; EDTA from Fisher Scientific; Luria Broth (Miller's LB Broth) from Chemie Brunschwig; acetonitrile (HPLC-grade) from Macherey Nagel; spectrophotometer semi-micro cuvettes from Reactolab; C4 HPLC column from Phenomenex; HisPrep 16/10 column from GE Healthcare; GCK kinase (0.5 µg/ul, cat. # BC047865) from MRC PPU Reagents; Uranyl formate (UO2(CHO2)2) and Formvar/carbon 200 mesh, Cu 50 grids from EMS; high-precision cell made of Quartz SUPRASIL 1 mm light path from Hellma Analytics; Dulbecco's Buffer Substance (PBS w/o Ca and Mg) ancienne ref. 47,302 (RT) SERVA from Witech; and 100-kD Microcon fast flow filters from Merck Millipore.

Expression and Purification of SUMO-Httex1-43Q

Expression of His6-SUMO-Httex1-Qn (n= 23 or 43). The expression and purification of His-SUMO-Httex1-43Q were performed as previously reported (Reif et al., 2018). Httex1-43Q with a His-SUMO tag at its N-terminal was cloned into the pTWIN1 plasmid with ampicillin (Amp) resistance. Then, the plasmid was transformed in *E-coli* ER2566, and the resulting transformed bacterial cells were plated onto an agar plate containing ampicillin (Froger and Hall, 2007). Next, 400 ml LB + Amp (100 µg/ml) medium was inoculated with a single colony and incubated at 37°C overnight to start the day after the 12L expression at an optical density (OD_{600}) density of 0.15. When the OD_{600} value was around 0.6, the culture was induced with 0.4 mM IPTG and incubated at 18°C overnight. The cells

were harvested by centrifugation (3,993 rpm, 4°C, 10 min) and resuspended in buffer A (50 mM Tris, 500 mM NaCl 30 mM Imidazole, pH 7,5, 0,65 um filtrated) with PMSF and protease inhibitors. The suspended bacterial pellet was subjected to sonication on ice (70% amplitude, total sonication time 5 min, intervals of 30 s sonication, 30 s pause), followed by centrifugation (39,191 rpm, 4°C, 60 min). The supernatant was filtered (0.45 µm, syringe filters) and passed through a Ni-NTA column. The protein was then eluted with 100% IMAC buffer B (50 mM Tris, 500 mM NaCl 15 mM Imidazole, pH 7.5, 0.65 µm filtrated). The purified fusion protein was kept ice for further reactions. The final yield of Httex1-43Q was 1.23 mg/L of culture.

Production of Httex1-43Q oxM8 and Httex1-43Q oxM8/pT3

SUMO-Httex1-43Q (10 mg in 18 ml) fusion protein was treated with 1 ml of 30% H_2O_2 to perform the oxidation directly on the SUMO fusion protein. After 2 h of the reaction, 30 µl was supplemented with 1 µl of ULP1 (1 mg/ml), and the extent of the oxidation was verified by LC/MS. 1 oxidation corresponds to +16 Da to the molecular weight of Httex1-43Q. When the completion of the oxidation was confirmed, 400 µL of ULP1 enzyme was added to the reaction solution to cleave the SUMO tag, and the reaction was directly injected in RP-HPLC, C4 column using a gradient of 25–35% solvent B (Acetonitrile +0.1 TFA) in solvent A (H_2O+ 0.1% TFA), over 40 min at 15 ml/min. The fractions were analyzed by LC/MS; the ones containing the protein were pooled and lyophilized. The quality of the of protein was analyzed using LC/MS, UPLC, and SDS-PAGE. The final yield of Httex1-43Q oxM8 was 0.76 mg/L of culture.

For the generation of Httex1-43Q oxM8 and Httex1-43Q oxM8/pT3, the oxidation was performed as above, and when completed, the oxidized fusion protein was dialyzed against 4 L of TBS buffer overnight at 4°C and then supplemented with 10x phosphorylation buffer to obtain a final buffer with the following composition and concentration: 50 mM Tris, 25 mM $MgCl_2$, 8 mM EGTA, 4 mM EDTA, 1 mM DTT. Mg-ATP (5 mM) was added to the solution, and the pH was adjusted, and finally, GCK was added at a ratio of 1:30 w/w (kinase:protein) to the Httex1-43 oxM8 (666 µL), and the reaction was incubated at 30°C overnight. Next, the extent of the phosphorylation was verified by LC/MS, and when it was completed, 400 µL of ULP1 was added to the phosphorylation reaction solution to cleave the SUMO tag. Finally, Httex1-43Q-oxM8 was separated from the SUMO tag, and other impurities by HPLC and the fractions containing the protein were pooled and analyzed by LC/MS, UPLC, and SDS-PAGE. The final yield of Httex1-43Q oxM8/pT3 was 0.6 mg/L of culture.

Semi-Synthesis of Httex1-43Q oxM8/AcK6

The semi-synthesis of Httex1-43Q oxM8/AcK6 was performed as previously reported (Chiki et al., 2017). Briefly, 5 mg of Htt-A10C-90-43Q with free N-terminal cysteine was dissolved in neat TFA, and the TFA was dried under nitrogen after 30 min. The dried Htt-A10C-90-43Q was redissolved in 5.0 ml ligation buffer (8 M urea, 0.5 M L-Proline, 30 mM D-Trehalose, 100 mM TCEP and 100 mM MPAA). The pH was adjusted to 7.0 with 10 M NaOH, and the peptide Ac-2-9Nbz AcK6/oxM8 peptide (2.5 mg) was added to the reaction mixture. The native chemical ligation was monitored by LC-ESI-MS until the complete consumption of A10C-90 43Q. Next, the ligation reaction was desalted with hitrap 26/10 desalting column, and fractions containing the Httex1-A10C-43Q oxM8/pT3 were pooled and lyophilized. To recover the native alanine, the protein mixture was disaggregated and dissolved in 100 mM TCEP, 40 mM L-methionine, 20 vol% acetic acid in H_2O. Freshly prepared nickel boride suspension (1.0 ml) was added to the resulting solution, and the resulting suspension was incubated at 37°C. After 2 h, a loss of 32 Da was observed. Insoluble nickel was removed from the reaction mixture via sedimentation (4°C, 4,000 g, 10 min). The supernatant was subjected to RP-HPLC purification with a gradient of 25–55% buffer B (MeCN +0.1% TFA) in buffer A (H_2O+ 0.1% TFA). The fractions containing Ac-Httex1-43Q oxM8/pT3 were analyzed by LC-MS and UPLC, pooled, and lyophilized. The final purity of the protein was checked by LC/MS, UPLC, and SDS-PAGE. The final of the semisynthesis reaction was 27% (1.35 mg of final protein/5 mg of Htt-A10C-90-43Q starting material).

UPLC-Based Sedimentation Assay

To perform aggregation studies, Httex1 was disaggregated as previously reported (Reif et al., 2018) using TFA, and after evaporating the TFA, the protein was filtered through a 100-kDa filter. The final volume was adjusted with PBS to have the desired concentration and incubated at 37°C. To determine the percentage of the remaining soluble monomer during the aggregation, 40 µL was taken at each indicated time point, and insoluble aggregates were removed by centrifugation (4°C, 20,000 g, 30 min). The supernatant was injected into the UPLC. The peak area was measured for each time point, and the changes in the area were used to calculate the fraction of soluble protein compared to t = 0 being 100%. The exponential decay function $y = Ae^{b(x)} + c$ was used to fit the data of different aggregation curves (O'Nuallain et al., 2006b). All the mutant Httex1 aggregated protein curves had an adjusted R^2 ranging from 0.97 to 0.99 (except Httex1-43Q AcK6/oxM8 with $R^2 = 0.78$).

Electron Microscopy

For TEM analysis, 5 µl of aggregation solution was spotted onto a Formvar/carbon-coated 200-mesh glow-discharged copper grid for 1 min. The grid was then washed 3x with water and stained for 3×10 s with 0.7% w/v uranyl formate. Imaging was performed on a Tecnai Spirit BioTWIN electron microscope equipped with a LaB6 gun and a 4 K × 4 K FEI Eagle CCD camera (FEI) and operated at 80 kV.

Atomic Force Microscopy (AFM) Imaging

AFM was performed on freshly cleaved mica discs that are positively functionalized with 1% (3-aminopropyl) triethoxysilane (APTS) in an aqueous solution for 3 min at room temperature. For fibril deposition on the substrates, a 20-µL aliquot of 5.2 µM protein solution was loaded on the

surface of mica discs at defined time-points. Deposition took 3 min and was followed by a gentle drying by nitrogen flow. The mica discs were stored in a desiccator for one day before imaging to avoid prolonged exposure to atmospheric moisture. AFM imaging was performed at room temperature by a Park NX10 operating in a true non-contact mode that was equipped with a super sharp tip (SSS-NCHR) Park System cantilever. The length and height quantifications were performed semi-automatically using XEI software developed by Park Systems Corp by taking advantage of the grain detection method and thresholding (masking) the background values. The quantified data was plotted and analyzed in GraphPad Prism 8.

Circular Dichroism

mHttex1 proteins (100 µL) during the aggregation or 120 µL of the Nt17 peptides (at 60 µM in PBS) were removed from the mixture and analyzed using a Jasco J-815 CD spectrometer and a 1.0 mm quartz cuvette. Ellipticity was measured from 195 to 250 nm at 25°C.

Origin was used to smooth the data through Savitzky-Golay fitting (8 points window). The mean residue molar ellipticity (θMRE) was plotted, and BeStSel (Micsonai et al., 2018) was used for fitting and calculating the helicity. Average helicity and standard deviation were calculated from 3 repeats.

Molecular Dynamics Simulations

The starting structure for molecular dynamics simulations was a fully extended form of Htt 19 peptide. The length of this extended conformation was 68 Å, and the buffer distance between each side of the peptide and the periodic boundary was 10 Å, resulting in a water box with a width of 88 Å. The HTT 19 peptide was solvated using explicit CHARMM36m (Huang et al., 2017) modified TIP3P water model (Jorgensen et al., 1983) and 20 mM K^+ and Cl^- ions. The simulation was conducted with GROMACS (Abraham et al., 2015), and CHARMM-GUI was used to generate inputs. A CHARMM36m forcefield was used with modified residues for phosphorylated threonine and acetylated lysine residues (Grauffel et al., 2010). The phosphate group of phosphorylated threonine was in a dianionic form. The oxidized methionine used in this simulation was an R diastereomer, and its parameters were obtained from previously published papers (Lockhart et al., 2020). The system was minimized with the steepest descent and equilibrated to 1 atm and 303.15 K with constraints on the backbone. The nonbonded interaction cut-off was 12 Å. The time step was 1 fs for equilibration and 2 fs for production. The Nosé–Hoover temperature (Nosé, 1984) coupling method was used to maintain temperature, and the isotropic Parrinello–Rahman (Parrinello and Rahman, 1981; Nosé and Klein, 1983) method was used for pressure coupling. LINCS algorithm (Hess et al., 1997) was used for H-bond. The production was run for 13 µs for each system.

Data Analyses

To perform dPCA (Mu et al., 2005), the dihedral angles were calculated for the Nt19 backbone, and a transformation from the space of dihedral angles (ϕ_n, ψ_n) to metric coordinate space was done by taking trigonometric function (cos ϕ_n, sin ϕ_n, cos ψ_n, sin ψ_n). Then, the transformed dihedral data of each post-translationally modified Nt19 was combined, and a PCA analysis was done. A free energy map was generated with the first two components. The secondary structure was obtained with the STRIDE (Frishman and Argos, 1995) algorithm implemented in VMD (Humphrey et al., 1996). The overall helicity for each peptide was calculated based on the sum of the total number of residues that are in *a*-helix or 3_{10} helical conformations at every frame and then divide by the total number of frames.

We analyzed and compared the effects of each post-translational modification of Nt1-19 on the dynamics using a statistical analysis of the distribution of pairwise distances between residues. For each MD of Nt1-19, we reduce the ensemble of conformations to a statistical description of the distances between all pairs of Cα in Nt1-19. For each pair of residues within Nt1–19, we compute the distribution of the Euclidean distances between the two Cα with a bin resolution of 0.5 Å. We, therefore, describe every MD with a corresponding histogram of distances summarizing the dynamics of the structure. The differences in residue-residue dynamics between MDs can be quantified using the Kullback-Leibler divergence (Kullback and Leibler, 1951) from a reference histogram of distances.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

HL and AC conceived the project, HL and MDP supervised all the work, HL, MP, AC, ZZ, and LA designed all the experiments, AC, ZZ, KR, and IR performed the experiments, ZZ, LA, and LK analyzed molecular simulation results, AC and ZZ prepared the first draft of the paper and all authors contributed to the writing and preparation of the final version.

FUNDING

HAL acknowledges CHDI for financial support. MDP acknowledges Swiss National Science Foundation for financial support (grant number 205321_192371) and the Swiss National Supercomputing Center (CSCS) for computational resource for molecular simulations (project ID: 152).

ACKNOWLEDGMENTS

We would like to thank Senthil Thangaraj, Ahmed Sadek, Pedro Santana Magalhães, Sylvain Träger for valuable discussion for article preparation. We would like to thank all members of

Laboratory of Molecular and Chemical Biology of Neurodegeneration and Laboratory for Biomolecular Modeling for their feedback and input on the project. The content of the manuscript has been posted on bioRxiv with doi: https://doi.org10.1101/2021.02.21.432155.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.686086/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

Aiken, C. T., Steffan, J. S., Guerrero, C. M., Khashwji, H., Lukacsovich, T., Simmons, D., et al. (2009). Phosphorylation of Threonine 3. *J. Biol. Chem.* 284 (43), 29427–29436. doi:10.1074/jbc.m109.013193

Arnesen, T., Starheim, K. K., Van Damme, P., Evjenth, R., Dinh, H., Betts, M. J., et al. (2010). The Chaperone-like Protein HYPK Acts Together with NatA in Cotranslational N-Terminal Acetylation and Prevention of Huntingtin Aggregation. *Mol. Cel Biol.* 30 (8), 1898–1909. doi:10.1128/mcb.01199-09

Barbaro, B. A., Lukacsovich, T., Agrawal, N., Burke, J., Bornemann, D. J., Purcell, J. M., et al. (2015). Comparative Study of Naturally Occurring Huntingtin Fragments in Drosophila Points to Exon 1 as the Most Pathogenic Species in Huntington's Disease. *Hum. Mol. Genet.* 24 (4), 913–925. doi:10.1093/hmg/ddu504

Bergström, A.-L., Chabry, J., Bastholm, L., and Heegaard, P. M. H. (2007). Oxidation Reduces the Fibrillation but Not the Neurotoxicity of the Prion Peptide PrP106-126. *Biochim. Biophys. Acta (Bba) - Proteins Proteomics* 1774 (9), 1118–1127. doi:10.1016/j.bbapap.2007.06.016

Cariulo, C., Azzollini, L., Verani, M., Martufi, P., Boggio, R., Chiki, A., et al. (2017). Phosphorylation of Huntingtin at Residue T3 Is Decreased in Huntington's Disease and Modulates Mutant Huntingtin Protein Conformation. *Proc. Natl. Acad. Sci. USA.* 114 (50), E10809–E10818. doi:10.1073/pnas.1705372114

Ceccon, A., Tugarinov, V., and Clore, G. M. (2018). TiO2 Nanoparticles Catalyze Oxidation of Huntingtin Exon 1-derived Peptides Impeding Aggregation: A Quantitative NMR Study of Binding and Kinetics. *J. Am. Chem. Soc.* 141 (1), 94–97. doi:10.1021/jacs.8b11441

Chen, C.-M., Wu, Y.-R., Cheng, M.-L., Liu, J.-L., Lee, Y.-M., Lee, P.-W., et al. (2007). Increased Oxidative Damage and Mitochondrial Abnormalities in the Peripheral Blood of Huntington's Disease Patients. *Biochem. biophysical Res. Commun.* 359 (2), 335–340. doi:10.1016/j.bbrc.2007.05.093

Chiki, A., DeGuire, S. M., Ruggeri, F. S., Sanfelice, D., Ansaloni, A., Wang, Z-M., et al. (2017). Mutant Exon1 Huntingtin Aggregation Is Regulated by T3 Phosphorylation-Induced Structural Changes and Crosstalk between T3 Phosphorylation and Acetylation at K6. *Angew. Chem. Int. Ed. Engl.* 56 (19), 5202–5207. doi:10.1002/anie.201611750

Chiki, A., Ricci, J., Hegde, R., Abriata, L. A., Reif, A., Boudeffa, D., et al. (2020). Site-specific Phosphorylation of Huntingtin Exon 1 Recombinant Proteins Enabled by the Discovery of Novel Kinases. *Chembiochem* 22 (1), 217–231. doi:10.1002/cbic.202000508

DeGuire, S. M., Ruggeri, F. S., Fares, M.-B., Chiki, A., Cendrowska, U., Dietler, G., et al. (2018). N-terminal Huntingtin (Htt) Phosphorylation Is a Molecular Switch Regulating Htt Aggregation, Helical Conformation, Internalization, and Nuclear Targeting. *J. Biol. Chem.* 293 (48), 18540–18558. doi:10.1074/jbc.ra118.004621

DiFiglia, M., Sapp, E., Chase, K. O., Davies, S. W., Bates, G. P., Vonsattel, J. P., et al. (1997). Aggregation of Huntingtin in Neuronal Intranuclear Inclusions and Dystrophic Neurites in Brain. *Science* 277 (5334), 1990–1993. doi:10.1126/science.277.5334.1990

DiGiovanni, L. F., Mocle, A. J., Xia, J., and Truant, R. (2016). Huntingtin N17 Domain Is a Reactive Oxygen Species Sensor Regulating Huntingtin Phosphorylation and Localization. *Hum. Mol. Genet.* 25 (18), 3937–3945. doi:10.1093/hmg/ddw234

Friedemann, M., Helk, E., Tiiman, A., Zovo, K., Palumaa, P., and Tõugu, V. (2015). Effect of Methionine-35 Oxidation on the Aggregation of Amyloid-β Peptide. *Biochem. Biophys. Rep.* 3, 94–99. doi:10.1016/j.bbrep.2015.07.017

Frishman, D., and Argos, P. (1995). Knowledge-based Protein Secondary Structure Assignment. *Proteins* 23 (4), 566–579. doi:10.1002/prot.340230412

Froger, A., and Hall, J. E. (2007). Transformation of Plasmid DNA into *E. coli* Using the Heat Shock Method. *J. Vis. Exp.* (6), 253. doi:10.3791/253

Gauthier, L. R., Charrin, B. C., Borrell-Pagès, M., Dompierre, J. P., Rangone, H., Cordelières, F. P., et al. (2004). Huntingtin Controls Neurotrophic Support and Survival of Neurons by Enhancing BDNF Vesicular Transport along Microtubules. *Cell* 118 (1), 127–138. doi:10.1016/j.cell.2004.06.018

Grauffel, C., Stote, R. H., and Dejaegere, A. (2010). Force Field Parameters for the Simulation of Modified Histone Tails. *J. Comput. Chem.* 31 (13), 2434–2451. doi:10.1002/jcc.21536

Gu, M., and Viles, J. H. (2016). Methionine Oxidation Reduces Lag-Times for Amyloid-B(1-40) Fiber Formation but Generates Highly Fragmented Fibers. *Biochim. Biophys. Acta (Bba) - Proteins Proteomics* 1864 (9), 1260–1269. doi:10.1016/j.bbapap.2016.04.009

Gu, X., Greiner, E. R., Mishra, R., Kodali, R., Osmand, A., Finkbeiner, S., et al. (2009). Serines 13 and 16 Are Critical Determinants of Full-Length Human Mutant Huntingtin Induced Disease Pathogenesis in HD Mice. *Neuron* 64 (6), 828–840. doi:10.1016/j.neuron.2009.11.020

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., et al. (1983). A Polymorphic DNA Marker Genetically Linked to Huntington's Disease. *Nature* 306 (5940), 234–238. doi:10.1038/306234a0

Hegde, R. N., Chiki, A., Petricca, P., Martufi, P., Arbez, N., Mouchiroud, L., et al. (2020). TBK1 Phosphorylates Mutant Huntingtin and Suppresses its Aggregation and Toxicity in Huntington's Disease Models. *EMBO J.* 39 (17), e104671. doi:10.15252/embj.2020104671

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 (12), 1463–1472. doi:10.1002/(sici)1096-987x(199709)18:12<1463:aid-jcc4>3.0.co;2-h

Hodgson, J. G., Agopyan, N., Gutekunst, C.-A., Leavitt, B. R., LePiane, F., Singaraja, R., et al. (1999). A YAC Mouse Model for Huntington's Disease with Full-Length Mutant Huntingtin, Cytoplasmic Toxicity, and Selective Striatal Neurodegeneration. *Neuron* 23 (1), 181–192. doi:10.1016/s0896-6273(00)80764-3

Hoshi, T., and Heinemann, S. (2001). Regulation of Cell Function by Methionine Oxidation and Reduction. *J. Physiol.* 531 (Pt 1), 1–11. doi:10.1111/j.1469-7793.2001.0001j.x

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., et al. (2017). CHARMM36m: an Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* 14 (1), 71–73. doi:10.1038/nmeth.4067

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual Molecular Dynamics. *J. Mol. graphics* 14 (1), 33–38. doi:10.1016/0263-7855(96)00018-5

Huntington, G. (1872). On Chorea, *Medical and Surgical Reporter*. PA, United States. (Accessed April 13, 1872.)

Jayaraman, M., Thakur, A. K., Kar, K., Kodali, R., and Wetzel, R. (2011). Assays for Studying Nucleated Aggregation of Polyglutamine Proteins. *Methods* 53 (3), 246–254. doi:10.1016/j.ymeth.2011.01.001

Jeong, H., Then, F., Melia, T. J., Mazzulli, J. R., Cui, L., Savas, J. N., et al. (2009). Acetylation Targets Mutant Huntingtin to Autophagosomes for Degradation. *Cell* 137 (1), 60–72. doi:10.1016/j.cell.2009.03.018

Ji, J. A., Zhang, B., Cheng, W., and Wang, Y. J. (2009). Methionine, Tryptophan, and Histidine Oxidation in a Model Protein, PTH: Mechanisms and Stabilization. *J. Pharm. Sci.* 98 (12), 4485–4500. doi:10.1002/jps.21746

Johansson, A.-S., Bergquist, J., Volbracht, C., Päiviö, A., Leist, M., Lannfelt, L., et al. (2007). Attenuated Amyloid-β Aggregation and Neurotoxicity Owing to

Methionine Oxidation. *Neuroreport* 18 (6), 559–563. doi:10.1097/wnr.0b013e3280b07c21

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 (2), 926–935. doi:10.1063/1.445869

Keck, R. G. (1996). The Use of T-Butyl Hydroperoxide as a Probe for Methionine Oxidation in Proteins. *Anal. Biochem.* 236 (1), 56–62. doi:10.1006/abio.1996.0131

Kremer, B., Goldberg, P., Andrew, S. E., Theilmann, J., Telenius, H., Zeisler, J., et al. (1994). A Worldwide Study of the Huntington's Disease Mutation: The Sensitivity and Specificity of Measuring CAG Repeats. *N. Engl. J. Med.* 330 (20), 1401–1406. doi:10.1056/nejm199405193302001

Kullback, S., and Leibler, R. A. (1951). On Information and Sufficiency. *Ann. Math. Statist.* 22 (1), 79–86. doi:10.1214/aoms/1177729694

Kumar, A., and Ratan, R. R. (2016). Oxidative Stress and Huntington's Disease: The Good, the Bad, and the Ugly. *Jhd* 5, 217–237. doi:10.3233/jhd-160205

Landles, C., Sathasivam, K., Weiss, A., Woodman, B., Moffitt, H., Finkbeiner, S., et al. (2010). Proteolysis of Mutant Huntingtin Produces an Exon 1 Fragment that Accumulates as an Aggregated Protein in Neuronal Nuclei in Huntington Disease. *J. Biol. Chem.* 285 (12), 8808–8823. doi:10.1074/jbc.m109.075028

Li, S., Schöneich, C., and Borchardt, R. T. (1995). Chemical Instability of Protein Pharmaceuticals: Mechanisms of Oxidation and Strategies for Stabilization. *Biotechnol. Bioeng.* 48 (5), 490–500. doi:10.1002/bit.260480511

Lockhart, C., Smith, A. K., and Klimov, D. K. (2020). Methionine Oxidation Alters the Mechanism of Aβ Interaction with DMPC Bilayers. *Biophysical J.* 118 (3), 61a. doi:10.1016/j.bpj.2019.11.508

Lunkes, A., Lindenberg, K. S., Ben-Haïem, L., Weber, C., Devys, D., Landwehrmeyer, G. B., et al. (2002). Proteases Acting on Mutant Huntingtin Generate Cleaved Products that Differentially Build up Cytoplasmic and Nuclear Inclusions. *Mol. Cel* 10 (2), 259–269. doi:10.1016/s1097-2765(02)00602-0

Luo, Q., Joubert, M. K., Stevenson, R., Ketchem, R. R., Narhi, L. O., and Wypych, J. (2011). Chemical Modifications in Therapeutic Protein Aggregates Generated under Different Stress Conditions. *J. Biol. Chem.* 286 (28), 25134–25144. doi:10.1074/jbc.m110.160440

MacDonald, M. (1993). A Novel Gene Containing a Trinucleotide Repeat that Is Expanded and Unstable on Huntington's Disease Chromosomes. *Cell* 72 (6), 971–983. doi:10.1016/0092-8674(93)90585-e

Mangiarini, L., Sathasivam, K., Seller, M., Cozens, B., Harper, A., Hetherington, C., et al. (1996). Exon 1 of the HD Gene with an Expanded CAG Repeat Is Sufficient to Cause a Progressive Neurological Phenotype in Transgenic Mice. *Cell* 87 (3), 493–506. doi:10.1016/s0092-8674(00)81369-0

Marder, K., Zhao, H., Myers, R. H., Cudkowicz, M., Kayson, E., Kieburtz, K., et al. (2000). Rate of Functional Decline in Huntington's Disease. *Neurology* 54 (2), 452. doi:10.1212/wnl.54.2.452

Martindale, D., Hackam, A., Wieczorek, A., Ellerby, L., Wellington, C., McCutcheon, K., et al. (1998). Length of Huntingtin and its Polyglutamine Tract Influences Localization and Frequency of Intracellular Aggregates. *Nat. Genet.* 18 (2), 150–154. doi:10.1038/ng0298-150

Micsonai, A., Wien, F., Bulyáki, É., Kun, J., Moussong, É., Lee, Y.-H., et al. (2018). BeStSel: a Web Server for Accurate Protein Secondary Structure Prediction and Fold Recognition from the Circular Dichroism Spectra. *Nucleic Acids Res.* 46 (W1), W315–W322. doi:10.1093/nar/gky497

Mitomi, Y., Nomura, T., Kurosawa, M., Nukina, N., and Furukawa, Y. (2012). Post-aggregation Oxidation of Mutant Huntingtin Controls the Interactions between Aggregates. *J. Biol. Chem.* 287 (41), 34764–34775. doi:10.1074/jbc.m112.387035

Mu, Y., Nguyen, P. H., and Stock, G. (2005). Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis. *Proteins* 58 (1), 45–52. doi:10.1002/prot.20310

Nosé, S. (1984). A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* 52 (2), 255–268. doi:10.1080/00268978400101201

Nosé, S., and Klein, M. L. (1983). Constant Pressure Molecular Dynamics for Molecular Systems. *Mol. Phys.* 50 (5), 1055–1076. doi:10.1080/00268978300102851

O'Nuallain, B., Thakur, A. K., Williams, A. D., Bhattacharyya, A. M., Chen, S., Thiagarajan, G., et al. , 2006, Kinetics and Thermodynamics of Amyloid Assembly Using a High-Performance Liquid Chromatography-Based Sedimentation Assay, *Methods Enzymol.* 413: p. 34–74.doi:10.1016/s0076-6879(06)13003-7

O'Nuallain, B., Thakur, A. K., Williams, A. D., Bhattacharyya, A. M., Chen, S., Thiagarajan, G., et al. (2006). Kinetics and Thermodynamics of Amyloid Assembly Using a High-Performance Liquid Chromatography-Based Sedimentation Assay. *Amyloid, Prions, and Other Protein Aggregates, Pt C* 413, 34–74. doi:10.1016/s0076-6879(06)13003-7

O'Rourke, J. G., Gareau, J. R., Ochaba, J., Song, W., Raskó, T., Reverter, D., et al. (2013). SUMO-2 and PIAS1 Modulate Insoluble Mutant Huntingtin Protein Accumulation. *Cel Rep.* 4 (2), 362–375. doi:10.1016/j.celrep.2013.06.034

Parrinello, M., and Rahman, A. (1981). Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* 52 (12), 7182–7190. doi:10.1063/1.328693

Pastore, A., Yalinca, H., Caroline Gehin, C. J., Oleinikovas, V., Lashuel, H. A., and Gervasio, F. L. (2019). The Role of post-translational Modifications on the Energy Landscape of Huntingtin N-Terminus. *Front. Mol. biosciences* 6, 95.

