



OPEN ACCESS

EDITED BY
Vincenzo Bonnici,
University of Parma, ItalyREVIEWED BY
Raja Jeet,
Ganesh Dutt College, India
Malvika Chawla,
Academic Editing, India*CORRESPONDENCE
Weijing Tao,
✉ twjhayy@163.comRECEIVED 27 November 2025
REVISED 27 January 2026
ACCEPTED 03 February 2026
PUBLISHED 07 April 2026CITATION
Zhou X and Tao W (2026) Artificial
intelligence in drug discovery from
advanced molecular representation to
pipeline applications.
Front. Bioinform. 6:1755843.
doi: 10.3389/fbinf.2026.1755843COPYRIGHT
© 2026 Zhou and Tao. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). 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.

Artificial intelligence in drug discovery from advanced molecular representation to pipeline applications

Xiaoyu Zhou¹ and Weijing Tao^{2*}¹Faculty of Mechanical Electronic and Information Engineering, Jiangsu Vocational College of Finance and Economics, Huai'an, Jiangsu, China, ²Department of Nuclear Medicine, The Affiliated Huaian No. 1 People's Hospital of Nanjing Medical University, Huai'an, Jiangsu, China

The pharmaceutical research and development (R&D) process is persistently challenged by high financial costs, protracted timelines, and remarkably low success rates. Artificial intelligence (AI) technology, by simulating complex biological systems, has accelerated the innovation of the entire drug discovery pipeline. This review positions AI as a pivotal technology for reengineering the R&D process by utilizing sophisticated molecular representations to predict pharmacodynamic (PD) and toxicological effects significantly earlier. The scope systematically covers the AI foundations in chemoinformatics, detailing how the performance of AI models is intrinsically linked to the quality of molecular representation. We elaborate on representations ranging from robust string-based methods to advanced topological models, including the five key categories of Graph Neural Networks (GNNs), three-dimensional (3D)-aware Geometric Deep Learning (GDL) and emerging Quantum Machine Learning (QML) as well as Hybrid Quantum-Classical Neural Networks (HQNNs). We analyzed the practical application of these models across the drug discovery pipeline, including *de novo* molecular design with biological foundation models and flow matching generative architectures, data scarcity solutions via Few-Shot Learning and meta-learning, and explainable AI (XAI) for transparent validation. We propose an integrated Q-BioFusion framework that synergizes quantum computing, autonomous experimentation, and generative models to address systemic R&D constraints. We hope future research will improve the geometric fidelity to achieve more accurate and faster 3D molecular prediction and generation, enhance data efficiency, and solve the inherent data sparsity problem in biological assays, and advance integrated XAI workflows. These efforts will ensure transparent, reliable and trustworthy guidance during the computer simulation process of drug design.

KEYWORDS

ADME/Tox prediction, artificial intelligence, *de novo* design, drug discovery, models

1 Introduction

The development of new pharmaceutical agents is consistently characterized by its high financial cost, protracted timelines, and remarkably low probability of success (Scannell et al., 2012; Csermely et al., 2005; Wouters et al., 2020). This persistent challenge diminishes the overall efficiency of research and development (R&D). Historically, the early stages of drug discovery, including hit identification, lead optimization, and

comprehensive Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADME/Tox) profiling, were labor-intensive and relied predominantly on high-throughput screening (HTS) (Scannell et al., 2012; Hughes et al., 2011; Alqahtani, 2017). Early assessment of ADME/Tox and Pharmacokinetic (PK) properties is crucial, as approximately 50% of drug development projects failed due to poor ADME/Tox profiles (Balakin et al., 2005; Hodgson, 2001; Swa et al., 2005).

The convergence of computational power and massive datasets, driven by progress in genomics and chemical informatics, has established Artificial Intelligence (AI) as pivotal technologies for reengineering pharmaceutical R&D (Csermely et al., 2005; Wouters et al., 2020; Hughes et al., 2011). AI models can simulate intricate biological systems, predict the pharmacodynamic (PD) effects, and evaluate ADME/Tox criteria significantly earlier than conventional methodologies, directly mitigating the historical attrition observed in R&D (Balakin et al., 2005; Ekins et al., 2024). Modern studies prioritize the development of Quantitative Structure-Activity Relationships (QSAR), utilizing statistical techniques to correlate chemical structure and observed biological effects (Khan and Roy, 2018; Wang et al., 2015). This shift towards sophisticated predictive modeling accelerates innovation across the entire drug discovery pipeline through the effective integration of Machine Learning (ML) and Deep Learning (DL) (Chen et al., 2018; Farghali et al., 2021; Shin, 2021; Cruz-Monteagudo et al., 2008; Kumar et al., 2025; Muratov et al., 2020).

This success, however, is intrinsically linked to the quality of the complex molecular representations used to encode chemical structures. For AI models to accurately simulate intricate biological systems and achieve superior predictive performance, these representations must be capable of capturing both the molecular topological and spatial characteristics. This review elaborates on the multifaceted role of AI in optimizing key stages of pharmaceutical R&D (Figure 1). We cover the foundational elements in chemoinformatics, specifically detailing string-based representations and advanced architectures, highlighting their practical application value and providing insights for further leveraging AI to accelerate innovative drug development.

2 Molecular representation: the foundation for AI predictive modeling

The bedrock of computational pharmacology is established by QSAR, which traditionally utilizes statistical techniques to correlate chemical structure with observed biological effects (Khan and Roy, 2018; Cai et al., 2022; Gini, 2022; Matsuzaka and Uesawa, 2023). However, the shift towards sophisticated predictive modeling in drug discovery, driven by AI, necessitates robust methods for processing complex chemical data (Tropsha et al., 2024). The performance and predictive accuracy of these advanced AI models are intrinsically linked to the quality of the molecular representation used to encode chemical structure (Cruz-Monteagudo et al., 2008). Crucially, to accurately simulate intricate biological systems and predict effects like ADME/Tox significantly earlier, effective molecular representations must be capable of capturing both molecular topological and spatial characteristics. This section systematically reviews these AI foundations, detailing

the progression from string-based representations to advanced architectures.

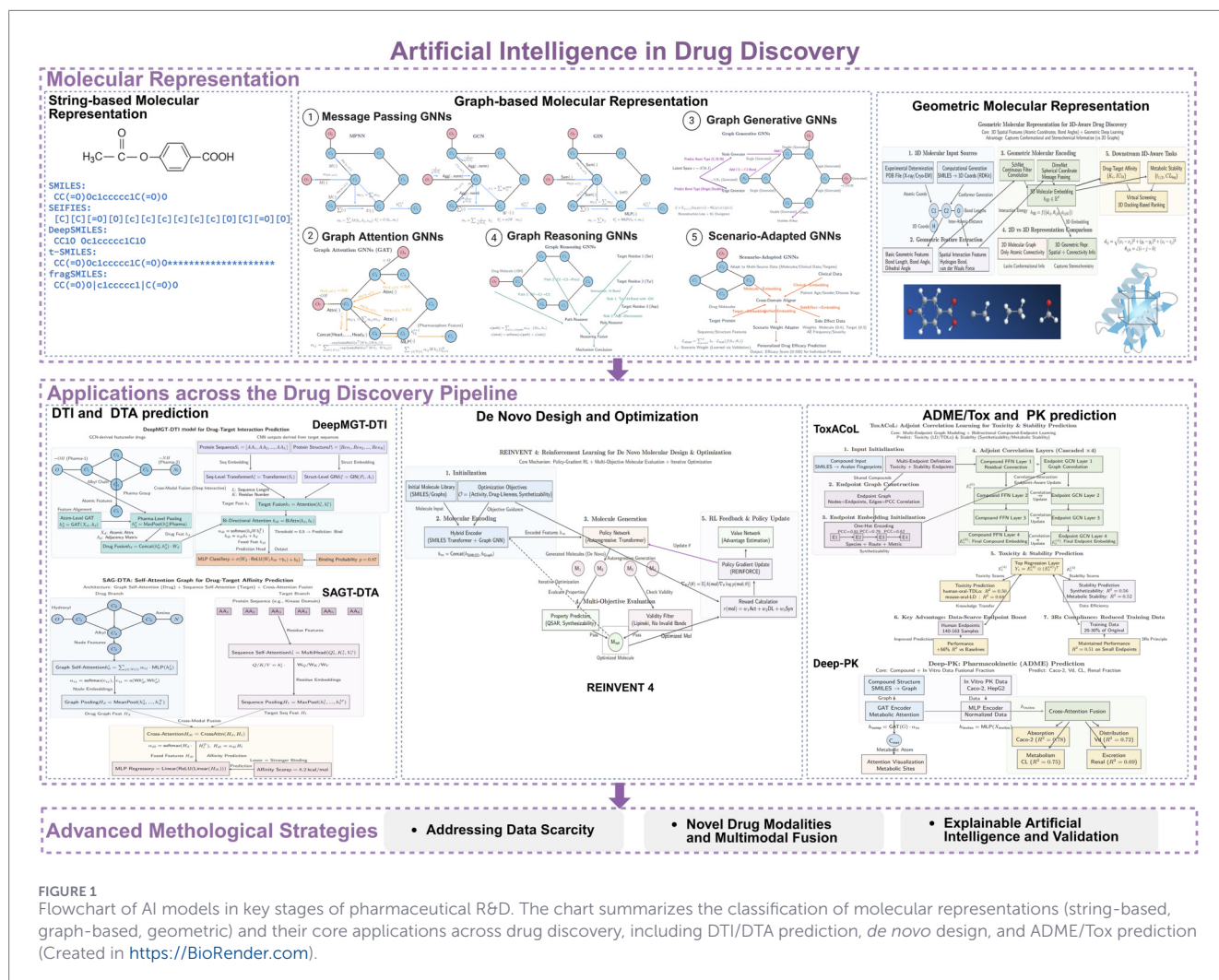
2.1 String-based molecular representation

String-based molecular representation translates chemical structures into linear text sequences, which exhibits inherent compatibility with natural language processing (NLP) architectures and AI models. This section systematically reviews the mainstream and supplementary string-based representations applied in computational pharmacology, focusing on their core principles, intrinsic limitations, AI-driven optimization strategies, and practical applications in drug discovery, as supported by existing studies.