Paulsen, J. S., Ready, R. E., Hamilton, J. M., Mega, M. S., and Cummings, J. L. (2001). Neuropsychiatric Aspects of Huntington's Disease. *J. Neurol. Neurosurg. Psychiatry* 71 (3), 310–314. doi:10.1136/jnnp.71.3.310

Polidori, M. C., Mecocci, P., Browne, S. E., Senin, U., and Beal, M. F. (1999). Oxidative Damage to Mitochondrial DNA in Huntington's Disease Parietal Cortex. *Neurosci. Lett.* 272 (1), 53–56. doi:10.1016/s0304-3940(99)00578-9

Reif, A., Chiki, A., Ricci, J., and Lashuel, H. A. (2018). Generation of Native, Untagged Huntingtin Exon1 Monomer and Fibrils Using a SUMO Fusion Strategy. *J. Vis. Exp.* (136), 57506. doi:10.3791/57506

Reiner, A., Albin, R. L., Anderson, K. D., D'Amato, C. J., Penney, J. B., and Young, A. B. (1988). Differential Loss of Striatal Projection Neurons in Huntington Disease. *Proc. Natl. Acad. Sci.* 85 (15), 5733–5737. doi:10.1073/pnas.85.15.5733

Rosas, H. D., Koroshetz, W. J., Chen, Y. I., Skeuse, C., Vangel, M., Cudkowicz, M. E., et al. (2003). Evidence for More Widespread Cerebral Pathology in Early HD: an MRI-Based Morphometric Analysis. *Neurology* 60 (10), 1615–1620. doi:10.1212/01.wnl.0000065888.88988.6e

Ross, C. A., and Tabrizi, S. J. (2011). Huntington's Disease: from Molecular Pathogenesis to Clinical Treatment. *Lancet Neurol.* 10 (1), 83–98. doi:10.1016/s1474-4422(10)70245-3

Sathasivam, K., Neueder, A., Gipson, T. A., Landles, C., Benjamin, A. C., Bondulich, M. K., et al. (2013). Aberrant Splicing of HTT Generates the Pathogenic Exon 1 Protein in Huntington Disease. *Proc. Natl. Acad. Sci.* 110 (6), 2366–2370. doi:10.1073/pnas.1221891110

Scherzinger, E., Sittler, A., Schweiger, K., Heiser, V., Lurz, R., Hasenbank, R., et al. (1999). Self-assembly of Polyglutamine-Containing Huntingtin Fragments into Amyloid-like Fibrils: Implications for Huntington's Disease Pathology. *Proc. Natl. Acad. Sci.* 96 (8), 4604–4609. doi:10.1073/pnas.96.8.4604

Sedighi, F., Adegbuyiro, A., and Legleiter, J. (2020). SUMOylation Prevents Huntingtin Fibrillization and Localization onto Lipid Membranes. *ACS Chem. Neurosci.* 11 (3), 328–343. doi:10.1021/acschemneuro.9b00509

Snell, R. G., MacMillan, J. C., Cheadle, J. P., Fenton, I., Lazarou, L. P., Davies, P., et al. (1993). Relationship between Trinucleotide Repeat Expansion and Phenotypic Variation in Huntington's Disease. *Nat. Genet.* 4 (4), 393–397. doi:10.1038/ng0893-393

Son, S., Bowie, L. E., Maiuri, T., Hung, C. L. K., Desmond, C. R., Xia, J., et al. (2019). High-mobility Group Box 1 Links Sensing of Reactive Oxygen Species by Huntingtin to its Nuclear Entry. *J. Biol. Chem.* 294 (6), 1915–1923. doi:10.1074/jbc.ra117.001440

Steffan, J. S., Agrawal, N., Pallos, J., Rockabrand, E., Trotman, L. C., Slepko, N., et al. (2004). SUMO Modification of Huntingtin and Huntington's Disease Pathology. *Science* 304 (5667), 100–104. doi:10.1126/science.1092194

Thompson, L. M., Aiken, C. T., Kaltenbach, L. S., Agrawal, N., Illes, K., Khoshnan, A., et al. (2009). IKK Phosphorylates Huntingtin and Targets it for Degradation by the Proteasome and Lysosome. *J. Cel Biol.* 187 (7), 1083–1099. doi:10.1083/jcb.200909067

Valley, C. C., Cembran, A., Perlmutter, J. D., Lewis, A. K., Labello, N. P., Gao, J., et al. (2012). The Methionine-Aromatic Motif Plays a Unique Role in Stabilizing Protein Structure. *J. Biol. Chem.* 287 (42), 34979–34991. doi:10.1074/jbc.m112.374504

Vieweg, S., Ansaloni, A., Wang, Z.-M., Warner, J. B., and Lashuel, H. A. (2016). An Intein-Based Strategy for the Production of Tag-free Huntingtin Exon 1 Proteins Enables New Insights into the Polyglutamine Dependence of

Httex1 Aggregation and Fibril Formation. *J. Biol. Chem.* 291 (23), 12074–12086. doi:10.1074/jbc.m116.713982

Warby, S. C., Chan, E. Y., Metzler, M., Gan, L., Singaraja, R. R., Crocker, S. F., et al. (2005). Huntingtin Phosphorylation on Serine 421 Is Significantly Reduced in the Striatum and by Polyglutamine Expansion *In Vivo*. *Hum. Mol. Genet.* 14 (11), 1569–1577. doi:10.1093/hmg/ddi165

Wasylaschuk, W. R., Harmon, P. A., Wagner, G., Harman, A. B., Templeton, A. C., Xu, H., et al. (2007). Evaluation of Hydroperoxides in Common Pharmaceutical Excipients. *J. Pharm. Sci.* 96 (1), 106–116. doi:10.1002/jps.20726

Yalinca, H., Caroline Gehin, C. J., Oleinikovasas, V., Lashuel, H. A., Luigi Gervasio, F., and Pastore, A. (2019). The Role of Post-translational Modifications on the Energy Landscape of Huntingtin N-Terminus. *Front. Mol. Biosci.* 6, 95. doi:10.3389/fmolb.2019.00095

Yanai, A., Huang, K., Kang, R., Singaraja, R. R., Arstikaitis, P., Gan, L., et al. (2006). Palmitoylation of Huntingtin by HIP14is Essential for its Trafficking and Function. *Nat. Neurosci.* 9 (6), 824–831. doi:10.1038/nn1702

Zhou, W., Long, C., Reaney, S. H., Di Monte, D. A., Fink, A. L., and Uversky, V. N. (2010). Methionine Oxidation Stabilizes Non-toxic Oligomers of α-synuclein through Strengthening the Auto-Inhibitory Intra-molecular Long-Range Interactions. *Biochim. Biophys. Acta (Bba) - Mol. Basis Dis.* 1802 (3), 322–330. doi:10.1016/j.bbadis.2009.12.004

Conflict of Interest: HL was the Founder and Chief Scientific Office of ND BioSciences SA.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Chiki, Zhang, Rajasekhar, Abriata, Rostami, Krapp, Boudeffa, Dal Peraro and Lashuel. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 09 August 2021
doi: 10.3389/fmolb.2021.633181

Exploring Covalent Docking Mechanisms of Boron-Based Inhibitors to Class A, C and D β-Lactamases Using Time-dependent Hybrid QM/MM Simulations

*Łukasz Charzewski, Krystiana A. Krzyśko and Bogdan Lesyng**

Department of Biophysics, Faculty of Physics, University of Warsaw, Warsaw, Poland

Recently, molecular covalent docking has been extensively developed to design new classes of inhibitors that form chemical bonds with their biological targets. This strategy for the design of such inhibitors, in particular boron-based inhibitors, holds great promise for the vast family of β-lactamases produced, *inter alia*, by Gram-negative antibiotic-resistant bacteria. However, the description of covalent docking processes requires a quantum-mechanical approach, and so far, only a few studies of this type have been presented. This study accurately describes the covalent docking process between two model inhibitors - representing two large families of inhibitors based on boronic-acid and bicyclic boronate scaffolds, and three β-lactamases which belong to the A, C, and D classes. Molecular fragments containing boron can be converted from a neutral, trigonal, planar state with sp^2 hybridization to the anionic, tetrahedral sp^3 state in a process sometimes referred to as *morphing*. This study applies multi-scale modeling methods, in particular, the hybrid QM/MM approach which has predictive power reaching well beyond conventional molecular modeling. Time-dependent QM/MM simulations indicated several structural changes and geometric preferences, ultimately leading to covalent docking processes. With current computing technologies, this approach is not computationally expensive, can be used in standard molecular modeling and molecular design works, and can effectively support experimental research which should allow for a detailed understanding of complex processes important to molecular medicine. In particular, it can support the rational design of covalent boron-based inhibitors for β-lactamases as well as for many other enzyme systems of clinical relevance, including SARS-CoV-2 proteins.

Keywords: NAMD, covalent docking, boron-based inhibitors, 3-nitrophenyl boronic acid, bicyclic boronate, antibiotic resistance, β-lactamases, time-dependent QM/MM

OPEN ACCESS

Edited by:
Peter J. Bond,
*Bioinformatics Institute (A*STAR), Singapore*

Reviewed by:
Roland G. Huber,
*Bioinformatics Institute (A*STAR), Singapore*
Hedieh Torabifard,
The University of Texas at Dallas, United States

****Correspondence:***
Bogdan Lesyng
Bogdan.Lesyng@fuw.edu.pl

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *24 November 2020*
Accepted: *19 July 2021*
Published: *09 August 2021*

Citation:
Charzewski Ł, Krzyśko KA and Lesyng B (2021) Exploring Covalent Docking Mechanisms of Boron-Based Inhibitors to Class A, C and D β-Lactamases Using Time-dependent Hybrid QM/MM Simulations. Front. Mol. Biosci. 8:633181. doi: 10.3389/fmolb.2021.633181

INTRODUCTION

Gram-negative bacteria can cause serious infections in immunocompromised patients, including urinary tract infections, pneumonia, hepatitis, sepsis, soft tissue infections, and peritonitis. It is well-known fact that these bacteria are resistant to such antibiotics as *Penicillin, Cephalosporins, Monobactams,* or *Carbapenems*, because of the ability to breakdown the β-lactam antibiotics by

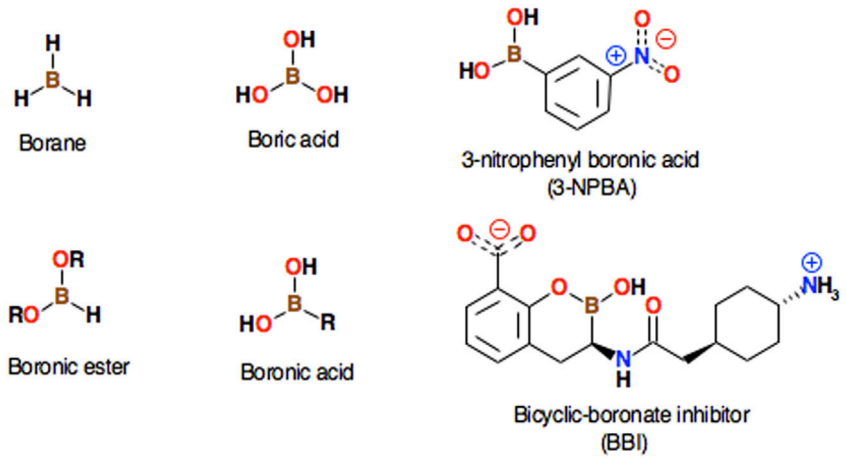


FIGURE 1 | Several well-known simple borane compounds are shown along with their systematic names. The structures of 3-nitrophenylboronic acid (3-NPBA) and the bicyclic boronate inhibitor ((3R)-3-[[2-[4-(amino)cyclohexyl]acetyl]amino]-2-hydroxy-3,4-dihydro-1,2-benzoxaborinine-8-carboxylic acid, BBI), which are the subject of this study, are also presented.

β-lactamases. The reaction mechanism of class A β-lactamases consists of acylation of an active site serine by the antibiotic molecule, followed by deacylation and release of the cleaved compound. Acylation is related to the formation of the tetrahedral intermediate. This reaction mechanism was described in detail using QM/MM and high-level *ab initio* calculations to model the rate-determining step for the formation of a tetrahedral intermediate in acylation (Hermann et al., 2009).

In recent years, significant progress has been made in the design and synthesis of a new class of antibiotics resistant to the β-lactamase mechanisms. In particular, boron-based inhibitors have proven to be effective inhibitors of β-lactamases due to chemical blockade of the tetrahedral transition state forced by these enzymes, and typically due to low toxicity. Boronic acid derivatives, as well as cyclic boronates, have been lately tested in their ability for inhibiting β-lactamases produced, among others, by Gram-negative pathogens *Escherichia coli* and *Klebsiella pneumoniae*. An overview of boron-based small molecules in disease detection and treatment is presented (Theraja, et al., 2017). The combination of theoretical and crystallographic approaches has provided important insight into the molecular stabilization of boron-based compounds in various target proteins, such as proteasomes, tyrosine kinases, histone deacetylases, GPCRs, glutamate racemases, amino acid transporters, autotaxin, and in particular β-lactamases (Calvopiña, et al., 2017; Bello, 2018; Cahill et al., 2019; Song et al., 2021). Based on crystallographic and NMR data, mechanism of proton transfer in class A β-lactamase catalysis and inhibition by *Avibactam* were described (Pemberton et al., 2020).

A number of boron-based chemicals are known as therapeutic agents or drugs, such as *Ixazomib, Tavaborole*, *Crisaborole, Vaborbactam* or *Taniborbactam* (VNRX-5133) Ban et al. (2015), Thareja et al. (2017), Krajnc et al. (2019a), Krajnc et al. (2019b), Lang et al. (2020), Lence and Gonzales-Bello (2021), for specifications see e.g. Drug Central 2021, https://drugcentral.org/. Two authors of this study (Ł.Ch. and K.A.K) participated in integrated studies involving the modeling, synthesis, and biological activity analysis of a large series of boron-based inhibitors of KPC/AmpC β-lactamases (Durka et al., 2019). One should also mention recent work on the evaluation of *Bortezomib* and other boron-containing compounds, as inhibitors of SARS-CoV-2 main protease, e.g. Vega-Valdez et al., 2020). A very comprehensive overview of the mechanisms of β-lactamase inhibition by boron-based inhibitors, including cyclic boronates, and their applications in biomedicine is reported (Brem et al., 2016; Cahill et al., 2017; Cahill et al., 2019; Tooke et al., 2019; Tooke et al., 2020; Song et al., 2021).

What is so special about the boron atom? The boron atom constitutes the basis of a huge number of chemical derivatives with very diverse chemical properties; some common borate derivatives including their systematic names are presented (**Figure 1**). Metalloid boron has many similarities to neighboring carbon and silicon. All three elements form covalent compounds. However, boron has one distinct difference as its outer $2s^2 2p^1$ electronic structure contains one less valence electron than it has valence orbitals. Boron atom is formally trivalent, but due to its empty p-orbital, it can act as a strong acid - the so-called Lewis acid. Its acidic properties will be important for the topic of this work. Lewis acid accepts electrons from electron pair donors, such as H_2O, HO^-, NH_3, or CH_3^-. The lone electron pair of the nucleophile interacts with the vacant p-orbital of boron that facilitates the chemical reaction. The interaction of boron with molecular groups having lone electron pairs leads to a tetrahedral transition state, with a valence bond configuration similar to one found in carbon (**Figure 2**). This way, boronic molecular fragments can easily be converted from the neutral, trigonal planar sp2 hybridization state to the anionic, tetrahedral sp3 state in a process sometimes called *morphing*, which may (or may not) lead to the formation of a chemical bond between boron and the electron donor moiety. It

FIGURE 2 | A simplified diagram of the Lewis acid-base model of boronic acid in water.

should also be mentioned that in some cases the boron derivatives may also act as Brønsted-Lowry acids. The term *Lewis acid* is more general than *Brønsted-Lowry acid*. In the Brønsted-Lowry model, an acid is characterized as a proton donor and a base as a proton acceptor. The Lewis acid-base model also describes reactions without proton transfer, but where single electron pairs can form new bonds. 3-nitrophenylboronic acid (3-NPBA) and bicyclic boronate derivatives are known covalent inhibitors of β-lactamases (Ke et al., 2012; Brem et al., 2016). 3-NPBA and a bicyclic boronate inhibitor (BBI) being the subjects of this study are shown (**Figure 1**). 3-NPBA can form a reversible chemical bond with a serine hydroxyl group of the target protein, providing additional covalent-stabilization compared to the initial nonbonded-type interaction, which is the key goal of the discussed activities. The process of formation of such a bond is, however, quite sensitive to the local electric field, depending on the relative orientations of the interacting molecular fragments. In addition, in the case of enzymes, the entire process runs through several transition states that have not been clearly defined so far. Therefore, intentional covalent docking may require a few non-obvious adjustments in order to yield reliable results. Until now, most of the covalent docking work has been supported by heuristic procedures Krajnc et al. (2019a), Krajnc et al. (2019b), and only a few studies of this type of process have been described based on approximate quantum-mechanical methods e.g. (Sgrignani et al., 2015; Krivitskaya and Khrenova, 2021; Lence and Gonzalez-Bello, 2021).

An important problem of boronic acids is their oxidative instability. Phenylboronic acid and its borate esters are oxidized by reactive oxygen species. Quite recently it has been found Graham et al. (2021) that reducing the electron density on the boron increases its oxidative stability. The oxidation of a boronic acid can be slowed down by the presence of a pendant carboxyl group, which is the ligand to the boron atom. The ensuing derivative, oxaborolone, is 10^4-fold more resistant to oxidation.

Organoboron compounds participate in free radical reactions (Renaud et al., 2007; Kumar et al., 2020). Recent studies describe C-B bond formation mechanisms *via* radical intermediates (Friese and Studer, 2019). Photoactivation processes of boron-species, as well as applications to the creation of C–H, C–C, C–O, B–C, and B–S bond, are discussed in review (Duret et al., 2015). It should be emphasized, however, that as far as the mechanisms of interaction of covalent inhibitors of boron with β-lactamases are concerned, so far there are no experimental or theoretical data that would indicate the radical mechanisms of the formation of chemical bonds of the boron atom with hydroxyl groups of serine residues.

The aim of this work is to accurately describe the covalent docking process between two model inhibitors - representing two large families of inhibitors based on boronic acid and bicyclic boronate scaffolds, and three β lactamases belonging to the A, C, and D enzyme classes, namely KPC-2 β-lactamase (class A), GC1 β-lactamase (class C), and OXA-24 β-lactamase (class D). Class B is not included in this study as it is an ortholog different structurally and mechanistically. As class B catalysis is based on Zn^{2+} ions instead of serine, these metallo-β-lactamases would require a separate study. For a more detailed β-lactamases classification description and properties see e.g. (Bush, 2018; Keshri et al., 2018; Philippon et al., 2016). As mentioned, the first inhibitor is 3-NPBA, and the second one is the inhibitor based on the bicyclic boronate scaffold, named in this study as BBI, (**Figure 1**). Systematic name of this compound is (3R)-3-[[2-[4-(amino)cyclohexyl]acetyl]amino]-2-hydroxy-3,4-dihydro-1,2-benzoxaborinine-8-carboxylic acid, and it is a close analog of the known *Taniborbactam* antibiotic (Krajnc et al., 2019a; Krajnc et al., 2019b; Liu et al., 2020).

Although the crystallographic structures (6TD1, 6YEN and 6RTN) of *Taniborbactam* complexes with β-lactamases are available Krajnc et al. (2019a), Krajnc et al. (2019b), Lang et al. (2020), Tooke et al. (2020), we decided to use BBI for which such crystallographic data are not available. Among other things, we wanted to prove that the multi-scale modeling and simulation procedures used in this study are so effective that they can be applied to other boron inhibitors of biomedical importance for which direct data from crystallographic or low-temperature electron microscopy of their complexes with protein targets are not available.

MATERIALS AND METHODS

In order to accurately simulate the covalent binding process, quantum mechanical and classical molecular mechanical approaches (QM/MM) have to be applied. In particular, a time-dependent molecular dynamics model with a quantum description of structural changes inside the active site should be used, applying quantum-classical molecular dynamics simulations. A number of multiscale molecular modeling and bioinformatics methods are applied in this study, which is briefly discussed below. An overview of multi-scale molecular modeling methods is presented among others in the following reviews

Lesyng and McCammon (1993), Lesyng and Rudnicki (2003), Gruziel et al. (2007), and a number of quantum-classical molecular dynamics models along with their numerical implementations are discussed in Bała et al. (1996), Lesyng (1999), Bała et al. (2000), Grochowski and Lesyng (2003). Selected methodologies described above have been used in this study.

Thus, in this study, motions of all atomic nuclei are governed by Newtonian dynamics, and electronic degrees of freedom are described by the evolution of the multi-electronic wave function in an adiabatic way. By adiabatic quantum dynamics, we mean an approximate theoretical model where the wave function of the quantum domain is always in the lowest electronic eigenstate, dependent exclusively on instantaneous positions of the atomic nuclei (molecular geometry based on the Born-Oppenheimer approximation). This approach will hereinafter be referred to as adiabatic quantum-classical molecular dynamics. When it comes to selecting the quantum method, its practical applications in drug design should rather be limited to fast semi-empirical or DFT methods. Within this theoretical framework, taking also into account the accuracy of the QM/MM methods and their computational MD performance, three semiempirical quantum mechanical models PM6, PM7 Stewart (2013), Stewart (2016), and DFTB3 Gaus et al. (2011) deserve attention. PM6 and PM7 have been developed on the basis of the NDDO (Neglect of Diatomic Differential Overlap) approximation. DFTB stands for Density Functional Tight-Binding method. Regarding PM6 and DFTB3 methods, *aposteriori* (D3H4) corrections were designed to improve the hydrogen bond and dispersion interaction energies, Řezáč and Hobza (2012), and Řezáč (2017). The PM6 + D3H4, PM7, and DFTB3 + D3H4 models, as well as a few others, have been tested, including in the analysis non-bonded interactions of inhibitors with enzymes (Križ and Řezáč, 2020).

In this study time-dependent QM/MM simulations were carried out using a high-performance NAMD molecular modeling environment with an implemented quantum potential energy generator MOPAC (Melo et al., 2018; Phillips et al., 2020). In more detail, the NAMD classical environment, VMD visualization environment, and the quantum mechanical package MOPAC Stewart (2016), were integrated into a comprehensive, customizable, and easy-to-use suite. Regarding technical aspects of the hybrid QM/MM model implemented in the NAMD environment, such as definitions of QM regions, mechanical and electrostatic embedding, link atoms, point charge alternations, link atom charge, and charge groups, see http://www.ks.uiuc.edu/Research/qmmm/.

When deciding to use the NAMD environment, the user has a choice of two semi-empirical quantum mechanical models: PM7, or PM6 with a mentioned above *a posteriori* D3H4. While PM6 + D3H4 is slightly better at predicting free energy differences of non-bonded inhibitors interacting with enzymes (Križ and Řezáč, 2020) PM7 has an advantage in predicting structures and heat of formation of chemical bonds. It was parameterized using experimental and advanced *ab initio* reference data, augmented by a new type of reference. In particular, molecular structures containing chemical bonds with boron are slightly better represented in the PM7 model, http://openmopac.net/PM7_and_PM6-D3H4_accuracy/Survey_of_Solids.html/. It has also been validated in the combined docking with classical the MMF94 force field (Sulimov et al., 2017). It should also be noted that time-dependent QM/MM simulations with PM7 Hamiltonian were successfully applied in the theoretical studies of boron-doped derivatives of graphene (Chaban and Prezhdo, 2016).

Therefore, in this work, a hybrid computational model was applied, consisting of the high-performance NAMD computational modeling environment with the classical CHARMM36 force field Vanommeslaeghe et al. (2010), and the MOPAC/PM7 quantum module–serving as the generator of the Born-Oppenheimer potential energy surface.

Structures deposited in Protein Data Bank Berman et al. (2000) of the following β-lactamases: KPC-2 β-lactamase (class A, PDB code 3RXX; Ke et al. (2012)), GC1 β-lactamase (class C, PDB code 1Q2Q; Venkatesan et al. (2004)), and OXA-24 (class D, PDB code 4F94; June et al. (2013)) were studied. With regard to OXA-24 β-lactamase, two forms were considered - wild type, containing carboxylated Lys 84 and its mutant - Lys 84 to Asp 84, denoted hereafter as K84D. The reason for considering the K84D mutant was that we were not able to observe covalent binding in the wild-type. Such a mutation has already been described Schneider et al. (2011), and as indicated the mutant was capable to bind and hydrolyze carbapenems, as well as to covalently bind inhibitors. Considering these biochemical properties and similar electrostatic potential the K84D mutant might effectively mimic the behavior of wild-type.

The covalent bond between 3-NPBA and Ser 70 was removed from KPC-2 β-lactamase structure, and substrate-state protonation was restored using Molecular Modeling Environment 2019 (MOE, 2019). The structure was optimized by the energy minimization, applying the AMBER10 force field Case et al. (2008) utilizing mild string restraints on aromatic ring carbons in order to preserve ring orientation. **Figure 3** presents KPC-2 β-lactamase with the docked 3-NPBA substrate molecule. As finding the structural core of considered proteins and optimizing structural alignment was not trivial, DAMA - a novel multiple structure alignment method was applied (*Essentia Proteomica* server, https://dworkowa.imdik.pan.pl/EP/DAMA/, Daniluk and Lesyng, 2011; Daniluk et al., 2021). Using DAMA, the KPC-2, GC1, and OXA-24 structures were compared, analyzed, and aligned with the identification of the biologically significant similarities. 3-NPBA in an orientation corresponding to the geometry of the optimized KPC-2 β-lactamase complex was introduced to GC1 and OXA-24. The BBI inhibitor was docked in a very similar way as 3-NPBA. Its initial conformation was derived from the structure of *Taniborbactam*. The inhibitor was inserted into the KPC-2, GC1, and OXA-24 binding pockets and oriented similarly to the *Taniborbactam* poses in the crystallographic 6TD1, 6YEN, and 6RTN structures Krajnc et al. (2019a), Krajnc et al. (2019b), Lang et al. (2020), Tooke et al. (2020), respectively. Covalent bonds between serine and the boron atom were removed. The resulting systems were optimized by applying the ligand and pocket sidechains minimization in the AMBER10 force field without

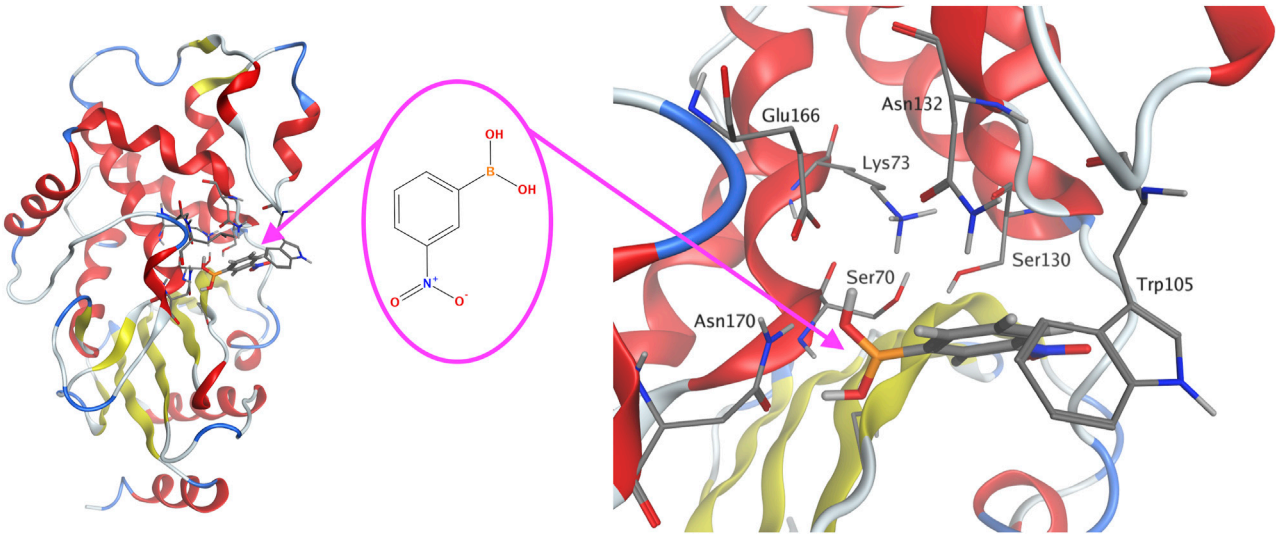


FIGURE 3 | KPC-2 β-lactamase with the docked 3-NPBA substrate molecule **(left)**. The chemical formula of 3-NPBA **(middle)**. The optimized molecular structure with an atomic resolution of KPC-2 β-lactamase focused on the active site with 3-NPBA docked **(right)**. The distance between Ser 70 hydroxyl oxygen and the boron atom of 3-NPBA was 2.57 Å. The β-lactamase α-helices are shown in red, β-strands in yellow and turns in blue.

TABLE 1 | Approximate time-periods characterizing the covalent docking process between 3NPBA and the members of the A, C, and D β-lactamases classes. 3-NPBA for simplicity is denoted as NPB. $T_{arrival}$ is the time to reach a given state on the reaction path, calculated from the beginning of the process. $T_{duration}$ is the average time to reach the next state, calculated from the previous one. All times, along with their standard deviations in parentheses, are given in ps. The atomic names follow the CHARMM force field nomenclature as well as the IUPAC nomenclature, and show which atom has moved towards which atom in another molecular fragment. The distances describing the changes in geometry are given in Å. Graphical representations of the reaction steps are shown in **Figures 4**, **6**. In class C, stage #A2 was observed only in the second simulation, reversing step #A1 and leading back to the structure as in step #1. The simulations in class C which were stabilized at stage #A1 resulted in blocking step #2.

Step	Description of the transfer process (CHARMM)	Description of the transfer process (IUPAC)	Change of the distance	$T_{arrival}$				
				First simul.	Second simul.	Third simul.	Average $T_{arrival}$	Average $T_{duration}$
			Class A – KPC-2					
#1	Lys73:HZ3 → Glu166:OE2	Lys73:N^6-H → Glu166:O^5	2.0 → 1.0	0.20	0.55	0.27	0.34 (±0.15)	0.34 (±0.15)
#2	Glu166 (proton rearrangement)	Glu166 (proton rearrangement)	–	9.05	2.95	24.28	12.09 (±8.97)	11.75 (±9.06)
#3	NPB:B → Ser70:OG	NPB:B^1 → Ser70:O^3	2.1 → 1.6	9.35	5.75	24.95	13.35 (±8.33)	1.26 (±1.10)
#4	Ser70:HG → Lys73:NZ	Ser70:O^3-H → Lys73:N^6	2.0 → 1.0	9.35	6.98	25.20	13.84 (±8.09)	0.49 (±0.53)
#5	NPB:B → Ser70:OG	NPB:B^1 → Ser70:O^3	1.6 → 1.5	10.10	7.00	25.40	14.17 (±8.04)	0.32 (±0.31)
			Class C – GC1					
#1	NPB:B → Ser64:OG	NPB:B^1 → Ser64:O^3	2.1 → 1.6	0.20	0.15	0.22	0.19 (±0.03)	0.19 (±0.03)
#A1	Lys318:HZ3 → Tyr150:OH	Lys318:N^6-H → Tyr150:O^4	1.7 → 1.1	0.67	0.20	0.30	0.39 (±0.20)	0.20 (±0.19)
#A2	Tyr150:HZ3 → Lys318:NZ	Tyr150:O^4-H → Lys318:N^6	1.5 → 1.1	–	0.95	–	0.95 (±0.00)	0.75 (±0.00)
#2	Ser64:HG → Tyr150:OH	Ser64:O^3-H → Tyr150:O^4	2.5 → 1.0	–	1.35	–	1.35 (±0.00)	0.40 (±0.00)
	NPB:B → Ser64:OG	NPB:B^1 → Ser64:O^3	1.6 → 1.5					
			Class D – OXA-24, mutant K84D					
#1	NPB:B → Ser81:OG	NPB:B^1 → Ser81:O^3	2.1 → 1.6	0.25	0.30	0.43	0.33 (±0.07)	0.33 (±0.07)
#2	H2O:H1 → Asp84:OD2	H_2O:O-H → Asp84:O^4	2.0 → 1.0	0.38	0.50	0.55	0.48 (±0.07)	0.16 (±0.04)
	Ser81:HG → H_2O	Ser81:O^3-H → H_2O	2.0 → 1.0					
#3	NPB:B → Ser81:OG	NPB:B^1 → Ser81:O^3	1.6 → 1.5	0.43	0.50	0.57	0.50 (±0.06)	0.03 (±0.02)

TABLE 2 | Approximate time-periods characterizing the covalent docking process between the bicyclic boronate inhibitor (BBI) and the members of the A, C, and D β-lactamases classes. $T_{arrival}$ is the time to reach a given state on the reaction path, calculated from the beginning of the process. $T_{duration}$ is the average time to reach the next state, calculated from the previous one. All times, along with their standard deviations in parentheses, are given in ps. The atomic names follow the CHARMM force field nomenclature as well as the IUPAC nomenclature, and show which atom has moved towards which atom in another molecular fragment. The distances describing the changes in geometry are given in Å. Graphical representations of the reaction steps are shown in **Figures 7**, **8**. In class D, two distinct paths involving one or two H_2O molecules in step #3 were observed, and are denoted as #A3 and #B3.

Step	Description of the transfer process (CHARMM)	Description of the transfer process (IUPAC)	Change of the distance	$T_{arrival}$				Average $T_{duration}$
				First simul	Second simul	Third simul.	Average $T_{arrival}$	
			Class A – KPC-2					
#1	BBI:B → Ser70:OG	BBI:B^2 → Ser70:O^3	2.8 → 1.7	0.26	0.37	0.57	0.40 (±0.13)	0.40 (±0.13)
#2	Ser70:HG → H_2O	Ser70:O^3-H → H_2O	2.0 → 1.0	0.28	0.42	0.61	0.44 (±0.14)	0.03 (±0.01)
	H_2O:H_2 → Glu166:O	H_2O:O-H → Glu166:O^5	1.7 →1.0					
#3	BBI:B → Ser70:OG	BBI:B^2 → Ser70:O^3	1.7 → 1.5	0.28	0.42	0.62	0.44 (±0.14)	0.003 (±0.004)
			Class C – GC1					
#1	BBI:B → Ser64:OG	BBI:B^2 → Ser64:O^3	2.8 → 1.7	0.35	0.32	0.38	0.35 (±0.02)	0.35 (±0.02)
#2	Ser64:HG → Tyr150:OH	Ser64:O^3-H → Tyr150:O^4	2.5 → 1.0	0.65	0.35	0.47	0.49 (±0.12)	0.14 (±0.12)
	BBI:B → Ser64:OG	BBI:B^2 → Ser64:O^3	1.7 → 1.5					
			Class D – OXA-24, mutant K84D					
#1	BBI:B → Ser81:OG	BBI:B^2 → Ser81:O^3	2.9 → 1.7	0.45	0.37	0.44	0.42 (±0.03)	0.42 (±0.03)
#2	Ser81:HG → H_2O(1)	Ser81:O^3-H → H_2O(1)	1.7 → 1.0	0.47	0.43	0.47	0.46 (±0.02)	0.03 (±0.01)
#A3	H_2O(1):H1 → Asp84:OD1	H_2O(1):O-H → Asp84:$O^{4'}$	1.7 → 1.0	0.55	0.55	–	0.55 (±0.00)	0.07 (±0.05)
#B3	H_2O(1):H1 → H_2O(2)	H_2O(1):O-H → H_2O(2)	1.7 → 1.0	–	–	0.45	0.45 (±0.00)	0.004 (±0.00)
	H_2O(2):H1 → Asp84:OD2	H_2O(2):O-H → Asp84:O^4	1.7 → 1.0					
#4	BBI:B → Ser81:OG	BBI:B^2 → Ser81:O^3	1.7 → 1.5	0.55	0.55	0.57	0.56 (±0.01)	0.04 (±0.01)

any restraints. Next, structures of all complexes were carefully virtually titrated at pH = 7. For titration purposes, we used the generalized Born model Labute (2008), Labute (2009) implemented in Protonate 3D of the MOE modeling environment. The method uses a modified version of the Generalized Born/Volume Integral (GB/VI) formalism for implicit solvent electrostatics. Based on our previous research and experience with generalized Born models Wojciechowski and Lesyng (2004) we judge the above choice as sufficiently precise. The applied procedure resulted in assigning protons to all dissociable functional groups. Protonate 3D includes side-chain rotamer, tautomer, and ionization states of all chemical groups during the course of the free energy minimization of a system. It is very important modeling element in this project. In particular, effective electrostatic interactions determine the zwitterionic structure of the boron-based inhibitors, as well as the protonation states of selected water molecules located in the enzyme active sites. The refined structures were used for the initial conditions to start time-dependent QM/MM simulations using NAMD. Structures were solvated, Na+ and Cl-ions were introduced to neutralize the total charge of the systems and provide ionic strength to 0.05 M. Periodic boundary conditions were applied. The quantum domain consisted of the 3-NPBA (or BBI) inhibitor and all protein residues and water molecules containing at least one atom at a distance smaller than 5 Å from 3-NPBA or 7 Å from BBI. BBI is larger and more elongated than 3-NPBA, therefore its quantum domain has been assumed to be larger. As mentioned, the PM7 Hamiltonian was applied. Prior to productive time-dependent QM/MM simulations, the molecular systems were once more optimized, next thermalized, and equilibrated for 20 ps The productive simulations typically consisted of 20 independent 50 ps runs with a 0.5 fs time-step. The MD trajectories were analyzed using the VMD graphical interface (Humphrey et al., 1996).

The energy barriers for both ligands and all three classes of complexes were estimated applying the steered MD technique (SMD), which has been executed applying the NAMD Adaptive Biasing Forces (ABF) module. For the QM domain, a minimal subset of residues required to construct the reaction pathway has been selected (i.e., only the residues listed in **Table 1** and **Table 2**). The distance bias has been applied to the inhibitor's boron atom and the target serine hydroxyl oxygen atom (Ser 70, Ser 64, Ser 81 for the A, C, and D classes, respectively). A force constant of 5,000 kcal/mol/$Å^2$ has been applied during 5 ps simulations to reduce the B–Ser oxygen distance up to 1.4 Å from that one obtained during the docking procedure. All other conditions were kept exactly the same as in the production runs. Since the propagation of reaction in SMD simulation is significantly slowed down, the QM subsystem energy has been smoothened by the moving average applying a 250 fs window.

Regarding the optimal simulation procedures, the so-called computational assays, for enzymatic activities, including also β-lactamases, several valuable publications dealing with

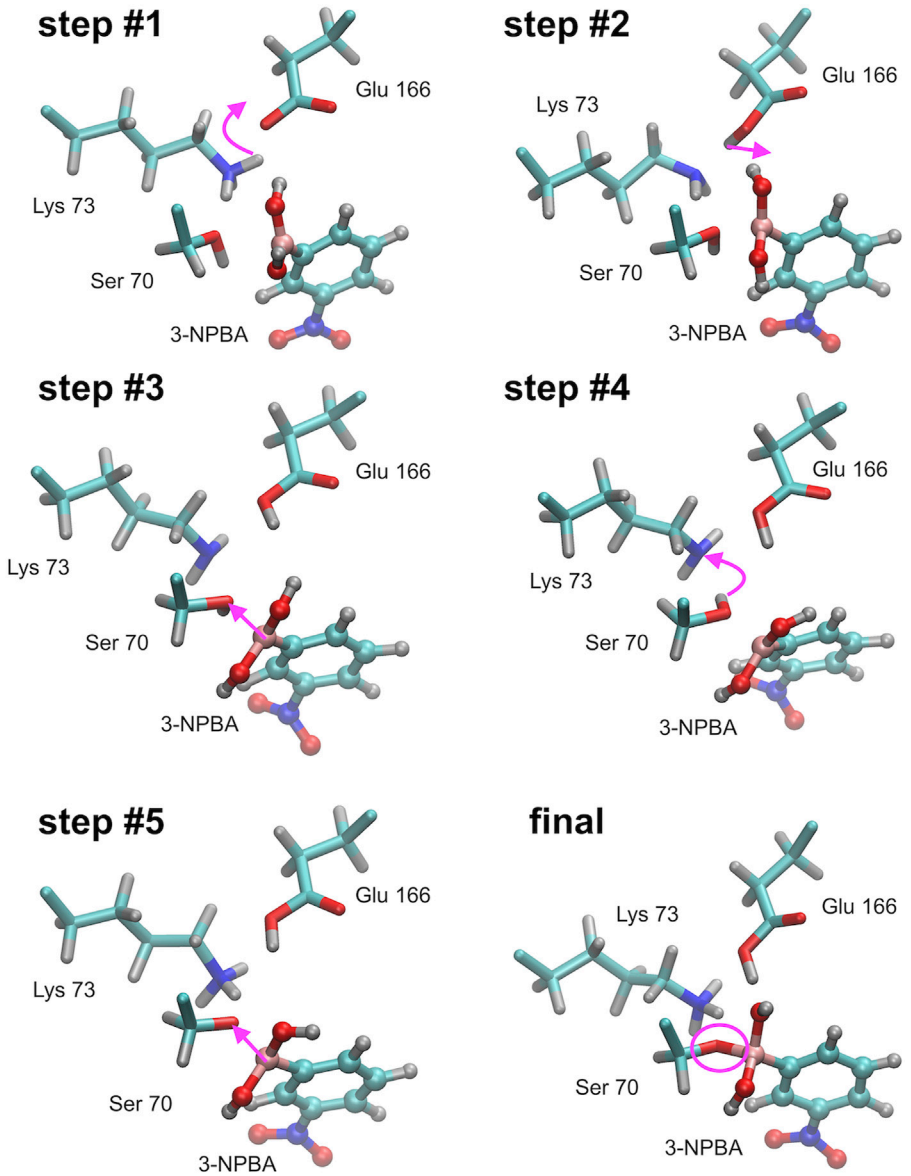


FIGURE 4 | Successive steps of the chemical reaction leading to the bond formation between the Ser 70 hydroxyl group of KPC-2 β-lactamase, and the boron atom of 3-NPBA. Most important molecular and atomic labels are given in **Figure 6** (Class A). Curved arrows indicate the proton transfer processes, straight arrows indicate distance reduction. Step #1 – proton transfer from Lys 70 to Glu 166 (Ser 70 - 3-NPBA-boron distance, 2.1 Ă). Step #2 – proton location rearrangement leading to hydrogen bond formation with 3-NPBA-boron (Ser 70 - 3-NPBA-boron distance, 2.1 Ă). Step #3 – the boron atom moves towards the Ser 70 hydroxyl group (Ser 70 - 3-NPBA-boron distance, 1.6 Ă). Step #4 – proton transfer between Ser 70 and Lys 73 (Ser 70 - 3-NPBA-boron distance, 1.6 Å). Step #5 – a movement of the boron atom towards the Ser 70 hydroxyl group (Ser 70 - 3-NPBA-boron distance, 1.5 Å). Final structure–the chemical bond is formed between the hydroxyl oxygen of Ser 70 and the boron atom of 3-NPBA (Ser 70 - 3-NPBA-boron distance, 1.50 Å). Step #4 and step #5 can be described as representatives of the leading transition state on the reaction path.

formulation simulation standards for QM/MM methods should be mentioned, in particular (Mulholland, 2005; Fonseca et al., 2012; Van der Kamp and Mulholland., 2013; Chudyk, et al., 2014; Lence, et al., 2018; Hirvonen et al., 2019). Special attention should be devoted to an overview of biomolecular simulations and computational assays (Huggins et al., 2019). However, with regard to covalent docking with boron-based inhibitors, it would be very difficult to directly apply a simulation protocol discussed in the mentioned above works. Also, quite recently an electronegativity equalization method at the σπ level (ABEEMσπ) polarizable force field (ABEEMσπ PFF) of boronic acid and β-lactamase has been developed to determine the potential energy functions and parameters (Lu et al., 2020). This should work for the complex of boronic acid with β-lactamases, and may, however, not be working well for the complex of the bicyclic boronate inhibitor with β-lactamases.

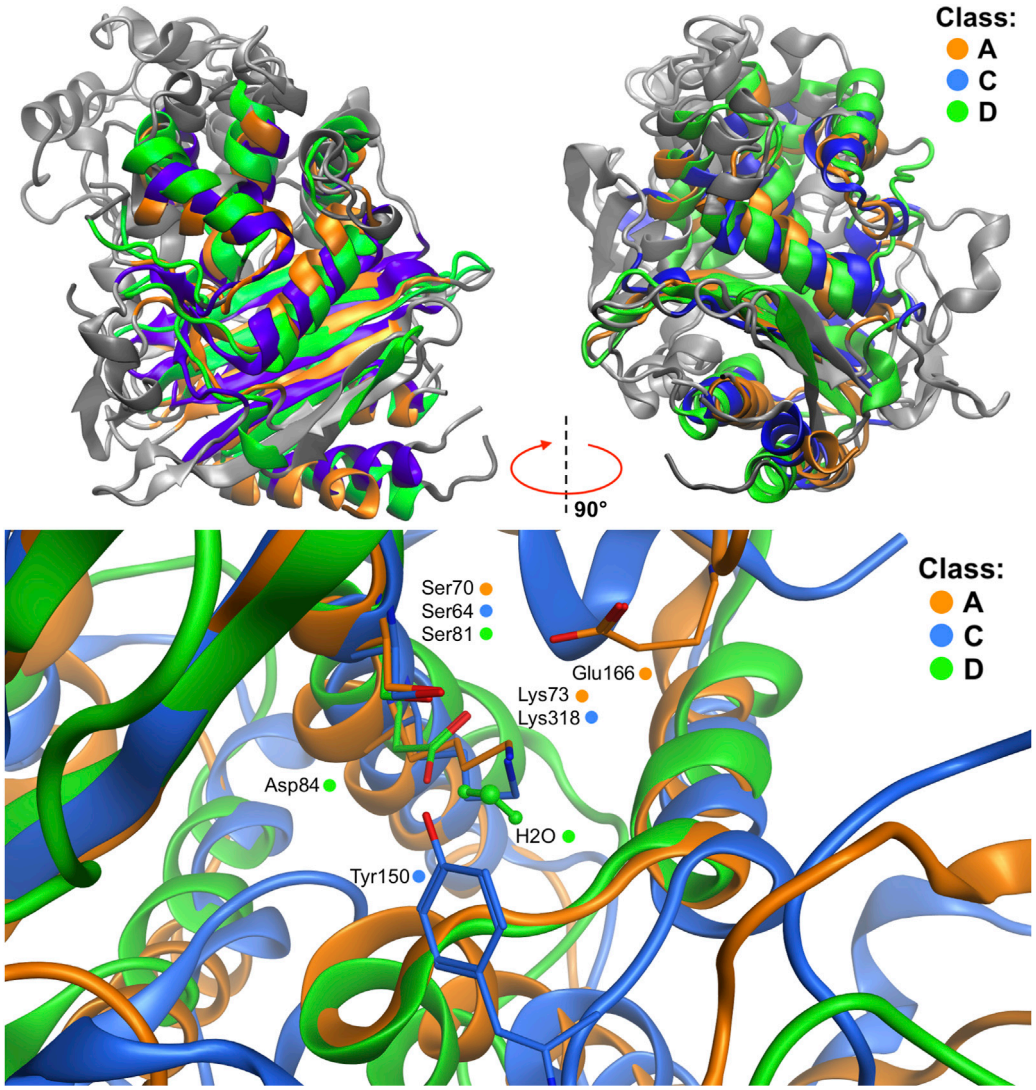


FIGURE 5 | Structural multiple-alignment of the β-lactamases members for the A (KPC-2 β-lactamase), C (CG1 β-lactamase), and D (OXA-24 β-lactamase) enzyme classes. Global alignment with two orientations **(top)**, and presentation of the active sites **(bottom)** resulting from the global alignment. Water residue in class D active site is depicted with balls and sticks representation. Residues crucial in the QM/MM simulations are presented as sticks.

In conclusion, in our opinion, the semi-empirical Hamiltonian PM7 used in this work mimics quite well more advanced *ab initio* methods for enzymatic reactions involving boron atoms. Embedding the QM kernel in the NAMD environment is fully functional and effective. As mentioned, a very important element of this type of modeling is careful titration of the initial molecular state, followed by the energy minimization, thermalization, and equilibration tasks.

RESULTS

The QM/MM simulations for KPC-2, GC1, and K84D mutant of OXA-24 have successfully reproduced the entire enzymatic process leading to the chemical bond formation between the boron atom of the inhibitor and the hydroxyl group of the catalytic Ser 70 residue (Ser 64 and Ser 81 for GC1 and OXA-24, respectively). The reaction didn't occur in the case of the wild type OXA-24 enzyme. We suspect that the presence of carboxylated lysine might be connected with some electrostatic/protonation changes, which we were not able to identify and include in the simulations.