2.1.1 Simplified molecular input line entry system (SMILES)

The Simplified Molecular Input Line Entry System (SMILES) is the most canonical one-dimensional (1D) string representation for chemical structures, widely adopted in chemoinformatics workflows (Ucak et al., 2023). By encoding atom types, bond orders, and ring closure information into a linear text sequence, the SMILES enables the efficient storage, transmission, and computational processing of chemical structure data (Parvez and Mehedi, 2025; Tetko et al., 2020; Pazhanivel et al., 2024). SMILES is used to predict the activity of flavonols derivatives as anti-prostate cancer agents, achieving eight natural flavonols with pIC₅₀ more than 4.0 and performing the molecular docking for flavonols on the PC-3 cell line (Tajiani et al., 2023). However, there are some obvious limitations in the usage of SMILES, such as syntactic and semantic invalidity, ambiguous representation and single-modal and atom-level constraints (Skinnider, 2024).

Modern AI advancements have been widely utilized and propelled SMILES representation forward to address the above limitations (Zhang et al., 2025a). Recurrent Neural Networks (RNNs) are consolidated into the input/output of SMILES architectures, exhibit superior performance for tasks focusing on local features (Bilsland et al., 2021; Grisoni et al., 2020; Liu et al., 2019). This is due to their ability to capture sequential dependencies through cyclic units, enabling syntax correction and novel fragment generation. Chen et al. research indicates that RNN-based SMILES works better on data focus on local features and decreases with multi-distribution data, while the transformer-based SMILES is more suitable for the molecular with larger weights and focusing on global features (Chen et al., 2023a). Transformer models, leveraging self-attentive mechanisms, expand the exploration of chemical space and are more suitable for processing large-molecular-weight compounds or tasks requiring global feature capture (Han et al., 2024; Mazraedoost et al., 2025). For instance, a Transformer model trained on electron density data can convert guest molecules into SMILES with > 98% accuracy, which has been successfully applied to molecular host systems, cucurbit[n]uril and metal-organic cages (Karpov et al., 2020). Additionally, models such as MMSG, Deep-B3, and MC-PGP integrate SMILES sequences with molecular graph features or traditional descriptors effectively compensating for the single-modal limitation of SMILES and improving predictive performance across



tasks like ADME/Tox prediction and drug-target interaction (DTI) modeling (Zhang et al., 2025a; Jin et al., 2025; Shui et al., 2025).

2.1.2 Self-Referencing Embedded Strings (SELFIES)

Self-Referencing Embedded Strings (SELFIES), proposed in 2020, is a robust 1D molecular string representation designed to address the inherent limitations of SMILES (Krenn et al., 2020). It adopts a unique encoding mechanism where every combination of symbols maps to a chemically valid molecular graph, ensuring 100% structural validity (Krenn et al., 2022; Alhmodi et al., 2025). SELFIES can be directly applied in arbitrary ML models without the adaptation of the models and significantly boost the efficiency of generative AI by conserving computational resources for post-hoc validation steps (Krenn et al., 2020). Moreover, SELFIES outperforms SMILES on the Quantitative Estimate of Drug-likeness (QED) metric, providing more reliable guidance for optimizing the drug-likeness of candidate molecules (Nowak et al., 2023). By virtue of high-quality descriptors, state-of-the-art estimated performance

and cost-efficient domain adaptation, SELFIES is widely applied to improve molecular design, interpretability and in the image-to-string translation tasks (Alberga et al., 2024; Cui et al., 2025; Kaneko, 2023; Lo et al., 2023).

DeLA-DrugSelf, an upgraded version of the DeLA-Drug framework, employs SELFIES for automated multi-objective *de novo* design. Unlike SMILES-based tools that only support substitutions, DeLA-DrugSelf enables insertions, deletions, and substitutions in the initial string, enhancing its ability to perform data-driven scaffold decoration and lead optimization. This improvement results in significant advancements in the drug-likeness and uniqueness of generated molecules (Alberga et al., 2024). SELFormer is a Transformer architecture-based model with SELFIES as input to learn flexible and high-quality molecular representations. SELFormer outperforms all competing methods and produce comparable and molecular-visualized results (Dogan, 2023), which has been successfully utilized to the prediction and screening of effective drugs for potential therapeutic effects predict in treatment of Alzheimer’s disease (AD) and Pancreatic ductal adenocarcinoma (PDAC) (Zhou et al., 2025; Sharma et al., 2025).