The simulation results will be more precisely described for the complex of 3-NPBA with KPC-2. The movie of the whole process for this enzyme is available (**Supplementary Video S1**). During all independent QM/MM simulations the reaction scenario appeared to be the same, indicating intermediate stages and proton transfer processes in the active site. The only differences refer to the time-periods of successive processes. The activation process consists of three steps: first Lys 73 is

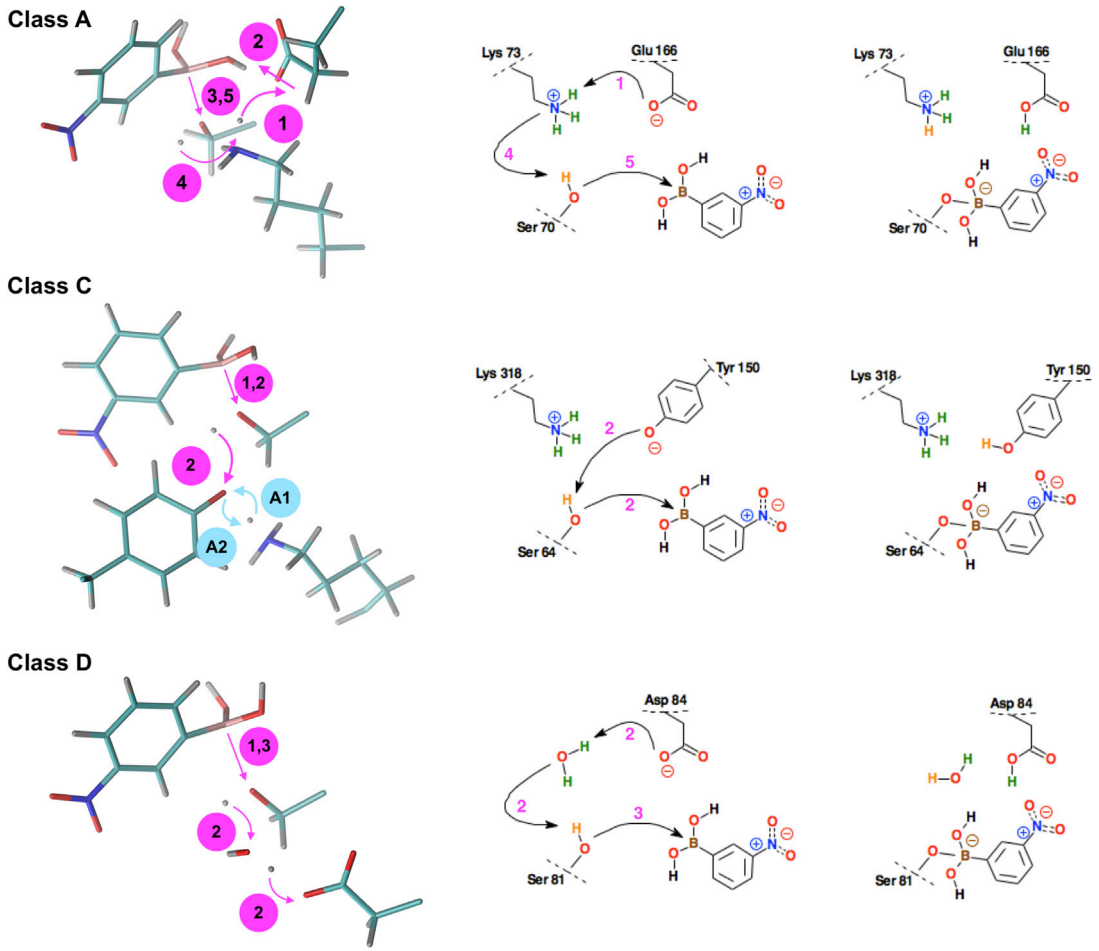


FIGURE 6 | Schematic representation of the successive steps in the formation of the chemical bond between the boron atom of the 3-NPBA inhibitor and the hydroxyl groups of serine β-lactamases. The active site residues belong to KPC-2 β-lactamase (class A), GC1 β-lactamase (class C), and the K84D mutant of OXA-24 β-lactamase (class D). On the left - curved arrows indicate the proton transfer processes, straight arrows indicate distance reduction. In class C, steps #A1 and #A2 (denoted by blue annotations) were observed, however, they are mutually reversible, leading finally to the structure as in step #1. On the right, conventional simplified diagrams showing reaction steps - the arrows symbolically indicate the electron transfer processes. Description of the time-periods related to the presented steps is given in **Table 1**.

deprotonated by Glu 166, then one observes rearrangement of the hydrogen bond system in the active site resulting in a hydrogen bond of Glu 166 with one of the hydroxyls of 3-NPBA, and then the key Ser 70 is deprotonated by Lys 73. This immediately allows the formation of the chemical bond between deprotonated hydroxyl of Ser 70 and the boron atom of 3-NPBA. The described above processes are visualized (**Figure 4**). Note that the hybridization of the boron atom has changed from sp2 (initial state) to sp3 (final state). Simulation correctly predicts changes of the distances between functional groups in the active site, resulting in the final structure very similar to the crystallographic 3RXX structure.

As for the QM/MM simulations of the covalent docking process to representatives of classes C and D β-lactamases, they were preceded by a detailed comparison of the structures of GC1 and OXA-24 with the reference enzyme KPC-2. Global structural multiple-alignment indicated all essential similarities (**Figure 5**). 10 long, 100 ps, QM/MM simulations of wild-type OXA-24 were carried out in order to check whether there were any unforeseen intuitive processes leading to covalent docking. The simulations, however, did not bring the system out of its initial state. Therefore, similar yet free from post-translational modifications K84D mutant was applied. In this case, covalent docking has proved successful. Scenarios for the covalent docking processes for the members of A, C, and D classes were identified and compared (**Figure 6**). A detailed description of the covalent docking reaction steps between 3-NPBA and the above-mentioned three β-lactamases, as well as their approximate time-periods, are described (**Table 1**). In the case of class C and D β-lactamases, one sees slightly different scenarios for the formation of a chemical bond between the boron atom of 3-NPBA and the oxygen of the Ser 64 (GC1) or Ser 81 (OXA-24) hydroxyl group, compared to class A. Specifically for GC1 β-lactamase, although the productive reaction path is shorter than

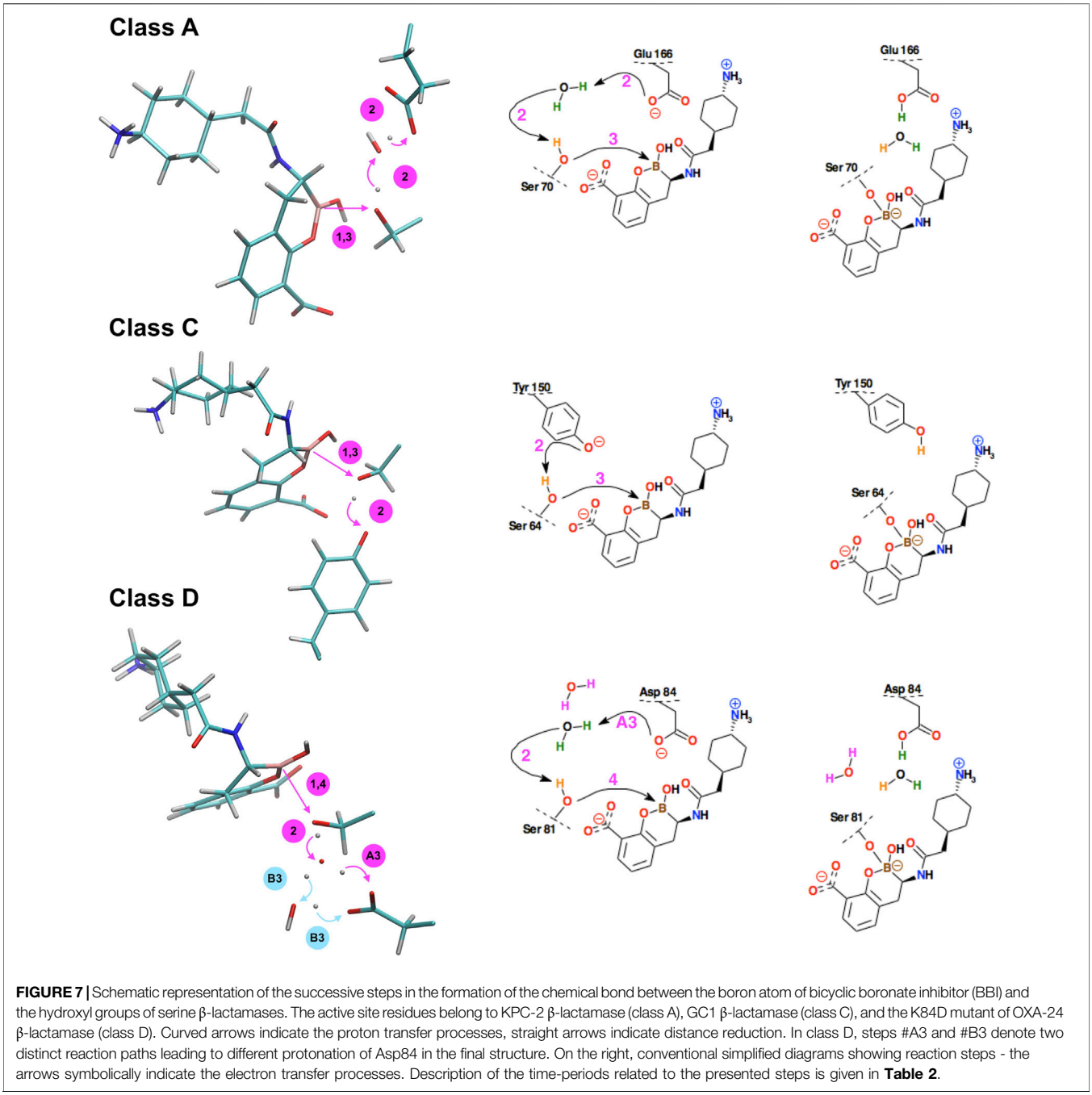


FIGURE 7 | Schematic representation of the successive steps in the formation of the chemical bond between the boron atom of bicyclic boronate inhibitor (BBI) and the hydroxyl groups of serine β-lactamases. The active site residues belong to KPC-2 β-lactamase (class A), GC1 β-lactamase (class C), and the K84D mutant of OXA-24 β-lactamase (class D). Curved arrows indicate the proton transfer processes, straight arrows indicate distance reduction. In class D, steps #A3 and #B3 denote two distinct reaction paths leading to different protonation of Asp84 in the final structure. On the right, conventional simplified diagrams showing reaction steps - the arrows symbolically indicate the electron transfer processes. Description of the time-periods related to the presented steps is given in **Table 2**.

for KPC-2, there is a possible dead-end involving Lys 318 (**Figure 6**, blue annotation), which does not lead to the final state, and only withdrawal from it may lead to the effective formation of the above-mentioned bond. In turn, for the K84D mutant of OXA-24 β-lactamase, an important role in the enzymatic process is played by the water molecule from the nearby environment of this enzyme. After deprotonation of the water molecule, the formed hydroxyl anion acts as a very strong base, ultimately leading to the proton dissociation from the hydroxyl group of Ser 81, and effective interaction of the deprotonated hydroxyl group with the boron atom, resulting in the formation of a stable chemical bond -B-O-, as in the previous cases. The most important fragments of the covalent 3-NPBA docking processes to GC1 and OXA-24 are visualized in **Supplementary Video S2**and in**Supplementary Video S3**, respectively.

Analyzing the reaction time-periods of the chemical bond formation in 3-NPBA systems, it can be seen that the bottleneck - which in the case of KPC-2 is stage #2, consists of the rearrangement of hydrogen bonds in the active site (11.75 ps). It should be emphasized, however, that there is a significant deviation of times assigned to this process (2.40–24.01 ps), so the

average value for the three analyzed trajectories should be treated as a very approximate one. The remaining steps are in all cases very fast and are in the range of 0.20–0.75 ps It should be noted, however, that in the case of GC1 only one third of the trajectories turned out to be productive, at least in terms of the simulation times. If the other two trajectories were significantly lengthened, they might reach the final state, but the average time of the catalytic process would be much longer. Also, one should note that Tyr 150 at the beginning of simulations is in anionic form, according to computational titration results. Similar simulations with neutral Tyr 150 form did not result in the -B-O- bond formation.

Regarding covalent docking of the BBI inhibitor, the simulation results are more precisely described for the complex of BBI with KPC-2. Successive steps of the chemical reaction leading to the bond formation between the Ser 70 hydroxyl group of KPC-2 β-lactamase, and the boron atom of BBI are presented (**Figure 7**). The catalytic site of class A beta-lactamases containing bicyclic boronates contains additional water molecule, which is absent in crystallographic structures of non-cyclic boronates (like 3-NPBA). The presence of this particular water molecule, often called *deacyling water* due to its function in beta-lactam catalysis, suggest that the course of bond formation with BBI might be different. Indeed, the reaction follows a different path, where water mediates the proton transfer form Ser 70 to Glu 166. This pathway is free from the bottleneck identified in 3-NPBA binding. In resulting structure Glu 166 forms hydrogen bond with the water molecule, which in turn forms another hydrogen bond with the ligand's hydroxyl group (using hydrogen transferred from Ser 70). The movie of the whole covalent docking process for this enzyme is available (**Supplementary Video S4**).

The covalent docking of BBI to GC1 took exactly the same course as observed in 3-NPBA, however the interference of Lys was not observed in any simulation. BBI docking to OXA-24 follows similar steps as 3-NPBA, which involve one water molecule mediating proton transfer from Ser81 to Asp84. However, an alternative pathway, involving two water molecules, has been observed in this system, which, interestingly, leads to slightly different final state. The proton transferred through both water molecules binds to different oxygen of Asp 84 carboxyl group.

Most important fragments of the covalent BBI docking processes to GC1 are visualized in**Supplementary Video S5**, and to OXA-24 in **Supplementary Video S6**as well as Supplementary Video_7 (two reaction paths for OXA-24, identified). A detailed description of the covalent docking reaction steps between the BBI covalent inhibitor and the above-mentioned three β-lactamases, as well as their approximate time-periods, are described (**Table 2**). Interestingly, the intermediate Ser oxygen–boron distance, in which the system waits for proton transfers to take place, is 0.1 Å shorter in BBI.

Due to the almost immediate timescale of the reaction steps, SMD simulations were used to shed a light on the energy profile of the reaction pathways. A few energy barriers has been observed. One should stress, however, the following. The dynamics of the reaction site is highly cooperative. In the case under study, the surface of the chemical reaction consists of a series of narrow maxima in the configuration space, that are traversed as a result of the strong coupling of the reaction site with the dynamics of the enzyme as a whole. In such a situation, an arbitrary choice of a one-dimensional reaction path in the SMD process must always lead to an overestimation of the energy barriers. The classical Potential of Mean Force (PMF) in this case is not a well-defined physical property. After all, the system does not have to choose the reaction path that we indicated, because the path is multidimensional, and we significantly disturb the natural dynamics of enzymes. The one-dimensional choice is justified when it concerns, for example, the slow degree of freedom. Although in this case, we chose a fairly natural reaction path, which is the distance between the boron atom and the oxygen of the serine hydroxyl group, it is probably not the slowest degree of freedom, because the boron atom strongly interacts with the oxygen atom. To summarize the applied SMD procedure gives us only an upper estimate of the PMF - the real barriers are smaller. Below are the upper estimates.

In the case of 3-NPBA ligand, a barrier below 11 kcal/mol occurs between 1.5–4.0 ps (B–Ser:O distance: 2.2 → 1.7 Å) for class A, below 7 kcal/mol between 1.6–3.0 ps (2.6 → 1.9 Å) for class C, and below 5 kcal/mol between 1.0–1.8 ps (2.3 → 2.1 Å) for class D.

For BBI a barrier below 9 kcal/mol occurs between 0.2–2.2 ps (B–Ser:O distance: 2.8 → 2.2 Å) for class A, below 12 kcal/mol between 0–2.5 ps (2.8 → 2.1 Å) for class C, and below 12 kcal/mol between 1.0–3.0 ps (2.7 → 2.0 Å) for class D. Representative energy profiles for each system are illustrated in **Supplementary Figures S3, S4**).

It should be added that in real systems, such as those studied in this paper, most likely quantum tunneling effects play an indispensable role due to cooperative proton transfer processes. This can lower the barriers even further. These effects can be accounted for by applying quantum dynamics, based explicitly on the time-dependent Schroedinger equation. This type of physics was, inter alia, used in a full description of the phospholipase A_2 reaction mechanism and model proton transfer systems Bala et al. (1996), Lesyng (1999), Bala et al. (2000), Grochowski and Lesyng (2003). This requires, however, much faster than PM6, PM7 or DFTB3 generators of the potential energy surface.

DISCUSSION

Multiscale molecular modeling methods have been applied in covalent docking studies using two boron-based inhibitors (3-NPBA and BBI, representing two large families of inhibitors, characterized by boronic acid and bicyclic boronate scaffolds), and three serine β-lactamases (KPC-2, GC1 and OXA 24, representing the A, C and D enzyme classes). This resulted in six successful covalent docking events and seven reaction paths (two paths identified for OXA-21). It should be noted that due to the presence of the functional -NO2 group in 3-NBPA, and -COO- in BBI the inhibitors are quite reactive.

In the main thread of this research, the hybrid, quantum-classical molecular mechanics and adiabatic quantum-classical molecular dynamics model were applied. The DAMA and MOE modeling environments were used in the analysis of the biomolecular structures, virtual titration, and their preparation to the simulations of the covalent docking processes. The QM/MM model implemented in the NAMD/VMD environment, with the approximate PM7 Hamiltonian describing the quantum domains of the enzymes, appeared to be a very effective approach to simulating the covalent docking processes. From the computational point of view, and with the current state of development of computing techniques, adiabatic time-dependent QM/MM simulations are fully acceptable for practical applications for molecular design or drug design experiments. 3D visualization with VMD and MOE tools, based on accurate physical simulations of this enzymatic process, provided detailed information on the very specific role of active site amino acid residues, water molecules and boron-containing molecular fragments, in the covalent docking process. Class B, which contains significantly different metallo (Zn^{+2}) β-lactamases, has not been studied.

The covalent 3-NPBA docking mechanism to KPC-2 has been described in great detail. Other systems with less accuracy, but the attached diagrams, tables and videos also allow to recreate all the details. In particular, results of the QM/MM simulations of the covalent reactions between the boron atom and hydroxyl of Ser 81 are well documented in a form of seven movies available in "Supplementary Material" (**Supplementary Video S1**. MOV, 3-NPBA–KPC-2; **Supplementary Video S2**. MOV, 3-NPBA–GC1; **Supplementary Video S3**. MOV, 3-NPBA–OXA 24; **Supplementary Video S4**. MOV, BBI–KPC-2; **Supplementary Video S5**. MOV, BBI–GC1; **Supplementary Video S6**. MOV and **Supplementary Video S7**. MOV present two reaction paths for the BBI–OXA 24 complex).

In the case of class D for the wild-type enzyme, the formation of the required -B-O- bond was not achieved due to the presence of a carboxylated Lys 84 derivative in the active site. However, it was shown that the K84D mutant very easily leads to the formation of the expected bond due participation of a water molecule from the solvent, which because of the proton dissociation acts as a strong catalytic base initiating the productive reaction path. The positioning of carboxylated Lys 84 in the active site suggests that in a wild-type system, the Ser 81 hydrogen transfer may occur directly to Lys carboxyl oxygen. For members of class A and C, deprotonated Lys 73 and Lys 318, act as such bases, respectively.

To the best of our knowledge, this study is the first one that describes in detail, the course of the covalent docking processes resulting from the quantum description of structural changes in the enzyme active sites, and with the participation of the 3-NPBA inhibitor with the boronic acid scaffold (**Figures 3–6** and **Table 1**), as well as the BBI inhibitor with the bicyclic boronate scaffold (**Figure 7** and **Table 2**). It should be noted that in class D, one or two water molecules play a key role in the deprotonation of Ser 84, which is a prerequisite for the final reaction to form a covalent bond between the boron atom and serine oxygen. The role of these water molecules can be related to the concept of "flip-flop" hydrogen bonds, and/or "water wires", which have been observed in many biomolecular systems, including enzyme systems. For a review of the first experimental observations and quantum-mechanical analysis of the "flip-flop" hydrogen bonds see (Saenger et al., 1985; Jeffrey and Saenger, 1991). Below are some representative publications analyzing the mechanisms of delocalized hydrogen bonds and "water wires" (Cui and Karplus, 2003; Cukierman, 2003; Marx, 2006; Yan et al., 2007; Sakti et al., 2019; Setny, 2020). The analysis of structural conformational changes coupled with proton transfer processes shows that the covalent docking reactions described in this study are highly cooperative processes, and future studies should probably also take into account the quantum-dynamical effects associated with cooperative proton transfer mechanisms.

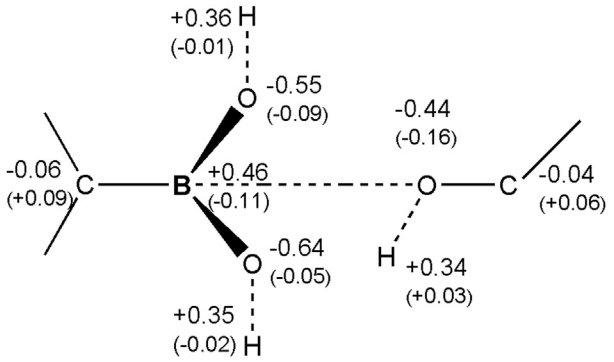


FIGURE 8 | Atomic Mulliken charges on the boron molecular fragment of the 3-NPBA inhibitor, as well as on the hydroxyl group of Ser 70. The charges were computed for the initial state, before the formation of the chemical -B-O- bond. In parentheses are changes of the charges for the final state, after the formation of the chemical bond.

In order to better understand electron charge rearrangements during the covalent docking processes, the Mulliken atomic charges have been computed as functions of time. These data are presented in **Supplementary Figures S3, S4**. There is no indication that any larger electron transfer processes assigned to protons take place. Atomic charges are quite typical and to our surprise, changes of the atomic charges assigned to protons are small. In conclusion, radical reactions involved in covalent docking mechanisms of boron-based inhibitors are highly unlikely - at least for the A, C, and D β-lactamases.

Changes of Mulliken atomic charges, related especially to the change of the hybridization state of atomic orbitals on the boron atom, from the sp^2 to sp^3 type and formation of the -C-B-O- bond are also presented (**Figure 8**). Boron is the electropositive atom in comparison to the oxygen atom. Before the formation of the chemical bond with oxygen in 3-NPBA system, its Mulliken charge was about +0.46 e. After the reaction, this charge drops down to about +0.35 e. In turn, oxygen of the serine hydroxyl group, after creating the chemical bond, increases its negative charge from −0.44 e to −0.60 e, and thus becomes similar to the other two oxygen atoms chemically bonded to the boron atom. In BBI systems similar boron charge drop was observed, from the

+0.49 e to +0.32 e after bond formation. Serine oxygens charge however acted differently and changed from −0.68 e to −0.54 e acquiring similar charge to the heterocyclic oxygen. The charge on oxygen in the BBI boron-attached hydroxyl group changed from −0.70 e to −0.48 e.

The energy profiles for the 3-NPBA–KPC-2 and BBI–KPC-2 complexes were assessed by monitoring the energy changes of the QM subsystem consisting of the residues and the inhibitors directly involved in the reaction, as well as by applying SMD simulations. Visual analysis of the SMD trajectories suggests that the observed energy barriers for both 3-NPBA and BBI in each β-lactamase class are amongst others associated with the tension created by bending boron substituents out of the primary trigonal plain when approaching the serine oxygen atom. We think that right after the system breaks this barrier the boron atom undergoes the morphing process, and the sp^3-associated tetragonal geometry relaxes this tension, resulting in a further energy decrease.

The analysis of structural changes in the enzyme active sites, involving also the 3-NPBA inhibitor, showed during the time-dependent QM/MM simulations several geometric preferences finally leading to the covalent docking processes. Thus, the distance between the boron atom and the oxygen atom of the serine hydroxyl group should be in the order of 3 Å, or slightly less before the chemical bond is formed. This distance can be compared to the sum of the Pauling atomic radii of those atoms, which are 1.65 and 1.42 Å, respectively. The resulting distance, due to long-range non-bonded interactions, and following covalent interactions, rapidly decreases. It should be emphasized that the nonbonded interactions after molecular structure optimization reduce this distance up to 2.6 Å (or to 2.8 Å for bicyclic boron inhibitors). The orientation of the inhibitor's C-B bond in relation to the HO-C serine bond, defined by the C-B-O angle should be close to 90°. It gives the possibility of the interaction of the "empty" $2p_z$ orbital of boron in its initial sp^2 hybridization, with the $2p_z$ orbital of oxygen. Deprotonation of the hydroxyl group allows donation of electrons to the boron atom and its transition to the sp^3 hybridization, thus increasing this angle to the range of 105–120°. After the formation of the chemical bond, the distance between the boron atom and the serine oxygen atom drops down to about 1.5 Å. As already noted, there should be at least one fairly strong basic group in the active site that will be able to sequentially deprotonate the hydroxyl groups bound to the boron atom, as well as the hydroxyl group of serine.

In order to obtain the highly specific effects of boron-based inhibitors, care must be taken that the optimal orientation of the boron atom to a particular serine group is correlated with as many other local interactions as possible. Such interactions can be hydrogen bonds, salt bridges, or localized hydrophobic interactions. The idea is that the chemically active inhibitor should only bind in a highly specific manner to a selected binding site belonging to a particular protein, and therefore that it cannot attack other hydroxyl groups, which could lead to serious side effects.

Presented results allow for a much deeper understanding of the complex covalent docking processes important for molecular medicine, related in particular to the design of novel drugs. They should support the rational design of covalent boron-based inhibitors for β-lactamases, which are the weapon of antibiotic-resistant bacteria, as well as for many other enzyme systems of clinical relevance.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ŁC designed the study, carried out the simulations, provided the results and participated in their analysis. KK participated in the analysis. All authors discussed the results and contributed to subsequent versions of the manuscript. BL supervised the project and prepared the final version of the manuscript.

FUNDING

These studies were supported by Faculty of Physics, University of Warsaw, 501-D111-01-1110102 (PP/BF). Computations were carried out using infrastructure financed by POIG.02.01.00-14-122/09.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.633181/full#supplementary-material

REFERENCES

Ban, H. S., and Nakamura, H. (2015). Boron-based Drug Design. *Chem. Rec.* 15, 616–635. doi:10.1002/tcr.201402100

Bała, P., Grochowski, P., Nowiński, K., Lesyng, B., and McCammon, J. A. (2000). Quantum-dynamical Picture of a Multistep Enzymatic Process: Reaction Catalyzed by Phospholipase A(2). *Biophys. J.* 79, 1253–1262. doi:10.1016/S0006-3495(00)76379-X

Bała, P., Grochowski, P., Lesyng, B., and McCammon, J. A. (1996). "Quantum-Classical Molecular Dynamics. Models and Applications," in *Quantum Mechanical Simulation Methods for Studying Biological Systems. Centre de Physique des Houches, vol 4*. Editors D. Bicout and M. Field (Berlin, Heidelberg: Springer), 119–156. doi:10.1007/978-3-662-09638-3_5

Bello, M. (2018). Advances in Theoretical Studies on the Design of Single boron Atom Compounds. *Cpd* 24, 3466–3475. doi:10.2174/1381612824666181102142432

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

Brem, J., Cain, R., Cahill, S., McDonough, M. A., Clifton, I. J., Jiménez-Castellanos, J.-C., et al. (2016). Structural Basis of Metallo-β-Lactamase, Serine-β-Lactamase

and Penicillin-Binding Protein Inhibition by Cyclic Boronates. *Nat. Commun.* 7, 12406–12414. doi:10.1038/ncomms12406

Bush, K. (2018). Past and Present Perspectives on β-lactamases. *Antimicrob. Agents Chemother.* 62, e01076–18. doi:10.1128/AAC.01076-18

Cahill, S. T., Cain, R., Wang, D. Y., Wareham, D. W., Oswin, H. P., et al. (2017). Cyclic Boronates Inhibit All Classes of β-lactamases. *Antimicrob. Agents Chemother.* 61, e02260-16. doi:10.1128/aac.02260-16

Cahill, S. T., Tyrrell, J. M., Navratilova, I. H., Calvopiña, K., Robinson, S. W., Lohans, C. T., et al. (2019). Studies on the Inhibition of AmpC and Other β-lactamases by Cyclic Boronates. *Biochim. Biophys. Acta (Bba) - Gen. Subjects* 1863, 742–748. doi:10.1016/j.bbagen.2019.02.004

Calvopiña, K., Hinchliffe, P., Brem, J., Heesom, K. J., Johnson, S., Cain, R., et al. (2017). Structural/mechanistic Insights into the Efficacy of Nonclassical β-lactamase Inhibitors against Extensively Drug resistantStenotrophomonas Maltophiliaclinical Isolates. *Mol. Microbiol.* 106, 492–504. doi:10.1111/mmi.13831

Case, D., Darden, T., Cheatham, T., Simmerling, C., Wang, J., Duke, R., et al. (2008). *AMBER 10*. San Francisco: University of California.

Chaban, V. V., and Prezhdo, O. V. (2016). Boron Doping of Graphene-Pushing the Limit. *Nanoscale* 8, 15521–15528. doi:10.1039/c6nr05309b

Chudyk, E. I., Limb, M. A. L., Jones, C., Spencer, J., Van der Kamp, M. W., and Mulholland, A. J. (2014). QM/MM Simulations as an Assay for Carbapenemase Activity in Class A β-lactamases. *Chem. Commun.* 50, 14736–14739. doi:10.1039/C4CC06495J/

Cui, Q., and Karplus, M. (2003). Is a "Proton Wire" Concerted or Stepwise? A Model Study of Proton Transfer in Carbonic Anhydrase. *J. Phys. Chem. B* 107, 1071–1078. doi:10.1021/jp021931v

Daniluk, P., and Lesyng, B. (2011). A Novel Method to Compare Protein Structures Using Local Descriptors. *BMC Bioinformatics* 12, 344–357. doi:10.1186/1471-2105-12-344/

Daniluk, P., Oleniecki, T., and Lesyng, B. (2021). DAMA – a Method for Computing Multiple Alignments of Protein Structures Using Local Structure Descriptors. *Submitted to Bioinformatics*.

Duret, G., Quinlan, R., Bisseret, P., and Blanchard, N. (2015). Boron Chemistry in a New Light. *Chem. Sci.* 6, 5366–5382. doi:10.1039/c5sc02207j

Durka, K., Laudy, A. E., Charzewski, Ł., Urban, M., Stępień, K., Tyski, S., et al. (2019). Antimicrobial and KPC/AmpC Inhibitory Activity of Functionalized Benzosiloxaboroles. *Eur. J. Med. Chem.* 171, 11–24. doi:10.1016/j.ejmech.2019.03.028

Fonseca, F., Chudyk, E. I., Van der Kamp, M. W., Correia, A., Mulholland, A. J., and Spencer, J. (2012). The Basis for Carbapenem Hydrolysis by Class A β-lactamases: A Combined Investigation Using Crystallography and Simulations. *J. Am. Chem. Soc.* 134, 18275–18285. doi:10.1021/ja304460j

Friese, F. W., and Studer, A. (2019). New Avenues for C-B Bond Formation via Radical Intermediates. *Chem. Sci.* 10, 8503–8518. doi:10.1039/c9sc03765a

Gaus, M., Cui, Q., and Elstner, M. (2011). DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *J. Chem. Theor. Comput.* 7, 931–948. doi:10.1021/ct100684s

Graham, B. J., Windsor, I. W., Gold, B., and Raines, R. T. (2021). Boronic Acid with High Oxidative Stability and Utility in Biological Contexts. *Proc. Natl. Acad. Sci. USA* 118 (10), e2013691118. doi:10.1073/pnas.2013691118

Grochowski, P., and Lesyng, B. (2003). Extended Hellmann-Feynman Forces, Canonical Representations, and Exponential Propagators in the Mixed Quantum-Classical Molecular Dynamics. *J. Chem. Phys.* 119, 11541–11555. doi:10.1063/1.1624062

Gruziel, M., Kmieć, P., Trylska, J., and Lesyng, B. (2007). "Selected Microscopic and Mezoscopic Modelling Tools and Models - an Overview," in *Molecular Materials with Specific Interactions – Modeling and Design. Challenges and Advances in Computational Chemistry and Physics*. Editor W. A. Sokalski (Dordrecht): Springer), Vol. 4, 203–223. doi:10.1007/1-4020-5372-X_3

Hermann, J. C., Pradon, J., Harvey, J. N., and Mulholland, A. J. (2009). High Level QM/MM Modeling of the Formation of the Tetrahedral Intermediate in the Acylation of Wild Type and K73A Mutant TEM-1 Class A β-lactamase. *J. Phys. Chem. A.* 113, 11984–11994. doi:10.1021/jp9037254

Hirvonen, V. H. A., Hammond, K., Chudyk, E. I., Limb, M. A. L., Spencer, J., Mulholland, A. J., et al. (2019). An Efficient Computational Assay for β-Lactam Antibiotic Breakdown by Class A β-lactamases. *J. Chem. Inf. Model.* 59, 3365–3369. doi:10.1021/acs.jcim.9b00442

Huggins, D. J., Biggin, P. C., Dämgen, M. A., Essex, J. W., Harris, S. A., Henchman, R. H., et al. (2019). Biomolecular Simulations: From Dynamics and Mechanisms to Computational Assays of Biological Activity. *Wires Comput. Mol. Sci.* 9, e1393. doi:10.1002/wcms.1393

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual Molecular Dynamics. *J. Mol. Graphics* 14, 33–38. doi:10.1016/0263-7855(96)00018-5

Jeffrey, G. A., and Saenger, W. (1991). *Hydrogen Bonding in Biological Structures*. Berlin Heidelberg New York: Springer-Verlag, 1–563. ISBN-13:978-3-540-57903-. doi:10.1007/978-3-642-85135-3

June, C. M., Bonomo, R. A., Leonard, D. A., Powers, R. A., and Powers, R. A. (2013). Structure of the Class D Beta-Lactamase OXA-24 K84D in Acyl-Enzyme Complex with Oxacillin. DOI: doi:10.2210/pdb4F94/pdb

Ke, W., Bethel, C. R., Papp-Wallace, K. M., Pagadala, S. R. R., Nottingham, M., Fernandez, D., et al. (2012). Crystal Structures of KPC-2 β-lactamase in Complex with 3-Nitrophenyl Boronic Acid and the Penam Sulfone PSR-3-226. *Antimicrob. Agents Chemother.* 56, 2713–2718. doi:10.1128/AAC.06099-11

Keshri, V., Panda, A., Levasseur, A., Rolain, J.-M., Pontarotti, P., and Raoult, D. (2018). Phylogenomic Analysis of β-lactamase in Archaea and Bacteria Enables the Identification of Putative New Members. *Genome Biol. Evol.* 10, 1106–1114. doi:10.1093/gbe/evy028

Kríž, K., and Rezać, J. (2020). Benchmarking of Semiempirical Quantum-Mechanical Methods on Systems Relevant to Computer-Aided Drug Design. *J. Chem. Inf. Model.*, 60, 1453–1460. doi: doi:10.1021/acs.jcim.9b01171

Krajnc, A., Brem, J., Hinchliffe, P., Calvopiña, K., Panduwawala, T. D., Lang, P. A., et al. (2019b). Bicyclic Boronate VNRX-5133 Inhibits Metallo- and Serine-β-Lactamases. *J. Med. Chem.* 62, 8544–8556. doi:10.1021/acs.jmedchem.9b00911

Krajnc, A., Lang, P. A., Panduwawala, T. D., Brem, J., and Schofield, C. J. (2019a). Will Morphing boron-based Inhibitors Beat the β-lactamases?. *Curr. Opin. Chem. Biol.* 50, 101–110. doi:10.1016/j.cbpa.2019.03.001

Krivitskaya, A. V., and Khrenova, M. G. (2021). Boronic Acids as Prospective Inhibitors of Metallo-β-Lactamases: Efficient Chemical Reaction in the Enzymatic Active Site Revealed by Molecular Modeling. *Molecules* 26, 2026–2039. doi:10.3390/molecules26072026

Kumar, N., Reddy, R. R., Eghbarieh, N., and Masarwa, A. (2020). α-Borylalkyl Radicals: Their Distinctive Reactivity in Modern Organic Synthesis. *Chem. Commun.* 56, 13–25. doi:10.1039/c9cc08027a

Labute, P. (2009). Protonate3D: Assignment of Ionization States and Hydrogen Coordinates to Macromolecular Structures. *Proteins* 75, 187–205. doi:10.1002/prot.22234

Labute, P. (2008). The Generalized Born/volume Integral Implicit Solvent Model: Estimation of the Free Energy of Hydration Using London Dispersion Instead of Atomic Surface Area. *J. Comput. Chem.* 29, 1693–1698. doi:10.1002/jcc.20933

Lang, P. A., Parkova, A., Leissing, T. M., Calvopiña, K., Cain, R., Krajnc, A., et al. (2020). Bicyclic Boronates as Potent Inhibitors of AmpC, the Class C β-lactamase from *Escherichia coli*. *Biomolecules* 10, 899–914. doi:10.3390/biom10060899

Lence, E., and González-Bello, C. (2021). Bicyclic Boronate β-lactamase Inhibitors: The Present Hope against Deadly Bacterial Pathogens. *Adv. Therap.* 4 (1 of 21), 2000246. doi:10.1002/adtp.202000246

Lence, E., Van der Kamp, M. W., González-Bello, C., and Mulholland, A. J. (2018). QM/MM Simulations Identify the Determinants of Catalytic Activity Differences between Type II Dehydroquinase Enzymes. *Org. Biomol. Chem.* 16, 4443–4455. doi:10.1039/c8ob00066b

Lesyng, B., and McCammon, J. A. (1993). Molecular Modeling Methods. Basic Techniques and Challenging Problems. *Pharmacol. Ther.* 60, 149–167. doi:10.1016/0163-7258(93)90004-W

Lesyng, B., and Rudnicki, W. (2003). "Molecular Modeling in Drug Design," in *Optimization of Aerosol Drug Delivery*. Editors L. Gradoń and J. Marijnissen (Dordrecht: Springer), 23–48. doi:10.1007/978-94-017-0267-6_2/

Lesyng, B. (1999). "Simulations of Biomolecular Systems and Processes: Perspectives and Limitations",". Modelling And Simulation: A Tool For the Next Millenium", *13th European Simulation Multiconference*. Editor H. Szczerbicki (Warsaw: Society for Computer Simulation International), 1, 26–32. ISBN: 1565551710 https://www.tib.eu/en/search/id/TIBKAT%3A270516883/.

Liu, B., Trout, R. E. L., Chu, G.-H., McGarry, D., Jackson, R. W., Hamrick, J. C., et al. (2020). Discovery of Taniborbactam (VNRX-5133): A Broad-Spectrum Serine- and Metallo-β-Lactamase Inhibitor for Carbapenem-Resistant Bacterial Infections. *J. Med. Chem.* 63, 2789–2801. doi:10.1021/acs.jmedchem.9b01518

Lu, L.-N., Liu, C., and Yang, Z.-Z. (2020). Systematic Parameterization and Simulation of Boronic Acid–β-Lactamase Aqueous Solution in Developing the ABEEMσπ Polarizable Force Field. *J. Phys. Chem. A*. 124, 8614–8632. doi:10.1021/acs.jpca.0c06806

Marx, D. (2006). Proton Transfer 200 Years after von Grotthuss: Insights from Ab Initio Simulations. *ChemPhysChem* 7, 1848–1870. doi:10.1002/cphc.200600128

Melo, M. C. R., Bernardi, R. C., Rudack, T., Scheurer, M., Riplinger, C., Phillips, J. C., et al. (2018). NAMD Goes Quantum: an Integrative Suite for Hybrid Simulations. *Nat. Methods* 15, 351–354. doi:10.1038/nmeth10.1038/nmeth.4638

Molecular Operating Environment (MOE) (2019). *Chemical Computing Group ULC, 1010 Sherbrooke St. West, Suite #910*. Montreal, QC, Canada. H3A 2R7, 2019.

Mulholland, A. J. (2005). Modelling Enzyme Reaction Mechanisms, Specificity and Catalysis. *Drug Discov. Today* 10, 1393–1402. doi:10.1016/S1359-6446(05)03611-1

Pemberton, O. A., Noor, R. E., Kumar M. V., V., Sanishvili, R., Kemp, M. T., Kearns, F. L., et al. (2020). Mechanism of Proton Transfer in Class A β-lactamase Catalysis and Inhibition by Avibactam. *Proc. Natl. Acad. Sci. USA* 117, 5818–5825. doi:10.1073/pnas.1922203117

Philippon, A., Slama, P., Dény, P., and Labia, R. (2016). A Structure-Based Classification of Class A β-lactamases, a Broadly Diverse Family of Enzymes. *Clin. Microbiol. Rev.* 29, 29–57. doi:10.1128/CMR.00019-15

Phillips, J. C., Hardy, D. J., Maia, J. D. C., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., et al. (2020). Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD. *J. Chem. Phys.* 153, 044130. doi:10.1063/5.0014475

Renaud, P., Beauseigneur, A., Brecht-Forster, A., Becattini, B., Darmency, V., Kandhasamy, S., et al. (2007). Boron: A Key Element in Radical Reactions. *Pure Appl. Chem.*, 79, 223–233. doi:10.1351/pac200779020223

Saenger, W., Betzel, C., Zabel, V., Brown, G. M., Hingerty, B. E., Lesyng, B., et al. (1985). Hydrogen Bonding Patterns and Dynamics in the Hydration of Biological Macromolecules. *J. Biosci.* 8, 437–450. doi:10.1007/BF02703995

Sakti, A. W., Nishimura, Y., and Nakai, H. (2019). Recent Advances in Quantum-Mechanical Molecular Dynamics Simulations of Proton Transfer Mechanism in Various Water-Based Environments. *Wires Comput. Mol. Sci.* 10, e1419. doi:10.1002/wcms.1419

Schneider, K. D., Ortega, C. J., Renck, N. A., Bonomo, R. A., Powers, R. A., and Leonard, D. A. (2011). Structures of the Class D Carbapenemase OXA-24 from Acinetobacter Baumannii in Complex with Doripenem. *J. Mol. Biol.* 406, 583–594. doi:10.1016/j.jmb.2010.12.042

Setny, P. (2020). Conserved Internal Hydration Motifs in Protein Kinases. *Proteins* 88, 1578–1591. doi:10.1002/prot.25977

Sgrignani, J., Novati, B., Colombo, G., and Grazioso, G. (2015). Covalent Docking of Selected boron-based Serine Beta-Lactamase Inhibitors. *J. Comput. Aided Mol. Des.* 29, 441–450. doi:10.1007/s10822-015-9834-7/

Song, S., Gao, P., Sun, L., Kang, D., Kongsted, J., Poongavanam, V., et al. (2021). Recent Developments in the Medicinal Chemistry of Single boron Atom-Containing Compounds. *Acta Pharmaceutica Sinica B*. in press. doi:10.1016/j.apsb.2021.01.010

Stewart, J. J. P. (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. (2016). Stewart Computation Chemistry, *MOPAC2016, Colorado Springs*. USA: CO. HTTP://OpenMOPAC.net. doi:10.1061/9780784480250.012

Sulimov, A. V., Kutov, D. C., Katkova, E. V., and Sulimov, V. B. (2017). Combined Docking with Classical Force Field and Quantum Chemical Semiempirical Method PM7. *Adv. Bioinformatics* 2017, 7167691. Article ID 7167691. doi:10.1155/2017/7167691/

Thareja, S., Zhu, M., Ji, X., and Wang, B. (2017). Boron-based Small Molecules in Disease Detection and Treatment (2013-2016). *Heterocycl. Commun.* 23, 137–153. doi:10.1515/hc-2017-0086/

Tooke, C. L., Hinchliffe, P., Bragginton, E. C., Colenso, C. K., Hirvonen, V. H. A., Takebayashi, Y., et al. (2019). β-Lactamases and β-lactamase Inhibitors in the 21st Century. *J. Mol. Biol.* 431, 3472–3500. doi:10.1016/j.jmb.2019.04.002

Tooke, C. L., Hinchliffe, P., Krajnc, A., Mulholland, A. J., Brem, J., Schofield, C. J., et al. (2020). Cyclic Boronates as Versatile Scaffolds for KPC-2 β-lactamase Inhibition. *RSC Med. Chem.* 11, 491–496. doi:10.1039/c9md00557a

Van der Kamp, M. W., and Mulholland, A. J. (2013). Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. *Biochemistry* 52, 2708–2728. doi: doi:10.1021/bi400215w

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

Vega-Valdez, I. R., Melvin N., R., José M., S.-Q., D., F.-G. E., and Marvin A., S.-U. (2021). Docking Simulations Exhibit Bortezomib and Other Boron-containing Peptidomimetics as Potential Inhibitors of SARS-CoV-2 Main Protease. *Ccb* 14, 279–288. doi:10.2174/2212796814999201102195651

Venkatesan, A. M., Gu, Y., Santos, O. D., Abe, T., Agarwal, A., Yang, Y., et al. (2004). Structure–Activity Relationship of 6-Methylidene Penems Bearing Tricyclic Heterocycles as Broad-Spectrum β-lactamase Inhibitors: Crystallographic Structures Show Unexpected Binding of 1,4-Thiazepine Intermediates. *J. Med. Chem.* 47, 6556–6568. doi:10.1021/jm049680x

Wojciechowski, M., and Lesyng, B. (2004). Generalized Born Model: Analysis, Refinement, and Applications to Proteins. *J. Phys. Chem. B* 108, 18368–18376. doi:10.1021/jp046748b

Yan, S., Zhang, L., Cukier, R. I., and Bu, Y. (2007). Exploration on Regulating Factors for Proton Transfer along Hydrogen-Bonded Water Chains. *ChemPhysChem* 8, 944–954. doi:10.1002/cphc.200600674

Řezáč, J. (2017). Empirical Self-Consistent Correction for the Description of Hydrogen Bonds in DFTB3. *J. Chem. Theor. Comput.* 13, 4804–4817.

Řezáč, 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

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Charzewski, Krzyśko and Lesyng. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 24 August 2021
doi: 10.3389/fmolb.2021.661322

Coarse-Grained Modeling and Molecular Dynamics Simulations of Ca^{2+}-Calmodulin

*Jules Nde[1,2], Pengzhi Zhang[1], Jacob C. Ezerski[1], Wei Lu[2], Kaitlin Knapp[2], Peter G. Wolynes[2] and Margaret S. Cheung[1,2]**

[1]Department of Physics, University of Houston, Houston, TX, United States, [2]Center for Theoretical Biological Physics, Rice University, Houston, TX, United States

Calmodulin (CaM) is a calcium-binding protein that transduces signals to downstream proteins through target binding upon calcium binding in a time-dependent manner. Understanding the target binding process that tunes CaM's affinity for the calcium ions (Ca^{2+}), or vice versa, may provide insight into how Ca^{2+}-CaM selects its target binding proteins. However, modeling of Ca^{2+}-CaM in molecular simulations is challenging because of the gross structural changes in its central linker regions while the two lobes are relatively rigid due to tight binding of the Ca^{2+} to the calcium-binding loops where the loop forms a pentagonal bipyramidal coordination geometry with Ca^{2+}. This feature that underlies the reciprocal relation between Ca^{2+} binding and target binding of CaM, however, has yet to be considered in the structural modeling. Here, we presented a coarse-grained model based on the Associative memory, Water mediated, Structure, and Energy Model (AWSEM) protein force field, to investigate the salient features of CaM. Particularly, we optimized the force field of CaM and that of Ca^{2+} ions by using its coordination chemistry in the calcium-binding loops to match with experimental observations. We presented a "community model" of CaM that is capable of sampling various conformations of CaM, incorporating various calcium-binding states, and carrying the memory of binding with various targets, which sets the foundation of the reciprocal relation of target binding and Ca^{2+} binding in future studies.

Keywords: calmodulin dynamics, calcium-binding protein, conformational changes, intrinsic disorder, AWSEM, community model

OPEN ACCESS

Edited by:
Vojtech Spiwok, University of Chemistry and Technology in Prague, Czechia

Reviewed by:
Jianhan Chen, University of Massachusetts Amherst, United States
Manuel N. Melo, Universidade Nova de Lisboa, Portugal
Sunita Negi, GD Goenka University, India

****Correspondence:***
Margaret S. Cheung
mscheung@uh.edu

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *30 January 2021*
Accepted: *21 July 2021*
Published: *24 August 2021*

Citation:
Nde J, Zhang P, Ezerski JC, Lu W, Knapp K, Wolynes PG and Cheung MS (2021) Coarse-Grained Modeling and Molecular Dynamics Simulations of Ca^{2+}-Calmodulin. Front. Mol. Biosci. 8:661322. doi: 10.3389/fmolb.2021.661322

INTRODUCTION

Calmodulin (CaM) is a calcium-binding protein that is present in all eukaryotic cells (Berchtold and Villalobo, 2014; Villalobo et al., 2018; Chin and Means, 2000). CaM is composed of two globular domains separated by a central linker. The two domains (N- and C-domains) are constituted of two calcium-binding helix-loop-helix motifs each, namely, EF-hand motifs (Chattopadhyaya et al., 1992; Fernandes and Oliveira-Brett, 2017). Upon sufficient increase in Ca^{2+} concentration, Ca^{2+}-free CaM transitions to the Ca^{2+}-loaded CaM (**Figure 1**), exposing the hydrophobic target-binding surfaces of CaM to the solvent (**Figure 2**) (Barton et al., 2002; Vetter and Leclerc, 2003; Park et al., 2008; Gromiha and Gromiha, 2010; Wu et al., 2012; Fernandes and Oliveira-Brett, 2017). This transition is accompanied by large conformational changes mutually induced by the conformational changes in the CaM-binding target (CaMBT) peptides (Wang et al., 2013; Liu et al., 2017a; Archer et al., 2019;

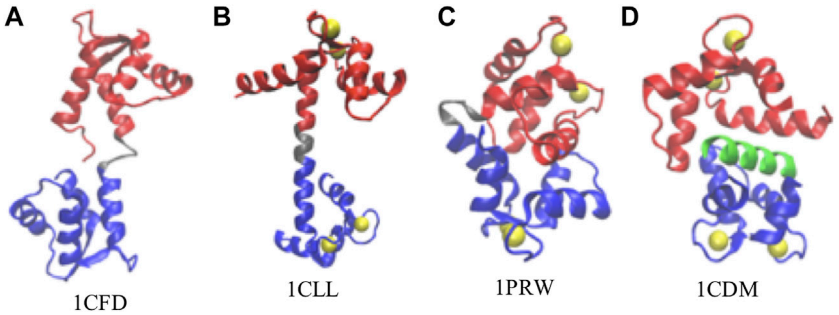


FIGURE 1 | Representative conformations of calmodulin (CaM). **(A)** Ca^{2+}-free CaM, or apoCaM (PDB ID: 1CFD); **(B)** Ca^{2+}-bound CaM, or holoCaM, in the extended conformation (PDB ID: 1CLL); **(C)** holoCaM in the collapsed conformation without a target peptide (PDB ID: 1PRW); **(D)** holoCaM with a target peptide (PDB ID 1CDM). We visualized the structures using the Visual Molecular Dynamics (VMD) program (Humphrey et al., 1996). Ca^{2+} ions are shown in yellow spheres, the N-terminal and C-terminal domains of CaM are shown in red and blue, respectively, the central linker is shown in grey, and the target peptide is in green.

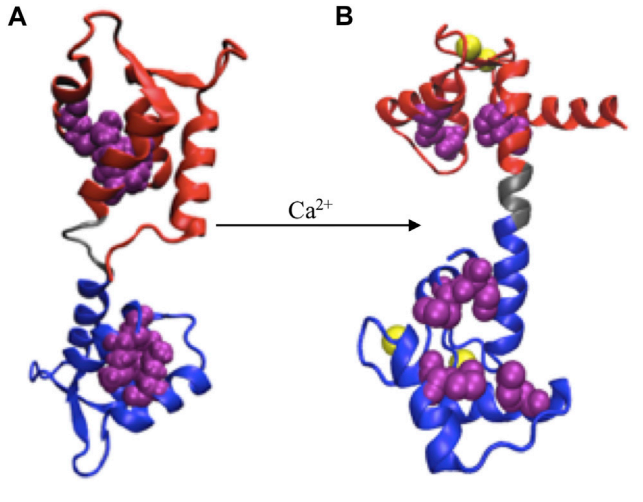


FIGURE 2 | Ca^{2+} induced conformational change. **(A)** apoCaM, **(B)** holoCaM. Ca^{2+} ions are shown in yellow spheres, the N-terminal and C-terminal domains of CaM are shown in red and blue, respectively, and the central linker is shown grey. Upon conformational change, the four methionine residues (shown as beads in purple), initially buried inside the EF-hands, are exposed to the solvent.