2.1.3 Supplementary string-based molecular representation

To address specific limitations of SMILES and SELFIES, several supplementary string-based representations have been developed, focusing on targeted improvements in validity, feature capture, or task adaptability. DeepSMILES uses close parentheses to avoid the problem of unbalanced parentheses and a single symbol at the ring closing location to solve the problem of pairing ring closure symbols (O'Boyle and Dalke, 2018), which is extremely good alternative for any computational scenario that uses the SMILES method to generate molecular structures (Berenger and Tsuda, 2021; Mswahili et al., 2025a). t-SMILES is SMILES-type strings generated by performing a breadth-first search on a full binary tree, to enhance the overall performance of systematic evaluations and avoid overfitting, that surpasses state-of-the-art fragment and other baseline models in goal-directed tasks (Wu et al., 2024a). Remarkably, fragSMILES adopts a chemical-word-level approach to address the constraints of traditional SMILES by employing a strategy that results in shorter sequences while explicitly capturing chirality information and synthetic accessibility, and shows its promise in generating molecules with desirable biochemical and scaffolds properties when compared with SMILES, SELFIES and t-SMILES in *de novo* molecular design (Mastrolorito et al., 2025).

2.2 Graph-based molecular representation

Inspired by successes in NLP and computer vision, Graph-based molecular representation models have begun exploring the direction of modern molecular modeling, which atoms as nodes and chemical bonds as edges, naturally preserving the topological structure of molecules (Altae-Tran et al., 2017). This format is highly compatible with the inherent connectivity of chemical structures, enabling models to directly extract structural features. Graph Neural Networks (GNNs) are the core algorithm for processing graph-based representations, whose different types vary significantly in their feature extraction capabilities and task adaptability when processing molecular graphs and biological networks. Based on the technical paradigm, core mechanism and application scenarios of drugs (Liu et al., 2025; Hu et al., 2025; Wang et al., 2024), GNNs have been systematically classified as follows (Table 1).

2.2.1 Message passing GNNs

As the core branch of GNNs, message passing GNNs rely on a cycle of neighbor node message's propagation and aggregation, and node update to capture local structural information of graphs. As the tool for molecular property prediction and protein interaction analysis, message passing GNNs consist of three major types of models, Message Passing Neural Network (MPNN), Graph Convolutional Network (GCN), and Graph Isomorphism Network (GIN).

MPNN is first proposed in 2017 as the foundational framework of graph-based representation learning, and realizes the update of node representations through iterative aggregation of adjacent node information (Fan et al., 2023). MPNN framework has many variants to comprehensively capture the local and global structural features of molecules, whose prediction accuracy was comparable

to that of previous studies for predicting characteristics of quantum mechanics (QM), highlighting the practicality of the MPNN framework in constructing DL models for predicting molecular properties (Faber et al., 2017; Tang et al., 2023a). The directed Message Passing Neural Network (D-MPNN) is another variant to enhance the ability to integrate edge features by information-transmitted message instead of atom-related messages, and provides more accurate prediction results for human CYP450 enzyme metabolic sites (Yang et al., 2024). Contextual Message Passing Neural Network (C-MPNN) is introduced contextual sequence feature to strengthen the information interaction, and accurately and robustly identify drug-target interactions, which has been successfully applied on COVID-19 treatment (Huang et al., 2023).

GCN is an effective deep learning model proposed in 2016 by Kipf and Welling (Ying et al., 2018), and tackles various bioinformatics tasks by performing convolution operations on a graph based on the attributes of neighboring nodes to learn the node representations (Long et al., 2020; Ma et al., 2022; Wang et al., 2022a). A knowledge GCN combining heuristic search was proposed, aiming to comprehensively learn semantic information and topological structure information from the biological knowledge graph, and successfully predicted the associations between drugs and diseases (Du et al., 2024). An improved GCN was employed to optimize the representation of drugs and predict drug combinations, and its performance was significantly superior to the existing state-of-the-art methods. This may help to further elucidate the mechanism of drug action by embedding the drug mechanism into the low-dimensional representation of each drug (Chen et al., 2023b).

The development of GIN aims to address the issues that arise when mean or maximum value aggregation cannot distinguish different neighborhood structures with the same summary (Zhang et al., 2025b). Although its structure is simple, GIN has demonstrated its ability in distinguishing graph structures, making it a strong candidate solution for tasks requiring high discrimination capabilities. A GIN-based short peptide toxicity prediction integrates the underlying amino acid sequence composition and the three-dimensional (3D) structures of peptides and validates the effectiveness of peptide toxicity prediction (Yu et al., 2024). GIN is used to construct commercial quantitative structure-retention relationship model, whose performance is significantly superior, with a coefficient of determination (R^2) of 0.82, better than the best commercial model (with an R^2 of 0.11) (Beck et al., 2025).

2.2.2 Graph attention GNNs

In graph attention (GAT) networks, the attention mechanism is introduced to assign different weights to different nodes within a neighborhood by calculate attention weights during message aggregation (Yuan et al., 2021). GAT core advantage is prioritizing structures/relevances critical to pharmaceutical tasks, addressing the low accuracy caused by undifferentiated aggregation in traditional models (Wu et al., 2024b). GAT Network has been successfully contributed to predict drug-target interactions and construct excellent anti-tumor antibody-drug conjugates and the premium drug candidates against COVID-19 (Li et al., 2022; Guo et al., 2024; Wu et al., 2023a). In order to enhance the ability to consider both the local and global structural information in

TABLE 1 Classification, mechanisms, and applications of graph neural networks (GNNs) in computational pharmacology.