Bekker et al., 2020). The most unique feature of CaM is its reciprocal relation between Ca^{2+} binding and target binding (Meador et al., 1993; Weinstein and Mehler, 1994; Wall et al., 1997; Wriggers et al., 1998; Chin and Means, 2000; Hoeflich and Ikura, 2002; Westerlund and Delemotte, 2018), where the conformational changes in the CaM/CaMBT compound further tune CaM's affinity for the Ca^{2+} ions (Hoffman et al., 2014; Zhang et al., 2017). The important feature that the net charges of Ca^{2+} ions vary with the conformations of a calcium-binding loop (Zhang et al., 2021), however, has not been accurately captured in molecular dynamics simulations on calcium-binding proteins.

The 148-amino acid CaM is an evolutionarily conserved protein across all vertebrates (Davis and Thorner, 1989; Tripathi et al., 2015a), while it is capable of binding more than 300 variations of CaMBTs (Yamniuk and Vogel, 2004). Such structural variability permits its regulation upon the stimulation of calcium ions processes (Clapham, 2007; Carafoli, 2002; Bootman, 2012) in a wide range of biological activities (Chin and Means, 2000; Hoeflich and Ikura, 2002) including cellular motility, neurogenesis, memory formation, muscle contraction, and neuronal transmission (Bootman, 2012; Marambaud et al., 2009; Berridge, 2006). A large and growing body of literature has investigated the structures and dynamics of CaM upon calcium binding (Wriggers et al., 1998; Mehler et al., 1991; Her et al., 2018; Anthis et al., 2011). The crystal structure of the Ca^{2+}-CaM shows a dumbbell-like structure with a central α-helical linker (**Figure 1B**). However, in solution, the central linker demonstrates higher flexibility (Wriggers et al., 1998; Her et al., 2018; Anthis et al., 2011; Ikura et al., 1992; Chou et al., 2001), which enables the two globular domains of CaM to orient independently from each other (Wriggers et al., 1998; Shepherd and Vogel, 2004; Qin and Squier, 2001). Moreover, this flexibility of the central linker could justify the compact structure of the Ca^{2+}-CaM (**Figure 1D**) upon target binding (Barton et al., 2002; Shepherd and Vogel, 2004). Additionally, Ca^{2+}-CaM has been shown to crystallize in the collapsed conformation in the absence of the target (**Figure 1C**), indeed providing more evidence of the backbone plasticity of the Ca^{2+}-CaM (Fallon and Quiocho, 2003) and domain-domain interactions. The structural distribution between the collapsed and extended conformations is roughly 1:9 (Anthis et al., 2011).

A myriad of computational approaches, ranging from all-atom to coarse-grained (Weinstein and Mehler, 1994; Spoel et al., 1996; Wriggers et al., 1998; Vigil et al., 2001; Yang et al., 2001; Barton et al., 2002; Komeiji et al., 2002; Yang et al., 2004; Zuckerman, 2004; Monticelli et al., 2008; Nandigrami and Portman, 2016; Liu et al., 2017b; Gong and Sun, 2017; Robustelli et al., 2018; Delfino et al., 2019; Sun and Kekenes-Huskey, 2021), have been employed to model the structural dynamics of CaM. However, these models have yet to capture the reciprocal relation between Ca^{2+} binding and target binding. The modeling of CaM is often limited for a specific purpose to either bind the target or the calcium ions. As

such, CaM is treated as a folded protein (Shepherd and Vogel, 2004; Yang et al., 2004; Zuckerman, 2004) or a folded protein with intrinsically disordered regions (Robustelli et al., 2018), where, in most cases, Ca^{2+} effects are ignored for simplicity (Spoel et al., 1996; Yang et al., 2001; Shepherd and Vogel, 2004; Robustelli et al., 2018). When Ca^{2+} is considered, it is treated as a particle with fixed charges as in an aqueous solution (Monticelli et al., 2008; Gong and Sun, 2017). When Ca^{2+} ions are implicitly included, their charges are often distributed to all the negatively charged residues present in the EF-hand motifs (Nandigrami and Portman, 2016) without any consideration of the coordination geometry of the Ca^{2+} ion, which is influenced by the conformation of a calcium-binding loop (Zhang et al., 2021). Our group has recently proposed a coarse-grained model to study the dynamics of CaM that accounts for both considerations (Wang et al., 2013; Zhang et al., 2017), but it is yet to reproduce the reciprocal relationship between calcium binding and target binding as both require intensive data-driven efforts.

One of the main challenges for developing an appropriate force field involving Ca^{2+} is that CaM is a structurally flexible protein about its central linker while the binding of Ca^{2+} to the two domains of CaM increases the rigidity of the local domains. A compromise between rigidity and softness has to be considered in order for CaM to accurately select, bind, and activate its binding targets. We, therefore, turned to the coarse-grained approach that adopts the Associative memory, Water mediated, Structure, and Energy Model (AWSEM) force field (Wang et al., 2011; Zhang et al., 2017) to fulfill these requirements. The next challenge is to properly include the Ca^{2+} ions in the force field of a coarse-grained model since this would allow us to understand the reciprocal relationship between calcium binding to CaM and binding of a target (Wang et al., 2013; Tripathi et al., 2015b). Although in separate studies, there have been development in AWSEM force fields on structurally flexible proteins (Wu et al., 2018) or on di-valent ions (Tsai et al., 2016), both features, however, have yet to be combined for modeling CaM. Particularly, the coordination chemistry of Ca^{2+} ions and the structural flexibility allowing CaM to flex according to its binding targets and surrounding environment is important in modeling the adaptive behavior of CaM (Wang et al., 2013; Liu et al., 2017a; Archer et al., 2019; Bekker et al., 2020). By taking into account the coordination chemistry of the Ca^{2+} ions as well as the flexible nature of the central linker of CaM, our proposed model for Ca^{2+}-CaM elucidates the many-body effects of calcium binding on the geometrical shape of the calcium-binding loop and allows large conformational changes for binding various targets.

MATERIALS AND METHODS

Coarse-Grained Modeling Using the AWSEM Force Field

We used the AWSEM force field (Davtyan et al., 2012; Tsai et al., 2016) to build the model of Ca^{2+}-CaM. AWSEM is a transferable coarse-grained protein model, which uses three beads (C_α, C_β, and O atom from the peptide bond) to represent each amino acid residue (except for glycine, which lacks C_β). The positions of all the rest of the other atoms along the backbone are deduced by assuming an ideal geometry of the system. The total energy function used in AWSEM is given in **Eq. 1**,

$$V = V_{BB} + V_{PMF} + V_{DH} + V_{FM} \tag{1}$$

which is composed of two main terms: the physics-based terms, $V_{BB} + V_{PMF} + V_{DH}$, and the knowledge-based term, V_{FM}.

(a) The backbone term, V_{BB}, from **Eq. 1**, maintains the backbone geometry of protein. Its expression is given in **Eq. 2**,

$$V_{BB} = V_{con} + V_{chain} + V_{\chi} + V_{rama} + V_{excl}. \tag{2}$$

V_{con} and V_{chain} represent the connectivity and chain terms, respectively. The former links neighboring residues, while the latter maintains ideal bond angles around the C_α atoms of each residue. The chirality potential, V_χ, which is applied to all residues except for glycine, maintains the local residue chirality. V_{rama} represents the Ramachandran potential, which biases the protein chain conformation toward the allowed Ramachandran regions. The excluded volume potential, V_{excl}, is used to avoid chain collapse and unphysical entanglements.

(b) The many-body potential of mean force V_{PMF}, from **Eq. 1**, takes into account the nature of the interacting residues, and its expression is given in **Eq. 3**:

$$V_{\mathrm{PMF}} = V_{\mathrm{contact}} + V_{\mathrm{burial}} + V_{\mathrm{HB}} \tag{3}$$

The contact term V_{contact} represents the pairwise additive and the many-body terms that consider interactions between residues far apart in sequence. The burial term V_{burial} accounts for a particular residue's type propensity. The hydrogen bond term V_{HB} is composed of three other terms given in **Eq. 4**:

$$V_{HB} = V_{\beta} + V_{P-AP} + V_{helical}, \tag{4}$$

where V_β and V_{P-AP} are for the hydrogen bonding interactions in β-sheet conformations. The former stabilizes already formed β hydrogen bonds, while the latter allows a protein to undergo parallel and anti-parallel β-sheet conformations before the stabilization of the hydrogen bond. The helical term $V_{helical}$ ensures the formation of α-helices.

(c) The electrostatic interactions in solution with implicit solvent are described by the Debye–Hückel term (Tsai et al., 2016) V_{DH} from **Eq. 1**. The expression is given in **Eq. 5**:

$$V_{DH} = K_{Elec} \sum_{i<j} \frac{q_i q_j}{\varepsilon_r r_{ij}} \exp\left(-\frac{r_{ij}}{l_D} \right), \tag{5}$$

where q_i and q_j represent the charges of beads i and j; r_{ij} is the distance between the two beads; $K_{Elec} = (4\pi\varepsilon_0)^{-1} = 332.24\ kcal\ mol^{-1} e^{-2}$Å; ε_r represents the dielectric constant of the media; and l_D is the Debye–Hückel screening length, which is given by $l_D = \sqrt{\varepsilon_r \varepsilon_0 k_B T / 2e^2 I}$. k_B is the Boltzmann constant, T is the temperature, e represents the elementary electric charge, and I is the ionic strength of the implicit solvent.

TABLE 1 | The PDB IDs of the 60 non-redundant CaM memory structures for the AWSEM model and the Rg of the CaM. In the case of CaM complexes, only the coordinates of CaM were used for memory. *m* is the index for the memories.

m	PDB ID	Rg (Å)	*m*	PDB ID	Rg (Å)	*M*	PDB ID	Rg (Å)
1	1PRW	14.60	21	3GP2	15.96	41	2BBM	17.03
2	3EWT	15.17	22	3DVM	16.00	42	2BCX	17.25
3	2LGF	15.23	23	1CKK	16.05	43	2KNE	18.67
4	3DVJ	15.31	24	2FOT	16.05	44	4DCK	18.87
5	3BYA	15.38	25	2JZI	16.12	45	1SK6	19.30
6	2LHI	15.38	26	1CM1	16.20	46	1XFU	19.31
7	2HQW	15.41	27	3DVK	16.22	47	2L1W	19.46
8	2L7L	15.43	28	2VAY	16.26	48	1K9O	19.46
9	3EWV	15.52	29	2O60	16.28	49	1PKO	19.49
10	1IWQ	15.57	30	2WEL	16.35	50	1S26	19.50
11	1L7Q	15.59	31	1QTX	16.43	51	1LVC	19.70
12	3DVE	15.61	32	2O5G	16.47	52	1CFF	19.78
13	1CDM	15.66	33	1QS7	16.50	53	4G27	20.79
14	3SUI	15.68	34	3GOF	16.50	54	2L53	21.00
15	3BXL	15.80	35	3BXK	16.51	55	2KDU	25.65
16	1IQ5	15.90	36	1MXE	16.55	56	4EHQ	22.07
17	1ZUZ	15.93	37	2Y4V	16.62	57	4DJC	22.31
18	3HR4	15.94	38	2BE6	16.65	58	2YGG	23.66
19	2KOF	15.95	39	2F3Y	16.68	59	1G4Y	25.32
20	2LL6	15.95	40	1SY9	16.83	60	1CLL	21.80

(d) V_{FM} from **Eq. 1**, is a bioinformatic fragment memory potential that structurally biases short fragments of the protein chain, typically composed of 3–9 residues at a time, towards conformations that are based on "memory" structures. As listed in **Table 1**, the memory structures selected from the Protein Data Bank (PDB) were used to speed up the conformational search of the native states. Its expression is given in **Eq. 6**:

$$V_{FM} = -\lambda_{FM} \sum_{m=1}^{60} \cdot \sum_{n=1}^{3} W_n^m \sum_{i,j \ni 3 \le |i-j| \le 9} \gamma_{ij} exp\left[-\frac{\left(r_{ij} - r_{ij}^m\right)^2}{2\sigma_{ij}^2} \right], \quad (6)$$

where the outer sum is over the aligned memory (from $m = 1$ *to* 60), the sum in the middle with a variable n is over the three segments of CaM ($n = 1$ is N-domain, $n = 2$ is central linker, and $n = 3$ is C-domain), and the inner sum is over all possible pairs of C_α and C_β atoms within the memory fragments that are separated by two or more residues; $|i - j| \le 9$ gives the maximum sequence separation of interacting residues, which is either the length of the memory or the maximum cutoff, whichever is shorter. r_{ij} and r_{ij}^m represent the instantaneous distance and the corresponding distance in the memory fragment between the atoms i and j; γ_{ij} is the residue type dependence interaction strength; σ_{ij} represents the sequence separation dependent width. Its expression is given by $\sigma_{ij} = |i - j|^{0.15}$.

We used 60 non-redundant structures of CaM taken from our prior study (Tripathi et al., 2015c) (**Table 1**), either free or in complex with the target peptides, to build the memory fragments for the coarse-grained model of CaM. We used the 60 sequence homologs of Ca^{2+}-CaM since we are interested in studying the dynamics and the folding properties of CaM. We divided each memory into three segments: N-domain of CaM, from residue 5 to residue 76; the flexible central linker, from residue 77 through residue 81; and C-domain of CaM, from residue 82 to residue 147 (as illustrated in **Figure 1**). The weight of each segment is controlled by a memory weight parameter W_n^m and a global parameter λ_{FM} (**Eq. 6**). Because the central linker has its own memories of large variations in the structures, essentially it was modeled with more structural flexibility than the two globular domains of CaM.

We summarized CaM models using various values of λ_{FM} and W_n^m parameters in **Table 2** for our molecular simulations. All other parameters for AWSEM were set by default as defined in the original ASWEM study (Davtyan et al., 2012; Tsai et al., 2016). The parameters of the default setting are also provided in the Supplemental Information (SI 1).

Modeling Ca^{2+} Ions in ASWEM Force Field

Calcium-binding proteins in most cases contain the EF-hand calcium-binding motif which is formed by the pentagonal bipyramidal geometry (**Figure 3**; **Supplementary Figure S1** in the SI) (Drake et al., 1997; Yang et al., 2002). The literature on modeling those proteins, in general, does not clearly explain how Ca^{2+}-binding effects are considered during the simulations. We developed an implicit Ca^{2+} model that evenly splits the Ca^{2+} charges to a selection of residues. Partial charges on the sidechain beads of the negatively charged residues that coordinate Ca^{2+} collectively in the Ca^{2+} binding loops (**Figure 3**) were adjusted to reflect the Ca^{2+}- binding effect (**Table 3**). To note, charges on those acidic residues which do not coordinate (Ca^{2+}) remain intact (−1 e). This simple approach incorporates the coordination geometry of the Ca^{2+} and allows us to improve the modeling of many-body effects in the local Ca^{2+} binding loops which furthermore controls the open/close state of the helix-loop-helix EF-hand and the global conformation of calmodulin.

In this study, we compared the abovementioned approach (approach III) with two other calcium models:

I. Splitting the Ca^{2+} charges evenly to the neighboring negatively charged residues (Tsai et al., 2016).

II. Splitting the Ca^{2+} charges evenly to the residues that participate with the calcium coordination and the radius of gyration of the sidechain beads are constrained by a stiff harmonic potential with a force constant of 30 kcal/mol/Å^2.

III. The same with approach II except that the harmonic constraint is removed.

Comparing to approach I of Ca^{2+} in AWSEM (Tsai et al., 2016), in approach III calcium charges were split according to the coordination chemistry of calcium; comparing to approach II, in approach III more degrees of freedom to the Ca^{2+} coordination residues allow sampling of more different conformations of CaM.

Simulation Details

We performed the coarse-grained simulations with the open-source MD package LAMMPS, in which AWSEM code was implemented (Davtyan et al., 2012). We used the periodic boundary conditions on the cubic box of 400 Å on each side so that even unfolded CaM can fit in the box. Initial velocities were chosen randomly from a Boltzmann distribution with the average squared velocity equal to 3 k_BT/m, where k_B is the

TABLE 2 | Summary of Ca^{2+}-CaM models with different values of the global and local memory weight and the average Rg ($\overline{Rg}$) and the ratio between the collapsed and extended states ($\gamma_{c:e}$) from the umbrella sampling simulations. The models that best reproduce experimental results from our simulations are marked with *. The experimental measured Rg = 20.0 ~ 22.5 Å and $\gamma_{c:e}$ ~ 0.11.

Models	λ_{FM}	W_n^m				$\overline{Rg}$ (Å)	$\gamma_{c:e}$
		W_n^{others}	W_1^{1CLL}	W_2^{1CLL}	W_3^{1CLL}		
I.1	1	1	1	1	1	22.05	0.001
I.2	0.1	1	1	1	1	20.51	0.48
I.3	0.01	1	1	1	1	19.86	0.64
II.2	0.1	1	2	2	2	20.34	0.59
II.3	0.1	1	3	3	3	20.66	0.47
II.4	0.1	1	4	4	4	20.43	0.50
II.5	0.1	1	5	5	5	20.70	0.37
II.6	0.1	1	6	6	6	20.52	0.38
II.7	0.1	1	7	7	7	21.11	0.16
*II.8	0.1	1	8	8	8	21.15	0.14
*II.9	0.1	1	9	9	9	21.26	0.10
III.1	0.1	1	5	1	5	19.91	0.73
III.2	0.1	1	5	2	5	19.70	0.82
III.3	0.1	1	5	3	5	19.84	0.77
III.4	0.1	1	5	4	5	20.03	0.58
III.6	0.1	1	5	6	5	20.43	0.48
III.7	0.1	1	5	7	5	21.20	0.16
*III.8	0.1	1	5	8	5	21.35	0.12
III.9	0.1	1	5	9	5	21.59	0.07

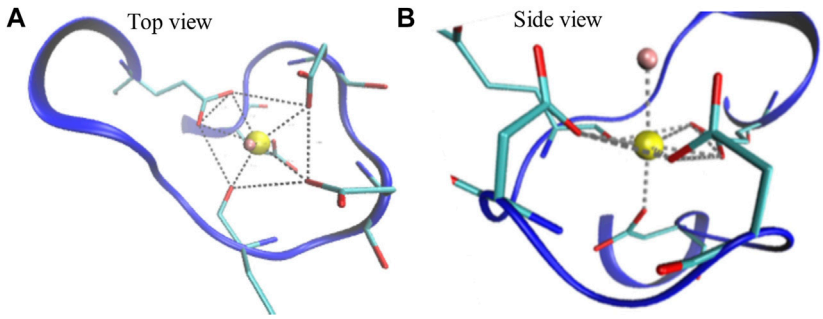


FIGURE 3 | Illustration of a Ca^{2+} binding loop at the fourth EF-hand of CaM (PDB ID: 1CLL). The pentagonal bipyramidal geometry of the Ca^{2+} coordination is represented with the dotted lines in the top view (**A**) and side view (**B**). The red sticks linked with the dotted lines represent oxygen atoms that participate in coordination. The water molecule, shown in pink, completes the spherical coordination through hydrogen bonding.

TABLE 3 | The adjusted partial charges of the negatively charged residues participated in calcium coordination in the four calcium-binding loops of CaM. The amino acid sequences of the calcium-binding loops are provided as a reference. The subscript indices stand for the positions of the starting/ending residues in the CaM sequence. Underscored residues are the selected amino acids with negative charges as shown in the third column. We adjusted the partial charges on the sidechain beads of those residues based on the bipyramidal pentagonal coordination geometry of the calcium ion.

Calcium loop	Amino acid sequence	Adjusted partial charges on selected amino acids
Loop 1	$_{20}$D̲K̲D̲G̲D̲G̲TITTK̲E̲$_{31}$	-0.5, -0.5, -0.5, -0.5
Loop 2	$_{56}$D̲A̲D̲G̲N̲G̲TID̲F̲P̲E̲$_{67}$	-0.5, -0.5, -0.5, -0.5
Loop 3	$_{93}$D̲K̲D̲G̲N̲G̲YISAAE̲$_{104}$	-0.33, -0.33, -0.33
Loop 4	$_{129}$D̲I̲D̲G̲D̲G̲QVNYEE̲$_{140}$	-0.5, -0.5, -0.5, -0.5

Boltzmann constant, m is the mass, and the temperature is T. We also used the cutoff distance of 3.5 Å and κ (inverse of the Debye length) = 0.0127 $Å^{-1}$ in the electrostatic term.

Simulated annealing: firstly, we denatured the protein at 450 K using the initial structure built from the crystal structure (PDB ID: 1CLL) (Chattopadhyaya et al., 1992); then, we performed simulated annealing on Ca^{2+}-CaM by cooling down the system to a low temperature (280 K) for 2,000,000 time steps for each value of λ_{FM} (models I.1–3 in **Table 2**).

Production simulation: we employed the umbrella sampling (US) method (Roux, 1995) to evaluate the thermodynamic

properties of Ca^{2+}-CaM using different memory parameters. We conducted the production simulations in canonical ensemble (NVT) at the constant temperature of 300 K. Radius of gyration (Rg) of Ca^{2+}-CaM was used as the reaction coordinate, which was restrained by a harmonic potential $E^{i}_{Restr} = \frac{1}{2}k(Rg^{i} - Rg^{i}_{0})^{2}$ with a force constant k = 50 kcal/Å^2. The equilibrium positions of the harmonic potential Rg_0 range from 13.00 Å to 26.00 Å with a bin size of 0.25 Å, making up a total of 53 windows (i = 0, 1, 2, . . ., 52). The setup of the force constant and the equilibrium positions of the windows is justified by the fluctuations of Rg around the corresponding equilibrium positions as well as the overlap in the Rg distribution between neighboring windows, which are provided in **Supplementary Figures S4–S5**, respectively, in the SI. For each window, 2,000,000 time-steps of constrained molecular simulations were conducted using different initial conditions. To generate the initial structures at each window for the US simulations, molecular dynamics simulations were carried out for the Ca^{2+}-CaM from the crystal structure of CaM (PDB ID: 1CLL, in the coarse-grained model) at a temperature T = 300 K. Structures with Rg closest to Rg^{i}_{0} were selected as the initial configurations. A set of 5 US simulations were carried out at each window with the same initial configuration and different random initial velocities. The integration time step is 2 fs. It is important to note that with the smoothed CG potential, the realistic time represented by a time step is longer than 2 fs. Coordinates and Rg were recorded every 1,000 time steps for analyses, making up 10, 000 frames of data for analysis at each window. We analyzed the data using a multi-state Bennett acceptance ratio (MBAR) estimator with the *pymbar* program (Shirts and Chodera, 2008).

Data Analysis

We described the conformational changes of the Ca^{2+}-CaM during the simulations by calculating the radius of gyration (Rg) and the pairwise comparison (Q_w) from the crystal structure (PDB ID: 1CLL).

Radius of Gyration (Rg)

Rg describes the compactness of the protein, and its definition is given in **Eq. 7**:

$$Rg = \sqrt{\frac{\sum_{i=1}^{N} m_i \Delta r_i^2}{\sum_{i=1}^{N} m_i}}. \quad (7)$$

In the case of PDB structures, m_i is the mass of the i^{th} atom and Δr_i is the distance between the i^{th} atom and the center of mass of the system. In the case of coarse-grained structures, only the C_α beads are used, m_i is the mass of the C_α bead of the i^{th} residue, and Δr_i is the distance between the C_α bead of the i^{th} residue and the center of mass of the system.

Pairwise Comparison

Q_w measures the degree of similarity of conformations from the trajectories with respect to the native structure through the pairwise comparison. In other words, it compares the pairwise distances of C_α atoms (beads) among the residues in each instantaneous structure to counterpart in the native structure. This value is normalized to 1, with a higher value corresponding to the greater similarity to the native structure. In this case, Q_w is the extended crystal structure of Ca^{2+}-CaM (PDB ID: 1CLL), Q_w = 1. The expression of the order parameter Q_w is given in **Eq. 8**:

$$Q_w = \frac{2}{(N-2)(N-3)} \sum_{|j-i| \geq 2} exp\left[-\frac{\left(r_{ij} - r_{ij}^{m}\right)^2}{2\sigma_{ij}^2} \right], \quad (8)$$

where N is the total number of residues, r_{ij} represents the instantaneous distance between C_α atoms (beads) of residues i and j, r_{ij}^{m} is the same distance in the reference structure (PDB ID: 1CLL), and σ_{ij} is obtained from the following relation $\sigma_{ij} = (1 + |i-j|)^{0.15}$.

Definition of Collapsed and Extended Conformations of CaM

In our analyses, we used the parameter Rg to define a collapsed or extended conformation. The Rg cutoff was set according to the potential of mean force (PMF) and distribution of Rg in our simulations (**Figures 4–8**). A collapsed conformation is defined if Rg < 18.5Å; otherwise, an extended conformation is defined.

RESULTS AND DISCUSSION

The goal is to create a computational model for CaM accommodating its several key functions including divalent ion binding, conformational dynamics, and target recognition. In order to achieve this, we adopted a pool of 60 CaM or CaM complex structures from our previous study (Tripathi et al., 2015c) representing full-length CaM without mutations and its complex with 24 unique target proteins/peptides. As stated in the *Materials and Methods Simulation Details*, our approach is to tune the two scaling parameters λ_{FM} and W_n^m to adjust the global and each segment of the multiple memories, respectively, relative to other terms in the total energy function. After optimizing the memory parameters, three models of calcium are compared to describe the divalent ion binding.

Each of the 60 memories has a specific conformation of calmodulin which could be helpful in target recognition and target selection; therefore, clustering the 60 memories may leave out important conformations that accommodate recognition of a specific target protein. Potentially, there could be as many as 181 fragment memory parameters to be determined in the most complicated scenario. However, in the general practice of AWSEM modeling, a search in the available determined structures in the database according to the match in the desired amino acid sequence is conducted to generate a library of memory structures. Therefore, to avoid excessive tuning of the parameters, the individual memory weight W_n^m is kept as a constant default value. Because the unbound extended form of the Ca^{2+}/CaM (PDB code: 1CLL) was found to be most dominant experimentally, we adjusted its local memory weight parameter correspondingly, while keeping others the default $W_n^{others} = 1$. The actual parameters we tuned eventually reduce to λ_{FM} and W_n^{1CLL}(n = 1, 2, 3 to represent N-domain, central linker, and C-domain of the protein).