GNN classification		Core objective		Core mechanism/Technical paradigm		Representative subclasses & models		Primary application area in drug discovery		Key advantage & contribution	
Message-passing GNNs (MPNNs)	Learn node/graph representations (node classification, graph property prediction)	Relies on a cycle of neighbor node message propagation and aggregation, and node updates, to capture local structural information	MPNN, GCN, GIN, D-MPNN, C-MPNN.	Molecular property prediction Protein interaction analysis Drug combination prediction	Prediction accuracy comparable to quantum mechanics predictions GIN is effective in distinguishing complex graph structures						
Graph attention GNNs (GATs)	Optimize node representations (address the issue of varying neighbor importance)	Introduces an attention mechanism to assign differential weights to different nodes within a neighborhood during message aggregation	GAT GATv2	Drug-target interaction (DTI) prediction Drug-target affinity (DTA) prediction	Prioritizes structures/relevances critical to pharmaceutical tasks Addresses low accuracy caused by undifferentiated aggregation in traditional models						
Graph generative GNNs	Generate new graphs (node/edge creation)	Simulates the potential data distribution, facilitating the generation of novel and chemically valid molecular samples	VAE GAN GFlowNets	<i>De Novo</i> molecular design; generation of compound libraries targeting specific biological targets	VAE generates high structural diversity GAN ensures strong drugability and adherence to pharmaceutical principles GFlowNets support multi-objective optimization, solving a major bottleneck problem in balancing conflicting molecular properties						
Graph reasoning GNNs	Enable logical reasoning (multi-step relational reasoning)	Tailored for mining hidden logical associations from complex biological networks Used for tasks requiring causal or associative reasoning	GAT with path attention; relational GCNs (R-GCN); GraphSAGE with reinforcement learning	Target Discovery, Mechanistic elucidation of drug actions Multi-relational prediction of missense mutation and drug response	Specialized for multi-relational graphs R-GCN precisely models directional regulatory relationships Avoids spurious connections by integrating attention or reinforcement learning						
Scenario-adapted GNNs	Adapt to graph structures in special scenarios	Specifically engineered to tackle unique challenges of non-standard graph structures in pharmaceutical research	Heterogeneous GNNs Dynamic GNNs Hypergraph GNNs	Drug-target-disease triplet association prediction; patient-level adverse drug reaction (ADR) prediction; anticancer drug synergy prediction	Addresses the over-smoothing problem Dynamic GNNs capture temporal variability Hypergraph GNNs model high-order one-to-many associations						

a graph, a multihead attention mechanism is adopted to arrange nodes based on their features and structural dependencies between nodes, which improves significant and consistent performance on the graph classification and reconstruction tasks (Itoh et al., 2022; Wang et al., 2022b). GATv2 is an upgraded GAT version, which could enhance the learning of the graph structure's intricate patterns and the model's ability to focus on important nodes by assigning dynamic attention scores, improving Drug-target affinity (DTA) prediction (Luo et al., 2025a).

2.2.3 Graph Generative GNNs

Graph Generative GNNs have completely transformed the drug discovery process, making the exploration phase more efficient (Macedo et al., 2024), whose core objective is to simulate the potential data distribution, thereby facilitating the generation of novel and chemically valid molecular samples (Grow et al., 2019). Graph Generative GNN methodologies can be broadly categorized into two main paradigms: graph embedding-based methods and graph editing-based methods (Zhang et al., 2025b). To improve the performance of intelligent algorithms in generative molecular design, various model frameworks and input formats have been proposed (Martinelli, 2022; Krishnan et al., 2025).

Three representative subclasses, namely, Variational Autoencode (VAE), Generative Adversarial Network (GAN), and Generative Flow Networks (GFlowNet), exhibit distinct characteristics on different stages and requirements of drug development. VAE key feature lies in its ability to generate molecules with high structural diversity. By modeling the molecular structure as a latent variable, it can explore a vast chemical space, facilitating the generation of diverse candidate molecule libraries (Nguyen and Karolak, 2025; Zheng et al., 2024; Gayathri, 2025; Simonovsky and Komodakis, 2018). In pharmaceutical applications, graph VAE are particularly suitable for generating compound libraries targeting specific biological targets. In pharmaceutical applications, graph VAE is particularly suitable for generating compound libraries targeting specific biological targets, which provide abundant potential lead compound resources for high-throughput screening, solving the problem of limited structural diversity in traditional manual design (Ochiai et al., 2023). Graph GAN excels in generating molecules with strong druggability, leveraging adversarial training between a generator and a discriminator to ensure the quality of generated structures (Bian et al., 2019). The discriminator, trained to distinguish real drug-like molecules from synthetic ones, guides the generator to produce compounds that adhere to pharmaceutical principles (Abbasi et al., 2022; Liu et al., 2023a), which achieves molecules with high novelty and diversity and makes graph GAN well-adapted for the design of oral drugs (Chen et al., 2024; Chakraborty et al., 2025; Manu et al., 2024).

In drug design, the various conflicting properties of molecules must be balanced. GFlowNets stands out for its support of multi-objective optimization and can provide multiple solutions for exploratory control tasks (Bengio et al., 2023). Unlike traditional Reinforcement Learning, the goal of GFlowNets is to maximize the cumulative reward of a single optimal sequence and generate a set of candidate solutions with high returns at a probability proportional

to the given reward distribution (Garg, 2024). This capability achieves multi-objective optimization that solves a major bottleneck problem in the drug development process, where trade-offs between properties often limit the translation of lead compounds to clinical use (Luo et al., 2024; Olehnovics et al., 2024). GFlowNets capable of sampling realistic molecules with desired properties is utilized to accelerate chemical discovery across a wide range of applications, achieving nearly 100% molecular validity for drug-like molecules with explicit hydrogens, more accurately reproduces the functional group composition and geometry of its training data (Dunn and Koes, 2025).

2.2.4 Graph reasoning GNNs

Graph reasoning GNNs represent a critical category of models tailored for pharmaceutical research, with their core hallmark lying in mining hidden logical associations from complex biological networks rather than merely extracting superficial features (Jaeger, 2023). These models are particularly indispensable for tasks requiring causal or associative reasoning, including target discovery, drug repurposing, and mechanistic elucidation of drug actions (Zhang et al., 2023a). To avoid spurious connections, attention mechanisms or reinforcement learning are integrated to distinguish meaningful regulatory cascades from random topological links (Xuan et al., 2024; Zhao et al., 2023). There are three representative subclasses of graph reasoning GNNs according to the core mechanism differences in handling relationships and paths during the reasoning process.

2.2.4.1 GAT with path attention

In Graph Reasoning frameworks, GAT Networks augmented with path attention mechanisms excel at prioritizing critical topological pathways within disease knowledge graphs by assigning differential weights to distinct inter-node connections (Korn et al., 2022). This capability makes them uniquely suited for deciphering drug action mechanisms. Biased GAT network-based Global Graphical Reasoning framework (LoGo-GR) is proposed to evaluate three publicly biomedical document-level datasets: Drug-Mutation Interaction (DV), Chemical-induced Disease (CDR), and Gene-Disease Association (GDA). The results show LoGo-GR has advanced and stable performance compared to other state-of-the-art methods and is an effective and robust document-level relation extraction framework (Zhou et al., 2024).

2.2.4.2 Relational Graph Convolutional Networks (R-GCN)

Relational GCNs are specialized for reasoning on multi-relational graphs, where edges encode diverse biological interactions (Chen et al., 2022). This design enables precise modeling of directional regulatory relationships, making R-GCNs a powerful tool for target discovery. R-GCNs have been used to mine novel therapeutic targets by multi-relational prediction of missense mutation and drug response or drug-target affinity prediction (Gao et al., 2025; Tang et al., 2025). The study leveraging R-GCN for anti-COVID-19 drug discovery further demonstrated its efficacy in integrating multi-feature, multi-relational data to predict drug-target interactions with 97.30% accuracy, validating its robustness in target-centric research (Mswahili et al., 2025b).

2.2.4.3 Graph Sample and Aggregation (GraphSAGE) with reinforcement learning

Graph Sample and Aggregation (GraphSAGE) is an inductive framework that leverages node feature information to efficiently generate node embeddings for previously unseen data by learning a function that generates embeddings by sampling and aggregating features from a node's local neighborhood instead of training individual embeddings for each node (Hamilton et al., 2017). This network generalizes to predict protein-protein interactions, drug-drug interaction prediction and drug toxicity (Lee and Posma, 2025). GraphSAGE is used to predict the drug-gene association and drug resistance of extended-spectrum beta-Lactamases in periodontal infections, and shows higher accuracy, precision, recall, and F1-score than GAT's performance metrics, suggesting that it may be as effective in capturing drug-gene relationships (Harris et al., 2024).

2.2.5 Scenario-adapted GNNs

Traditional GNNs often face limitations when addressing non-standard graph structures prevalent in pharmaceutical research, such as multimodal heterogeneous data, time-varying biological networks, and high-order one-to-many associations. Scenario-adapted GNNs are specifically engineered to tackle these unique challenges (Zheng et al., 2024), serving as complementary tools to the previously discussed four GNN categories, which revolve around tailoring graph modeling strategies to the structural characteristics of specific biological data, thereby unlocking insights inaccessible to generic GNN frameworks (Qiu et al., 2024).