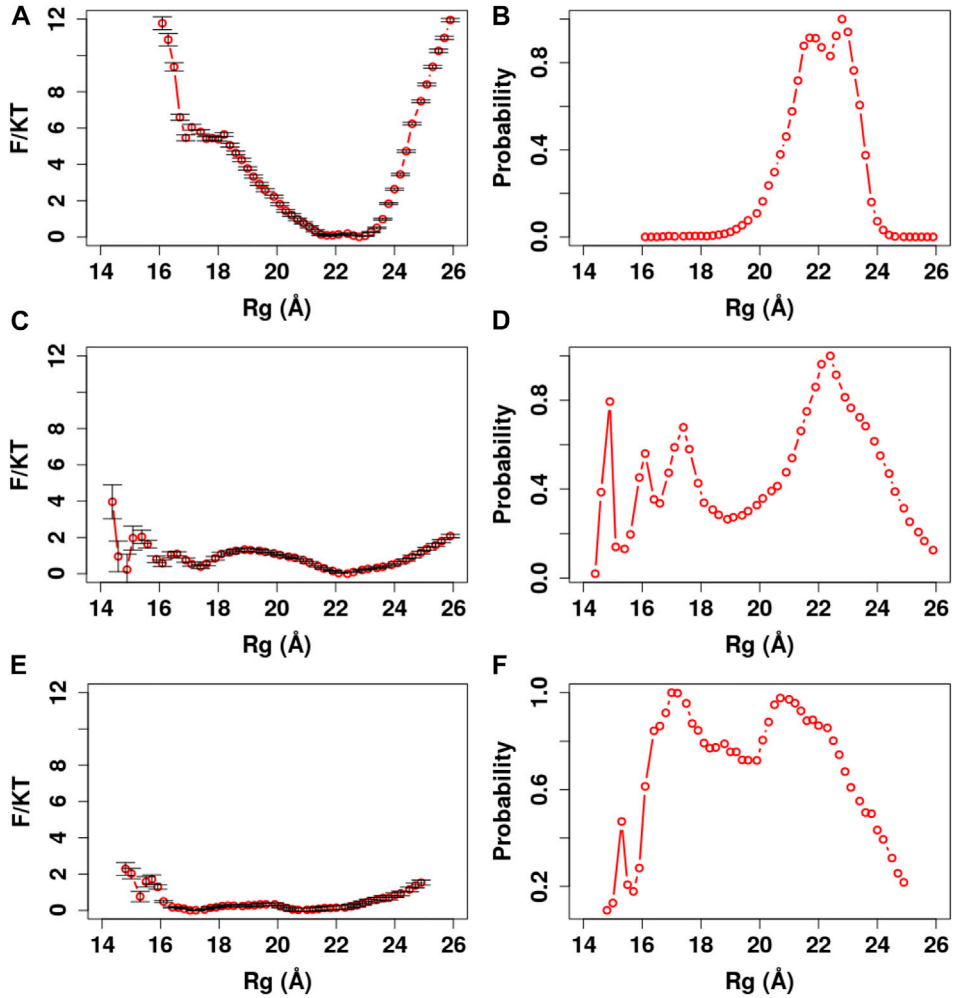


FIGURE 4 | PMF and Rg distribution of the Ca^{2+}-CaM models using the global memory weight of different order of magnitudes. The memory weight parameters of the Ca^{2+}-CaM models are listed in **Table 2** (model I; λ_{FM} = 1, 0.1, 0.01, and W_n^m = 1). **(A, C, E)** The reweighted free energy profile as a function of Rg from US simulations with λ_{FM} = 1, 0.1, 0.01, respectively. (B, D, F) The probability distribution as a function of Rg. The standard errors in the PMF are provided.

Parameterization of our CG model was guided by the existing following experimental measures: Rg of Ca^{2+}-CaM spans from 20.0 to 22.5 Å measured by small angle X-ray scattering (SAXS) experiments (Seaton et al., 1985; Heidorn and Trewhella, 1988; Kataoka et al., 1991); in addition, the population of CaM's collapsed conformation is about 10% measured by nuclear magnetic resonance (NMR) experiments (Anthis et al., 2011) (i.e., the ratio between the population of the collapsed conformation and that of the extended population is $\gamma_{c:e} \sim 0.11$).

The Global Parameter Weight Controls the Energy Barrier Between Extended and the Collapsed States of Ca^{2+}-CaM

Firstly, we determined the magnitude of the fragment memory interaction by investigating its effect on the folding properties of Ca^{2+}-CaM. We carried out simulated annealing using the unfolded structure of the Ca^{2+}-CaM ($Q_w \sim 0.2$ as shown in **Supplementary Figure S2** in the SI) with three orders of magnitudes of the global scaling parameter λ_{FM} = 1, 0.1, 0.01. At λ_{FM} = 1, the system transitions abruptly from the unfolded to the folded state (**Supplementary Figures S3A, S3B**) in the SI; at λ_{FM} = 0.1, smooth transitions of Ca^{2+}-CaM from the unfolded to the folded structure were observed with Q_w reaching a maximum value of 0.67 (**Supplementary Figures S3C, S3D**) in the SI; at λ_{FM} = 0.01, the maximum value of Q_w was less than 0.55 and the total energy of the system did not stabilize on a long timescale (**Supplementary Figures S3E, S3F**) in the SI.

In addition, we carried out US simulations to evaluate thermodynamic properties of Ca^{2+}-CaM with the global scaling parameter in three orders of magnitudes λ_{FM} = 1, 0.1, and 0.01, while keeping the local memory scaling parameters W_n^m = 1. The PMF and the probability distribution as a function of Rg were computed using the MBAR estimator (**Figure 4**) and the average Rg ($\overline{Rg}$) and the ratio between the population of collapsed conformation and that of the extended conformation ($\gamma_{c:e}$) were

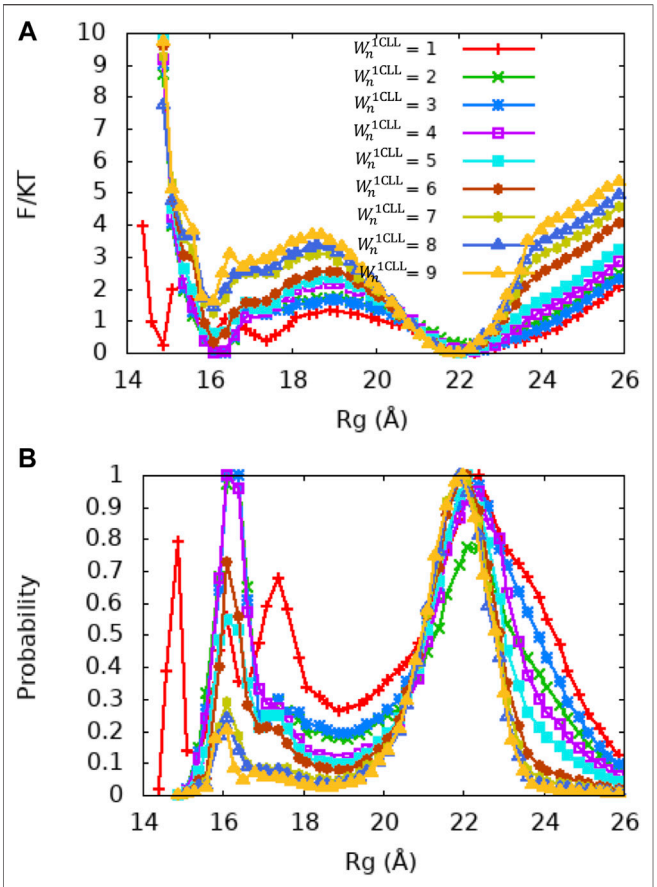


FIGURE 5 | PMF and Rg distribution of the Ca^{2+}-CaM models for different values of the local memory weight W_n^{1CLL} (n = 1, 2, 3). The memory weight parameters of the Ca^{2+}-CaM models are listed in **Table 2** (model II; λ_{FM} = 0.1, W_n^{others} = 1, and W_n^{1CLL} = 1 – 9). **(A)** The reweighted free energy profile as a function of Rg from US simulations **(B)** and the probability distribution as a function of Rg. The standard errors are too small to show and are provided in Figures S6 in the SI.

evaluated from the PMF. For λ_{FM} = 1, Ca^{2+}-CaM mostly samples the extended conformations (with Rg ~ 21.5 to 22.5 Å) and the free energy barrier between the extended and collapsed conformations is quite large (~6 k_BT); therefore, the system lacks the capacity of shifting between multiple conformations. For λ_{FM} = 0.1, there are four minima in the PMF of Ca^{2+}-CaM with Rg ranging from 15 to 22 Å (**Figure 4**), and the free energy barrier between the collapsed states and the extended state is ~ k_BT; the $\overline{Rg}$ of Ca^{2+}-CaM is within the experimental range ($\overline{Rg}$ = 20.51 Å); however, $\gamma_{c\text{-}e}$ (~ 0.48) is much higher than the experimental one (~ 0.11) (Anthis et al., 2011). It is worthwhile to note that the minimum at Rg ~ 15 Å resembles a completely collapsed structure as found in our previous study (Homouz et al., 2009). For λ_{FM} = 0.01, there are noticeably two minima located at Rg ~ 17 Å and Rg ~ 20.5 Å, sampling a somewhat collapsed conformation and a less extended conformation (comparing with λ_{FM} = 1, 0.1) respectively with free energy barrier < k_BT. $\overline{Rg}$ = 19.86 Å, below the experimentally measured Rg range (Seaton et al., 1985; Heidorn and Trewhella, 1988; Kataoka et al., 1991) and $\gamma_{c\text{-}e}$ (~ 0.64) is much higher than the experimental value (Anthis et al., 2011).

In summary, the global memory parameter λ_{FM} is effective for controlling the sampling of the collapsed and extended conformations of CaM as well as reproducing the averaged Rg. A large value of the scaling parameter λ_{FM} might tend to limit the malleability of Ca^{2+}-CaM in favor of the extended conformation (initial configuration of the simulations). Decreasing λ_{FM} leads to an increase in $\gamma_{c\text{-}e}$ and a decrease in $\overline{Rg}$. This is because decreasing the weight of the memory potential favors the hydrophobic interactions among residues of the two domains of Ca^{2+}-CaM and reduces the free energy barrier between the two major conformations of Ca^{2+}-CaM (extended and collapsed forms); consequently, the system becomes more flexible and can shift easily between the extended and collapsed conformations.

Single Memory Optimization Captures the Dynamics of Ca^{2+}-CaM

Although the global fragment memory weight of λ_{FM} = 0.1 resulted in a system that can sample the major two conformations of the Ca^{2+}-CaM and $\overline{Rg}$ is within the experimental range (Seaton et al., 1985; Heidorn and Trewhella, 1988; Kataoka et al., 1991), the ratio between

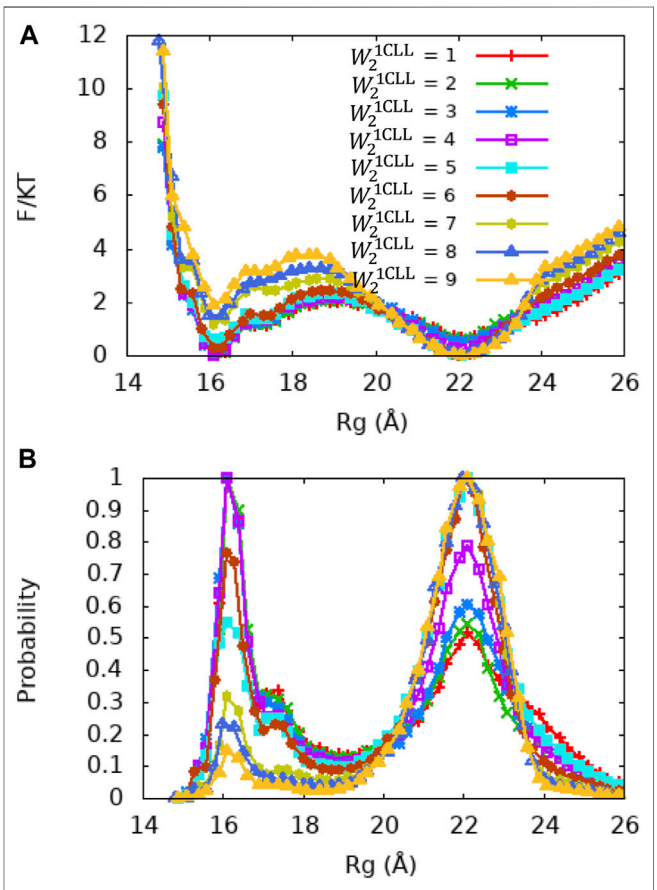


FIGURE 6 | PMF and Rg distribution of the Ca^{2+}-CaM models for different values of the local memory weight W_n^{1CLL} (n = 2) for the central linker of CaM. The memory weight parameters of the Ca^{2+}-CaM models are listed in **Table 2** model III; λ_{FM} = 0.1, W_n^{others} = 1, $W_{1,3}^{1CLL}$ = 5, W_2^{1CLL} = 1 – 9. **(A)** The reweighted free energy profile as a function of Rg from US simulations **(B)** and the probability distribution as a function of Rg. The standard errors in the PMF are provided in Figure S7 in the SI.

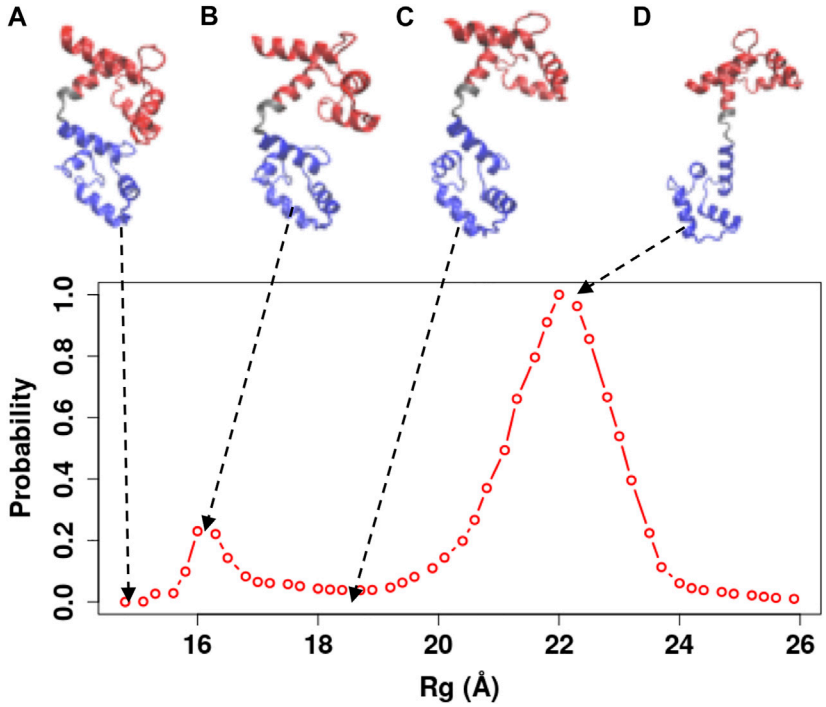


FIGURE 7 | Representative snapshots of the Ca^{2+}-CaM model obtained from our simulations. The model that best reproduces experimental measurements was used (λ_{FM} = 0.1, W_n^{others} = 1, $W_{1,3}^{1CLL}$ = 5, and W_2^{1CLL} = 8). **(A)** The completely collapsed conformation of the Ca^{2+}-CaM; **(B)** the collapsed conformation; **(C)** the intermediate conformation; and **(D)** the extended conformation. These structures are represented along the probability distribution of Rg.

the collapsed and the extended populations $\gamma_{c:e}$ did not match the experimentally measured value (Anthis et al., 2011). Therefore, finer tuning of the local memory parameters W_n^m is required for reproducing $\gamma_{c:e}$ in the abovementioned NMR experiment (Anthis et al., 2011).

Since the crystal structure of Ca^{2+}-CaM (PDB ID: 1CLL) demonstrates a value of Rg that is closest to the experimentally measured values and does not bind a target, we first examined whether variation in single memory weight of this structure (W_n^{1CLL}) could give rise to the conformational changes of the Ca^{2+}-CaM. We performed a series of molecular simulations using US with W_n^{1CLL} = 1, 2, . . . , 9 and maintaining λ_{FM} = 0.1 and W_n^{others} = 1 (**Table 2** Model II). From the PMF profiles and the probability distributions shown in **Figure 5**, we show that the Ca^{2+}-CaM samples both the extended and collapsed conformations in all the models we explored. In general, with the increase of W_n^{1CLL}, stability of the collapsed state decreases and the barrier between the extended and collapsed states increases, hence an increase in $\overline{Rg}$ and decrease in $\gamma_{c:e}$. Specifically, without tuning any other local memory weights, W_n^{1CLL} = 8, 9, and $\overline{Rg}$ and $\gamma_{c:e}$ match with experimental values (Seaton et al., 1985; Heidorn and Trewhella, 1988; Kataoka et al., 1991; Anthis et al., 2011) quite well (**Table 2**).

Central Linker of the Ca^{2+}-CaM Determines the Conformational Change of Ca^{2+}-CaM

Although the X-ray crystallography experiment shows that the central linker that connects the two domains of Ca^{2+}-CaM has a helical structure, numerous studies including NMR have shown its high flexibility in solution. To explore the relationship between the flexible linker and the dynamics of Ca^{2+}-CaM in more detail, we carried out a series of US molecular simulations with the memory weight of the central linker of the Ca^{2+}-CaM, W_2^{1CLL} = 1, 2, . . . , 9 while keeping λ_{FM} = 0.1 and W_n^{others} = 1, and W_n^{1CLL}(n = 1,3) = 5 (**Table 2** Model III). As shown in the PMF profiles (**Figure 6**), W_2^{1CLL} shows a similar effect on the thermodynamic properties of the Ca^{2+}-CaM; with the increase of W_2^{1CLL}, stability of the collapsed state decreases and the barrier between the extended and collapsed states increases, hence an increase in $\overline{Rg}$ and decrease in $\gamma_{c:e}$ (**Table 2**). This is likely because the two globular domains of CaM consist of stable secondary structures (mostly α-helices) and the involving fragments are fully represented and heavily weighted by W_n^{others} (n = 1, 3) in other 59 memories, whereas the central linker connecting the two domains of CaM is modeled as an intrinsically disordered peptide; hence, the stability of the Ca^{2+}-CaM is more sensitive to the parameter W_2^{1CLL} corresponding to the central linker of CaM. Therefore, tuning W_n^{1CLL} for the whole CaM is mostly equivalent to tuning W_2^{1CLL} for the central linker. This can also be seen in the three highlighted models in **Table 2**, which reproduce the experimentally measure Rg and $\gamma_{c:e}$ well (Seaton et al., 1985; Heidorn and Trewhella, 1988; Kataoka et al., 1991; Anthis et al., 2011).

In order to illustrate the conformational changes of Ca^{2+}-CaM in solution, we selected representative snapshots from the simulations using the Ca^{2+}-CaM model with λ_{FM} = 0.1, W_n^{others}

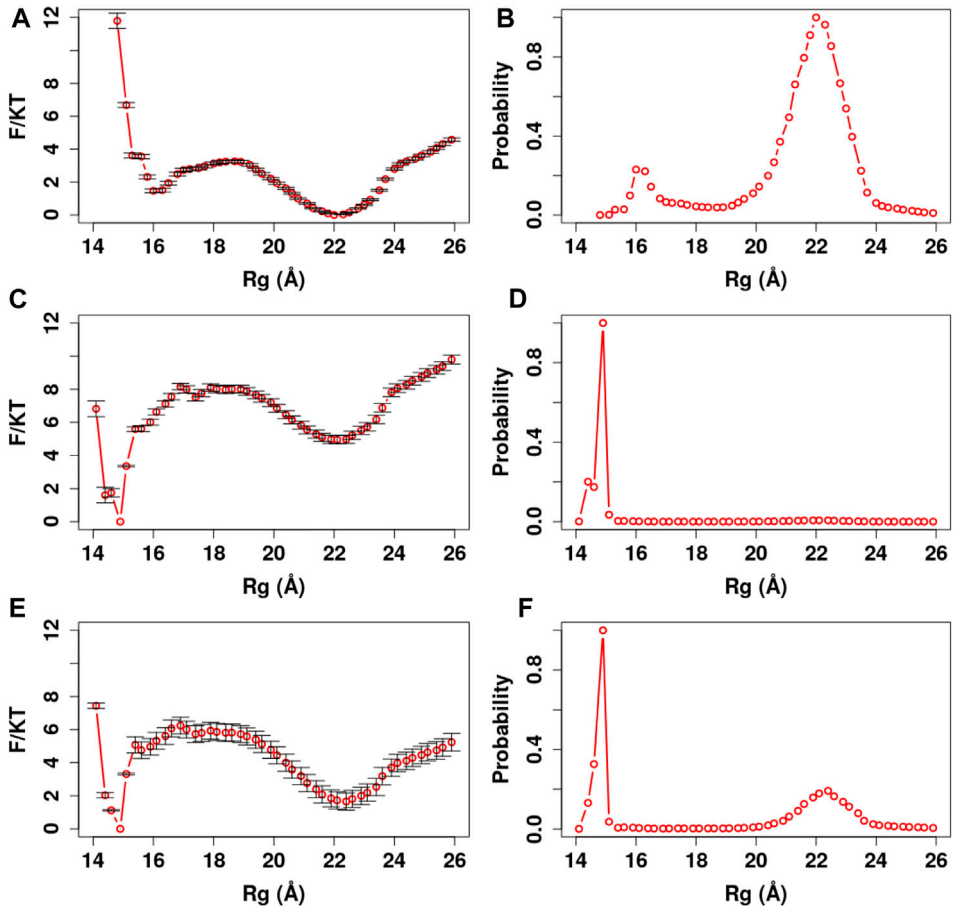


FIGURE 8 | PMF and Rg distribution of the Ca^{2+}-CaM for different Ca^{2+} models. The optimized memory parameters for the Ca^{2+}-CaM model were used (λ_{FM} = 0.1, W_n^{others} = 1, $W_{1,3}^{1CLL}$ = 5, and W_2^{1CLL} = 8). The three charge models are described in the *Method* section. **(A)** The reweighted free energy profile as a function of Rg from the US simulations **(B)** and the probability distribution as a function of Rg. The standard errors in the PMF are provided.

= 1, $W_{1,3}^{1CLL}$ = 5, and W_2^{1CLL} = 8, as highlighted in **Table 2**. The unstructured central linker of the extended structure (**Figure 7D**) unwraps and bends to allow the two domains of the Ca^{2+}-CaM to interact with each other, as shown in **Figure 7A–C**). **Figure 7A** represents a completely collapsed structure similar to the compact structure (PDB ID: 1PRW) or in CaM-target complexes, however, the chance for Ca^{2+}-CaM to sample this structure is low. **Figure 7B** represents a collapsed structure. **Figure 7C** represents a structure in the transition state between the collapsed and extended states.

Coordination Geometry Is Necessary for Representing Divalent Ions in Coarse-Grained Models

Although it can be tempting to use the available intrinsically disordered protein (IDP) force fields such as AWSEM-IDP (Wu et al., 2018) (due to the disordered nature of the central linker) to study the dynamics of Ca^{2+}-CaM, the heterogeneity of CaM needs to be considered. The high rigidity of the two domains upon calcium-binding limits the use of IDP force fields, thus making the computational investigation of CaM even more complex.

We further manifest the importance of coordination chemistry by comparing the three approaches to modeling Ca^{2+} as described in Method *Modeling Ca^{2+} Ions in ASWEM Force Field* in **Figure 8**. With approach I, the Ca^{2+} charges were evenly distributed to all the negatively charged residues (Tsai et al., 2016) and Ca^{2+}-CaM presents an equal probability of sampling the completely collapsed conformation and the extended conformation ($\gamma_{c:e}$ = 1.01). $\overline{Rg}$ ~ 18.64 Å is well below the experimental values (Seaton et al., 1985; Heidorn and Trewhella, 1988; Kataoka et al., 1991). With approach II by applying constraints on the Ca^{2+} loops, Ca^{2+}-CaM samples favorably the completely collapsed conformation with Rg ~ 15 Å ($\gamma_{c:e}$ = 22.87). Consequently, the binding selectivity of Ca^{2+}-CaM could be impeded due to the lack of the flexibility of calcium-binding loop in the system. Being confined in this completely collapsed conformation, Ca^{2+}-CaM lacks the capacity of recognizing, binding, and activating the targets. With approach III (w/coordination chemistry), Ca^{2+}-CaM samples both extended and collapsed conformations while preferring the former.

Force field development for Ca^{2+} ion is mostly limited to the aqueous solution rather than in protein-bound environments (Probst et al., 1985). One of the main challenges in computer simulations of macromolecules remains the development of force fields with the appropriate implementation of the metal ion (Gifford et al., 2007; Duarte et al., 2014; Li and Merz, 2017; Liao et al., 2017; Fracchia et al., 2018; Huang and MacKerell, 2018). Due to the high sensibility of the molecules upon bivalent metal ion binding, taking the Ca^{2+}-induced conformational changes in CaM as an example, a lack of appropriate approach to model its effects on the molecules could impede the capability of the target molecule to transduce its signals, thus, leading to the inactivation and/or malfunction of the signaling pathways (Gaertner et al., 2004; Wang et al., 2011).

Most calcium-binding proteins such as CaM contain the EF-hand (helix-loop-helix) structural domain as the calcium-binding motif (Lepšík and Field, 2007; Capozzi et al., 2006). In this motif, the calcium ion is coordinated in a pentagonal bipyramidal configuration (**Figure 3**). The canonical positions 1, 3, 5, 7, 9, and 12 represent the positions of the residues involved in calcium coordination (Clapham, 2007; Gifford et al., 2007). **Table 3** shows simple and efficient ways to consider the effect of the Ca^{2+} ions in the force field, where we distributed its charges to the negatively charged residues present in the pentagonal bipyramidal configuration. In contrast to other approaches where the harmonic constraint is applied to the calcium coordinated residues in the loop, or the charges are distributed to all the negatively charged residues in the system, our approach (III) not only considers the heterogeneity of CaM but also takes into account the coordination chemistry of calcium by giving the coordinated residues the freedom to participate to the different conformations of the system, which are relevant to the target selectivity of Ca^{2+}-CaM.