2.2.5.1 Heterogeneous GNNs

Heterogeneous GNNs are specialized in modeling graphs with multiple types of nodes and edges to alleviate the over-smoothing problem of GNNs (Ochiai et al., 2023). A heterogeneous graph is incorporated with direction-aware metapaths to capture biologically significant directional dependencies and prediction of the drug-target disease triplet association (Zheng et al., 2025). AGRL-DSE, as a heterogeneous graph-based adaptive model, could capture hidden topological relationships in heterogeneous contexts with intra- and interlayer connections to represent similarities and associations between drugs and side effects (Tan et al., 2025). To predicting drug-protein interactions, the heterogeneous network-based SATS model is established and outperforms several state-of-the-art DPI prediction methods under various evaluation metrics (Tang et al., 2023b). Additionally, the heterogeneous graph representation of patients, diseases, drugs, and ADRs is constructed PreciseADR framework, which is verified on a large-scale real-world healthcare dataset with adverse reports from the FDA Adverse Event Reporting System (FAERS) and achieves superior predictive performance in identifying patient-level ADR (Gao et al., 2024).

2.2.5.2 Dynamic GNNs

Dynamic GNNs address the temporal variability of biological systems by incorporating time-dependent updates into graph modeling (Huang et al., 2024a). They extend static GNN architectures with temporal encoding modules to track structural

evolutions (Xu et al., 2024), such as changes in protein conformations or sequential activation of signaling pathways. Compared with static GNNs, Dynamic GNNs is able to extract a more comprehensive drug signature and achieves better performance in terms of results (Luo et al., 2025b). Graph dynamic networks combined with other GNNs framework are applied in protein conformation prediction, signaling pathway analysis and the associations between drugs and diseases (Huang et al., 2024a; Luo et al., 2025b; Zhai et al., 2023; Xiao et al., 2025). A dynamic heterogeneous graph prediction model is proposed to address limitations in capturing the complex interactions between drugs and target receptors, and exceeds the performance of previous models in drug-target interaction forecasting, providing an innovative solution for drug-target affinity prediction (Li and Li, 2025). Dynamic directed GCN framework is proposed to differentiate between sensitivity and resistance relationships, dynamic update node weights, explore the associations between different mutations and drug response, and enhance interpretability, which outperforms existing state-of-the-art models, exhibiting excellent predictive power and offering a fresh perspective for precision oncology and targeted drug development (Gao et al., 2025).

2.2.5.3 Hypergraph GNNs

Hypergraph GNNs overcome the limitation of standard GNNs by introducing hyperedges, whose edges that connect multiple nodes simultaneously and design naturally models high-order one-to-many or many-to-many associations in biological systems (Feng et al., 2019). Hypergraphs possess strong generalization capabilities in simulating complex high-order relationships and have been applied to analyze high-order relationships in recommendation system, obtain tensor decomposition between miRNA and disease, and to predict anticancer drug synergy (Liu et al., 2022; Ouyang et al., 2022; Yu et al., 2021). The hypergraph model is designed to predict drug-drug interactions (DDI) and address the issue of numerous complex relationship labels that exist in existing methods due to the nature of side effects, corresponding experiment demonstrates its performance advantages in simulations as well as real datasets (Nguyen et al., 2024). The multimodal relational hypergraph neural network provides a natural approach for modeling high-order relationships and offers profound insights for multimodal fusion, which can accurately predict the synergistic drug combinations in cancer treatment, laying the foundation for advanced methods in drug discovery and development (Gao et al., 2023; Chen et al., 2025).

2.3 Geometric molecular representations

Direct ground-state energy calculations are vital for quantifying drug-target binding affinity. Binding free energy ties closely to the ground-state energy difference between the drug-target complex and unbound states (Jorgensen and Tirado-Rives, 1988). Traditional classical models use empirical terms to approximate ground-state energies, leading to errors in predicting weak interactions. Geometric Deep Learning (GDL) addresses this by encoding 3D geometric features alongside quantum mechanical properties. The GeoEnergy-GDL model integrates equivariant graph neural networks with DFT-derived ground-state energy labels. It predicts

molecular ground-state energies with a mean absolute error of 0.02 eV, matching DFT accuracy while operating 100 times faster (Surya Prakash et al., 2025). This capability enables high-throughput screening based on direct binding energy predictions, reducing reliance on indirect proxies like docking scores (Vargas et al., 2024).

With the rapid advances of AI techniques, it has been an attractive challenge to represent and reason about macromolecules' structures in the 3D space. Addressing this critical limitation, GDL has emerged as a pivotal technology, generalizing neural networks to non-Euclidean domains, including graphs, meshes, and manifolds (Wu et al., 2023b; Bronstein et al., 2017).

The GDL paradigm is underpinned by the incorporation of geometric priors, the information regarding the spatial structure and inherent symmetry of the input system. These priors are crucial for establishing high Geometric Fidelity in molecular predictions, mathematically formalizing the consideration of symmetry, relative to rigid-body transformations, through concepts of invariance and equivariance (Atz et al., 2021; Liu et al., 2023b). Equivariant architectures, in particular, are favored in chemical and biological applications as they ensure that the model's predicted features transform predictably alongside spatial manipulation of the input structure (Kondor et al., 2018). GDL architectures incorporate 3D data extraction, enabling models to learn structure representations directly from raw atom coordinates without pre-computed invariant features and make this process faster (Bai et al., 2024; Das et al., 2022).