The Distribution of Ca^{2+}-CaM Conformations Is Beyond the Extended and the Collapsed Structures

Ca^{2+}-CaM undergoes substantial conformational changes upon binding to a target, in most cases, the Ca^{2+}-CaM develops a collapsed form that wraps around the target peptide with the two domains coming in proximity (Meador et al., 1993; Westerlund and Delemotte, 2018; Tidow and Nissen, 2013). For various CaMBT molecules, CaM undergoes structural rearrangement to different degrees during the complex formation. Thus, the collapsed conformation of the intact Ca^{2+}-CaM (e.g., **Figure 7**) assembles a vast variety of structures including a structure that resembles the compact conformation of a specific Ca^{2+}-CaM/target complex (**Figure 1D**). This was also observed by X-ray crystallography (Fallon and Quiocho, 2003) and NMR studies (Anthis et al., 2011). This collapsed structure of the Ca^{2+}-CaM (**Figure 7A**) confirms the earlier experiments that show multiple conformations of the Ca^{2+}-CaM in solution, as well as the interaction between opposing domains of CaM (Johnson, 2006; Anthis et al., 2011; Her et al., 2018).

Previous studies have shown that the high flexibility of the linker in solution allows the two domains to independently undergo multiple conformations (Malmendal et al., 1999; Chou et al., 2001; Barton et al., 2002; Hoeflich and Ikura, 2002; Komeiji et al., 2002; Park et al., 2008; Wu et al., 2012). Structural differences can be inferred from crystallography, but the dynamical insights are pivotal to understand Ca^{2+}-CaM interactions with the targets. Experimental and computational studies including NMR spectroscopy, X-ray crystallography, small-angle X-ray scattering, and molecular dynamics simulations unveil important details of the dynamics of Ca^{2+}-CaM and paint a clear portrait of its conformational motions at the origin of the binding selectivity. The available structures of Ca^{2+}-CaM show that it crystallizes into two major conformations: extended (Chattopadhyaya et al., 1992) and collapsed (Fallon and Quiocho, 2003), while in solution, Ca^{2+}-CaM adopts multiple conformations (Project et al., 2006; Park et al., 2008; Wu et al., 2012; McCarthy et al., 2015; Fernandes and Oliveira-Brett, 2017; Jing et al., 2017; Her et al., 2018; Villalobo et al., 2018; Shimoyama, 2019). The conformational changes of CaM upon Ca^{2+} binding have been under intensive investigation (Weinstein and Mehler, 1994; Gilli et al., 1998; Project et al., 2006; Park et al., 2008; Wu et al., 2012). The goal of such studies is to understand the molecular mechanism that leads to the Ca^{2+}-CaM target binding selectivity. How Ca^{2+}-CaM interacts with its CaMBTs is of major interest. This presented community model of CaM allows transitions between a rich set of states beyond the two major extended or the collapsed conformations and are applicable for various Ca^{2+} bound conditions; this further could improve the interpretation of how target binding may tune CaM's affinity for Ca^{2+} ions (Gaertner et al., 2004; Zhang et al., 2017).

CONCLUSION

Here, we developed a "community model" of CaM that samples various conformations of CaM, incorporates various calcium-binding states, and carries the memory of binding with various targets. The model would be useful for studying the reciprocal relationship interplay between target binding and calcium binding through CaM's conformational changes as the next step. We demonstrated a workflow of the development of the AWSEM model for CaM guided by physical knowledge and experimental data. Should new structures of CaM (or CaM/CaMBT complex) structures become available, the model can be easily expanded by including these structures in the memory library.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

MC conceived and designed the research. JN performed the simulations and prepared the figures. MC, JN, PZ, JE, WL, and KK performed data analysis and wrote the article. PW provided critical advice. All authors contributed to the manuscript, read, and approve the submitted version.

ACKNOWLEDGMENTS

We thank Drs. Neal M. Waxham, Piotr Cieplak, and Qian Wang for stimulating discussions. The work was supported by a grant from the National Institutes of Health (2R01GM097553). MSC and PGW are grateful for the funding support from the National Science Foundation (PHY2019745). PGW is grateful for the support from the D. R. Bullard-Welch Chair at Rice University, Grant C-0016. The authors acknowledge the computing resources from the computing clusters uHPC and Sabine at the University of Houston.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.661322/full#supplementary-material

REFERENCES

Anthis, N. J., Doucleff, M., and Clore, G. M. (2011). Transient, Sparsely Populated Compact States of Apo and Calcium-Loaded Calmodulin Probed by Paramagnetic Relaxation Enhancement: Interplay of Conformational Selection and Induced Fit. *J. Am. Chem. Soc.* 133 (46), 18966–18974. doi:10.1021/ja2082813

Archer, C. R., Enslow, B. T., Taylor, A. B., De la Rosa, V., Bhattacharya, A., and Shapiro, M. S. (2019). A Mutually Induced Conformational Fit Underlies Ca2+-Directed Interactions between Calmodulin and the Proximal C Terminus of KCNQ4 K+ Channels. *J. Biol. Chem.* 294 (15), 6094–6112. doi:10.1074/jbc.ra118.006857

Barton, N. P., Verma, C. S., and Caves, L. S. D. (2002). Inherent Flexibility of Calmodulin Domains: A Normal-Mode Analysis Study. *J. Phys. Chem. B* 106 (42), 11036–11040. doi:10.1021/jp026692q

Bekker, G. J., Fukuda, I., Higo, J., and Kamiya, N. (2020). Mutual Population-Shift Driven Antibody-Peptide Binding Elucidated by Molecular Dynamics Simulations. *Sci. Rep.* 10 (1), 1406. doi:10.1038/s41598-020-58320-z

Berchtold, M. W., and Villalobo, A. (2014). The many Faces of Calmodulin in Cell Proliferation, Programmed Cell Death, Autophagy, and Cancer. *Biochim. Biophys. Acta (Bba) - Mol. Cel Res.* 1843 (2), 398–435. doi:10.1016/j.bbamcr.2013.10.021

Berridge, M. J. (2006). Calcium Microdomains: Organization and Function. *Cell Calcium* 40 (5), 405–412. doi:10.1016/j.ceca.2006.09.002

Bootman, M. D. (2012). Calcium Signaling. *Cold Spring Harbor Perspect. Biol.* 4 (7). doi:10.1101/cshperspect.a011171

Capozzi, F., Casadei, F., and Luchinat, C. (2006). EF-hand Protein Dynamics and Evolution of Calcium Signal Transduction: an NMR View. *J. Biol. Inorg. Chem.* 11 (8), 949–962. doi:10.1007/s00775-006-0163-0

Carafoli, E. (2002). Calcium Signaling: A Tale for All Seasons. *Proc. Natl. Acad. Sci.* 99 (3), 1115–1122. doi:10.1073/pnas.032427999

Chattopadhyaya, R., Meador, W. E., Means, A. R., and Quiocho, F. A. (1992). Calmodulin Structure Refined at 1.7 Å Resolution. *J. Mol. Biol.* 228 (4), 1177–1192. doi:10.1016/0022-2836(92)90324-d

Chin, D., and Means, A. R. (2000). Calmodulin: a Prototypical Calcium Sensor. *Trends Cel Biol.* 10 (8), 322–328. doi:10.1016/s0962-8924(00)01800-6

Chou, J. J., Li, S., Klee, C. B., and Bax, A. (2001). Solution Structure of Ca(2+)-Calmodulin Reveals Flexible Hand-like Properties of its Domains. *Nat. Struct. Biol.* 8 (11), 990–997. doi:10.1038/nsb1101-990

Clapham, D. E. (2007). Calcium Signaling. *Cell* 131 (6), 1047–1058. doi:10.1016/j.cell.2007.11.028

Davis, T. N., and Thorner, J. (1989). Vertebrate and Yeast Calmodulin, Despite Significant Sequence Divergence, Are Functionally Interchangeable. *Proc. Natl. Acad. Sci.* 86 (20), 7909–7913. doi:10.1073/pnas.86.20.7909

Davtyan, A., Schafer, N. P., Zheng, W., Clementi, C., Wolynes, P. G., and Papoian, G. A. (2012). AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. *J. Phys. Chem. B* 116 (29), 8494–8503. doi:10.1021/jp212541y

Delfino, F., Porozov, Y., Stepanov, E., Tamazian, G., and Tozzini, V. (2019). Structural Transition States Explored with Minimalist Coarse Grained Models: Applications to Calmodulin. *Front. Mol. Biosciences* 6 (104). doi:10.3389/fmolb.2019.00104

Drake, S. K., Zimmer, M. A., Kundrot, C., and Falke, J. J. (1997). Molecular Tuning of an EF-hand-like Calcium Binding Loop. *J. Gen. Physiol.* 110 (2), 173–184. doi:10.1085/jgp.110.2.173

Duarte, F., Bauer, P., Barrozo, A., Amrein, B. A., Purg, M., Åqvist, J., et al. (2014). Force Field Independent Metal Parameters Using a Nonbonded Dummy Model. *J. Phys. Chem. B* 118 (16), 4351–4362. doi:10.1021/jp501737x

Fallon, J. L., and Quiocho, F. A. (2003). A Closed Compact Structure of Native Ca2+-Calmodulin. *Structure* 11 (10), 1303–1307. doi:10.1016/j.str.2003.09.004

Fernandes, I. P. G., and Oliveira-Brett, A. M. (2017). Calcium-induced Calmodulin Conformational Change. Electrochemical Evaluation. *Bioelectrochemistry* 113, 69–78. doi:10.1016/j.bioelechem.2016.10.002

Fracchia, F., Del Frate, G., Mancini, G., Rocchia, W., and Barone, V. (2018). Force Field Parametrization of Metal Ions from Statistical Learning Techniques. *J. Chem. Theor. Comput.* 14 (1), 255–273. doi:10.1021/acs.jctc.7b00779

Gaertner, T. R., Putkey, J. A., and Waxham, M. N. (2004). RC3/Neurogranin and Ca2+/calmodulin-dependent Protein Kinase II Produce Opposing Effects on the Affinity of Calmodulin for Calcium. *J. Biol. Chem.* 279 (38), 39374–39382. doi:10.1074/jbc.m405352200

Gifford, J. L., Gifford, M. P., and Walsh, H. J. (2007). Structures and Metal-Ion-Binding Properties of the Ca2+-Binding helix-loop-helix EF-Hand Motifs. *Biochem. J.* 405 (2), 199–221. doi:10.1042/BJ20070255

Gilli, R., Lafitte, D., Lopez, C., Kilhoffer, M.-C., Makarov, A., Briand, C., et al. (1998). Thermodynamic Analysis of Calcium and Magnesium Binding to Calmodulin. *Biochemistry* 37 (16), 5450–5456. doi:10.1021/bi972083a

Gong, Z., and Sun, H. (2017). A Coarse-Grained Force Field Parameterized for MgCl2 and CaCl2 Aqueous Solutions. *J. Chem. Inf. Model.* 57 (7), 1599–1608. doi:10.1021/acs.jcim.7b00206

Gromiha, M. M. (2010). "Chapter 6 - Protein Stability," in *Protein Bioinformatics*. Editor M. M. Gromiha (Singapore: Academic Press).

Heidorn, D. B., and Trewhella, J. (1988). Comparison of the crystal and Solution Structures of Calmodulin and Troponin C. *Biochemistry* 27 (3), 909–915. doi:10.1021/bi00403a011

Her, C., Thompson, A. R., Karim, C. B., and Thomas, D. D. (2018). Structural Dynamics of Calmodulin-Ryanodine Receptor Interactions: Electron Paramagnetic Resonance Using Stereospecific Spin Labels. *Scientific Rep.* 8 (1), 10681. doi:10.1038/s41598-018-29064-8

Hoeflich, K. P., and Ikura, M. (2002). Calmodulin in Action. *Cell* 108 (6), 739–742. doi:10.1016/s0092-8674(02)00682-7

Hoffman, L., Chandrasekar, A., Wang, X., Putkey, J. A., and Waxham, M. N. (2014). Neurogranin Alters the Structure and Calcium Binding Properties of Calmodulin. *J. Biol. Chem.* 289 (21), 14644–14655. doi:10.1074/jbc.m114.560656

Homouz, D., Sanabria, H., Waxham, M. N., and Cheung, M. S. (2009). Modulation of Calmodulin Plasticity by the Effect of Macromolecular Crowding. *J. Mol. Biol.* 391 (5), 933–943. doi:10.1016/j.jmb.2009.06.073

Huang, J., and MacKerell, A. D., Jr. (2018). Force Field Development and Simulations of Intrinsically Disordered Proteins. *Curr. Opin. Struct. Biol.* 48, 40–48. doi:10.1016/j.sbi.2017.10.008

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual Molecular Dynamics. *J. Mol. Graphics* 14 (1), 33–38. doi:10.1016/0263-7855(96)00018-5

Ikura, M., Clore, G., Gronenborn, A., Zhu, G., Klee, C., and Bax, A. (1992). Solution Structure of a Calmodulin-Target Peptide Complex by Multidimensional NMR. *Science* 256 (5057), 632–638. doi:10.1126/science.1585175

Jing, Z., Qi, R., Liu, C., and Ren, P. (2017). Study of Interactions between Metal Ions and Protein Model Compounds by Energy Decomposition Analyses and the AMOEBA Force Field. *J. Chem. Phys.* 147 (16), 161733. doi:10.1063/1.4985921

Johnson, C. K. (2006). Calmodulin, Conformational States, and Calcium Signaling. A Single-Molecule Perspective. *Biochemistry* 45 (48), 14233–14246. doi:10.1021/bi061058e

Kataoka, M., Head, J. F., Vorherr, T., Krebs, J., and Carafoli, E. (1991). Small-angle X-ray Scattering Study of Calmodulin Bound to Two Peptides Corresponding to Parts of the Calmodulin-Binding Domain of the Plasma Membrane Calcium Pump. *Biochemistry* 30 (25), 6247–6251. doi:10.1021/bi00239a024

Komeiji, Y., Ueno, Y., and Uebayasi, M. (2002). Molecular Dynamics Simulations Revealed Ca2+ -dependent Conformational Change of Calmodulin. *FEBS Lett.* 521 (1), 133–139. doi:10.1016/s0014-5793(02)02853-3

Lepšík, M., and Field, M. J. (2007). Binding of Calcium and Other Metal Ions to the EF-Hand Loops of Calmodulin Studied by Quantum Chemical Calculations and Molecular Dynamics Simulations. *The J. Phys. Chem. B* 111 (33), 10012–10022.

Li, P., and Merz, K. M., Jr. (2017). Metal Ion Modeling Using Classical Mechanics. *Chem. Rev.* 117 (3), 1564–1686. doi:10.1021/acs.chemrev.6b00440

Liao, Q., Pabis, A., Strodel, B., and Kamerlin, S. C. L. (2017). Extending the Nonbonded Cationic Dummy Model to Account for Ion-Induced Dipole Interactions. *J. Phys. Chem. Lett.* 8 (21), 5408–5414. doi:10.1021/acs.jpclett.7b02358

Liu, F., Chu, X., Lu, H. P., and Wang, J. (2017). Molecular Mechanism of Multispecific Recognition of Calmodulin through Conformational Changes. *Proc. Natl. Acad. Sci. U S A.* 114, E3927. doi:10.1073/pnas.1615949114

Liu, F., Chu, X., Lu, H. P., and Wang, J. (2017). Molecular Mechanism of Multispecific Recognition of Calmodulin through Conformational Changes. *Proc. Natl. Acad. Sci. USA* 114 (20), E3927–E3934. doi:10.1073/pnas.1615949114

Malmendal, A., Evenäs, J., Forsén, S., and Akke, M. (1999). Structural Dynamics in the C-Terminal Domain of Calmodulin at Low Calcium Levels 1 1Edited by P. E. Wright. *J. Mol. Biol.* 293 (4), 883–899. doi:10.1006/jmbi.1999.3188

Marambaud, P., Dreses-Werringloer, U., and Vingtdeux, V. (2009). Calcium Signaling in Neurodegeneration. *Mol. Neurodegeneration* 4 (1), 20. doi:10.1186/1750-1326-4-20

McCarthy, M. R., Thompson, A. R., Nitu, F., Moen, R. J., Olenek, M. J., Klein, J. C., et al. (2015). Impact of Methionine Oxidation on Calmodulin Structural Dynamics. *Biochem. Biophysical Res. Commun.* 456 (2), 567–572. doi:10.1016/j.bbrc.2014.11.091

Meador, W., Means, A., and Quiocho, F. (1993). Modulation of Calmodulin Plasticity in Molecular Recognition on the Basis of X-ray Structures. *Science* 262 (5140), 1718–1721. doi:10.1126/science.8259515

Mehler, E. L., Pascual-Ahuir, J.-L., and Weinstein, H. (1991). Structural Dynamics of Calmodulin and Troponin C. *Protein Eng. Des. Sel* 4 (6), 625–637. doi:10.1093/protein/4.6.625

Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., and Marrink, S.-J. (2008). The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theor. Comput.* 4 (5), 819–834. doi:10.1021/ct700324x

Nandigrami, P., and Portman, J. J. (2016). Coarse-grained Molecular Simulations of Allosteric Cooperativity. *J. Chem. Phys.* 144 (10), 105101. doi:10.1063/1.4943043

Park, H. Y., Kim, S. A., Korlach, J., Rhoades, E., Kwok, L. W., Zipfel, W. R., et al. (2008). Conformational Changes of Calmodulin upon Ca2+ Binding Studied with a Microfluidic Mixer. *Proc. Natl. Acad. Sci.* 105 (2), 542–547. doi:10.1073/pnas.0710810105

Probst, M. M., Radnai, T., Heinzinger, K., Bopp, P., and Rode, B. M. (1985). Molecular Dynamics and X-ray Investigation of an Aqueous Calcium Chloride Solution. *J. Phys. Chem.* 89 (5), 753–759. doi:10.1021/j100251a007

Project, E., Friedman, R., Nachliel, E., and Gutman, M. (2006). A Molecular Dynamics Study of the Effect of Ca2+ Removal on Calmodulin Structure. *Biophysical J.* 90 (11), 3842–3850. doi:10.1529/biophysj.105.077792

Qin, Z., and Squier, T. C. (2001). Calcium-Dependent Stabilization of the Central Sequence between Met76 and Ser81 in Vertebrate Calmodulin. *Biophysical J.* 81 (5), 2908–2918. doi:10.1016/s0006-3495(01)75931-0

Robustelli, P., Piana, S., and Shaw, D. E. (2018). Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. *Proc. Natl. Acad. Sci. USA* 115 (21), E4758–E4766. doi:10.1073/pnas.1800690115

Roux, B. (1995). The Calculation of the Potential of Mean Force Using Computer-Simulations. *Comput. Phys. Commun.* 91 (1-3), 275–282. doi:10.1016/0010-4655(95)00053-i

Seaton, B. A., Head, J. F., Engelman, D. M., and Richards, F. M. (1985). Calcium-induced Increase in the Radius of Gyration and Maximum Dimension of Calmodulin Measured by Small-Angle X-ray Scattering. *Biochemistry* 24 (24), 6740–6743. doi:10.1021/bi00345a002

Shepherd, C. M., and Vogel, H. J. (2004). A Molecular Dynamics Study of Ca2+-Calmodulin: Evidence of Interdomain Coupling and Structural Collapse on the Nanosecond Timescale. *Biophysical J.* 87 (2), 780–791. doi:10.1529/biophysj.103.033266

Shimoyama, H. (2019). A Structural Comparison of 'real' and 'model' Calmodulin Clarified Allosteric Interactions Regulating Domain Motion. *J. Biomol. Struct. Dyn.* 37 (6), 1567–1581. doi:10.1080/07391102.2018.1462730

Shirts, M. R., and Chodera, J. D. (2008). Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* 129 (12), 124105. doi:10.1063/1.2978177

Spoel, D. V. D., Groot, B. L. D., Hayward, S., Berendsen, H. J. C., and Vogel, H. J. (1996). Bending of the Calmodulin central helix: a Theoretical Study. *Protein Sci.* 5 (10), 2044–2053. doi:10.1002/pro.5560051011

Sun, B., and Kekenes-Huskey, P. M. (2021). Assessing the Role of Calmodulin's Linker Flexibility in Target Binding. *Int. J. Mol. Sci.* 22 (9). doi:10.3390/ijms22094990

Tidow, H., and Nissen, P. (2013). Structural Diversity of Calmodulin Binding to its Target Sites. *Febs j* 280 (21), 5551–5565. doi:10.1111/febs.12296

Tripathi, S., Wang, Q., Zhang, P., Hoffman, L., Waxham, M. N., and Cheung, M. S. (2015). Conformational Frustration in Calmodulin-Target Recognition. *J. Mol. Recognit.* 28 (2), 74–86. doi:10.1002/jmr.2413

Tripathi, S., Waxham, M. N., Cheung, M. S., and Liu, Y. (2015). Lessons in Protein Design from Combined Evolution and Conformational Dynamics. *Sci. Rep.* 5, 14259. doi:10.1038/srep14259

Tripathi, S., Waxham, M. N., Cheung, M. S., and Liu, Y. (2015). Lessons in Protein Design from Combined Evolution and Conformational Dynamics. *Scientific Rep.* 5 (1), 14259. doi:10.1038/srep14259

Tsai, M.-Y., Zheng, W., Balamurugan, D., Schafer, N. P., Kim, B. L., Cheung, M. S., et al. (2016). Electrostatics, Structure Prediction, and the Energy Landscapes for Protein Folding and Binding. *Protein Sci.* 25 (1), 255–269. doi:10.1002/pro.2751

Vetter, S. W., and Leclerc, E. (2003). Novel Aspects of Calmodulin Target Recognition and Activation. *Eur. J. Biochem.* 270 (3), 404–414. doi:10.1046/j.1432-1033.2003.03414.x

Vigil, D., Gallagher, S. C., Trewhella, J., and García, A. E. (2001). Functional Dynamics of the Hydrophobic Cleft in the N-Domain of Calmodulin. *Biophysical J.* 80 (5), 2082–2092. doi:10.1016/s0006-3495(01)76182-6

Villalobo, A., Ishida, H., Vogel, H. J., and Berchtold, M. W. (2018). Calmodulin as a Protein Linker and a Regulator of Adaptor/scaffold Proteins. *Biochim. Biophys. Acta (Bba) - Mol. Cel Res.* 1865 (3), 507–521. doi:10.1016/j.bbamcr.2017.12.004

Wall, M. E., Clarage, J. B., and Phillips, G. N., Jr. (1997). Motions of Calmodulin Characterized Using Both Bragg and Diffuse X-ray Scattering. *Structure* 5 (12), 1599–1612. doi:10.1016/s0969-2126(97)00308-0

Wang, Q., Liang, K.-C., Czader, A., Waxham, M. N., and Cheung, M. S. (2011). The Effect of Macromolecular Crowding, Ionic Strength and Calcium Binding on Calmodulin Dynamics. *PLOS Comput. Biol.* 7 (7), e1002114. doi:10.1371/journal.pcbi.1002114

Wang, Q., Zhang, P., Hoffman, L., Tripathi, S., Homouz, D., Liu, Y., et al. (2013). Protein Recognition and Selection through Conformational and Mutually Induced Fit. *Proc. Natl. Acad. Sci.* 110 (51), 20545–20550. doi:10.1073/pnas.1312788110

Weinstein, H., and Mehler, E. L. (1994). Ca2+-Binding and Structural Dynamics in the Functions of Calmodulin. *Annu. Rev. Physiol.* 56, 213–236. doi:10.1146/annurev.ph.56.030194.001241

Westerlund, A. M., and Delemotte, L. (2018). Effect of Ca2+ on the Promiscuous Target-Protein Binding of Calmodulin. *Plos Comput. Biol.* 14 (4), e1006072. doi:10.1371/journal.pcbi.1006072

Wriggers, W., Mehler, E., Pitici, F., Weinstein, H., and Schulten, K. (1998). Structure and Dynamics of Calmodulin in Solution. *Biophysical J.* 74 (4), 1622–1639. doi:10.1016/s0006-3495(98)77876-2

Wu, G., Gao, Z., Dong, A., and Yu, S. (2012). Calcium-induced Changes in Calmodulin Structural Dynamics and Thermodynamics. *Int. J. Biol. Macromolecules* 50 (4), 1011–1017. doi:10.1016/j.ijbiomac.2012.02.017

Wu, H., Wolynes, P. G., and Papoian, G. A. (2018). AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins. *J. Phys. Chem. B* 122 (49), 11115–11125. doi:10.1021/acs.jpcb.8b05791

Yamniuk, A. P., and Vogel, H. J. (2004). Calmodulin's Flexibility Allows for Promiscuity in its Interactions with Target Proteins and Peptides. *Mb* 27 (1), 33–58. doi:10.1385/mb:27:1:33

Yang, C., Jas, G. S., and Kuczera, K. (2001). Structure and Dynamics of Calcium-Activated Calmodulin in Solution. *J. Biomol. Struct. Dyn.* 19 (2), 247–271. doi:10.1080/07391102.2001.10506736

Yang, C., Jas, G. S., and Kuczera, K. (2004). Structure, Dynamics and Interaction with Kinase Targets: Computer Simulations of Calmodulin. *Biochim. Biophys. Acta (Bba) - Proteins Proteomics* 1697 (1), 289–300. doi:10.1016/j.bbapap.2003.11.032

Yang, W., Lee, H.-W., Hellinga, H., and Yang, J. J. (2002). Structural Analysis, Identification, and Design of Calcium-Binding Sites in Proteins. *Proteins* 47 (3), 344–356. doi:10.1002/prot.10093

Zhang, P., Han, J., Cieplak, P., and Cheung, M. S. (2021). Determining the Atomic Charge of Calcium Ion Requires the Information of its Coordination Geometry in an EF-Hand Motif. *J. Chem. Phys.* 154 (12), 124104. doi:10.1063/5.0037517

Zhang, P., Tripathi, S., Trinh, H., and Cheung, M. S. (2017). Opposing Intermolecular Tuning of Ca2+ Affinity for Calmodulin by Neurogranin and CaMKII Peptides. *Biophysical J.* 112 (6), 1105–1119. doi:10.1016/j.bpj.2017.01.020

Zuckerman, D. M. (2004). Simulation of an Ensemble of Conformational Transitions in a United-Residue Model of Calmodulin. *J. Phys. Chem. B* 108 (16), 5127–5137. doi:10.1021/jp0370730

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Nde, Zhang, Ezerski, Lu, Knapp, Wolynes and Cheung. 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: frontiersin.org/about/contact

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