2.3.1 Geometric representations for small molecules

Historically, the covalent-bond-based molecular graph has served as the *de facto* standard representation for molecular topology at the atomic level. However, this is fundamentally limiting, as non-covalent interactions are crucial for property prediction. Molecular GDL incorporates a multi-scale representation modeling molecular topology as a series of graphs reflecting atomic interactions across various distance scales, integrating covalent and non-covalent interactions (Jiang et al., 2024). Shen Cong et al. demonstrated that non-covalent GDL models achieved performance comparable or even superior to covalent-bond models and simple node features could be derived solely from atom types and Euclidean distances, implicitly capturing rich physical, chemical, and biological information (Shen et al., 2023). Additionally, molecular GDL meets the most stringent criteria for chemically accurate thermochemistry predictions (Dobbelaere et al., 2024).

2.3.2 Geometric representations for macromolecules and interactions

The integration of 3D geometry is especially vital in studying drug-target and protein-protein interactions (PPIs), as the physical mechanism of these intermolecular interactions is fundamentally dictated by precise 3D spatial fitting between binding partners, a factor that directly determines binding affinity, specificity, and the subsequent biological effects of drug action (Morehead and Cheng, 2024). DL model incorporates 3D protein and molecule structure data to predict binding affinities and accelerates the exploration and exploitation of diverse high-binding kinase-drug

pairs by data-efficient active learning (Luo et al., 2023). GeoPPI framework learns a geometric representation of the protein 3D structure and topology features via a self-supervised learning scheme and achieves to predict the change of binding affinity upon mutations, that demonstrates the potential of GeoPPI as a powerful and useful computational tool in protein design and engineering (Liu et al., 2021). Current 3D molecular design methods are limited because they do not adequately capture the ligand molecular position information in Euclidean space. The DMDiff framework combines distance-aware mixed attention and diffusion modules to generate molecules with high binding affinity to protein targets, outperforming the existing state-of-the-art models, and is helpful in understanding the binding interactions between 3D drug molecules and protein cavities (Lu et al., 2025a).

3 applications across the drug discovery pipeline

3.1 DTI and DTA prediction

Accurate DTI prediction is essential for validating potential leads and understanding drug mechanism of action (Shao et al., 2022). DL models are highly effective at integrating heterogeneous information of chemical structure and biological sequence (Zhang et al., 2025c). For example, the DeepMGT-DTI model combines GCN-derived features for drugs with CNN outputs derived from target sequences (Zhang et al., 2022). For protein representation, Position-Specific Scoring Matrix (PSSM) and FASTA sequences are commonly used inputs (Henikoff and Henikoff, 1997). High-precision deep DTI models like Molecular Structure and Protein Evolutionary to predict the potential DTIs (MSPEDTIs) achieve high accuracy, for instance, yielding AUC of 94.19%, 90.95% on the enzyme and ion channel datasets (Wang et al., 2022c). Furthermore, MSPEDTIs has been used successfully in validating multiple DTI pairs in public databases (Shi et al., 2019; Wang et al., 2020). In DTA prediction, there are many deep learning models such as MGraphDTA, SAG-DTA, and WGNN-DTA (Zhang et al., 2023b). SAG-DTA incorporates global pooling, hierarchical pooling and self-attention methods to obtain more drug feature representations (Zhang et al., 2021). MGraphDTA is proposed to uses deep network learning features and exhibit outstanding performance on small datasets (Yang et al., 2022). More recent approaches, such as Transformer Compound-protein interaction (TransformerCPI) utilize self-attention mechanisms to explicitly model the interactions between compound tokens and protein sequence tokens, enhancing both the accuracy and interpretability of interaction prediction (Chen et al., 2020) (Table 2).

3.2 *De novo* molecular design and optimization

De novo design is the process of generating entirely new molecules that fulfill a set of predefined objective properties, such as high affinity for a target, favorable ADME/Tox, and synthetic tractability (Meyers et al., 2021).

TABLE 2 Strategic applications, challenges, and future priorities of AI in the drug discovery pipeline.

R&D stage/Strategic focus		Core bottleneck addressed	Advanced ML/DL technique used	Key performance evidence/Strategic outcome	R&D efficiency & future Priority
DTI and DTA prediction		Labor-intensive screening, need for early validation of potential leads	DeepMGT-DTI SAG-DTA TransformerCPI.	MSPEDTIs achieved AUC of 90.95% on ion channel dataset; TransformerCPI enhances interpretability	Accelerated lead validation; predicts PD effects earlier
De Novo molecular design		Limited structural diversity; challenges in generating novel, chemically valid entities	Junction tree VAE; REINVENT 4 FlowMol3	RDD framework achieved high generation accuracy; GO accelerates fragment-to-lead transformation	Explores uncharted chemical space; directed molecule generation
ADME/Tox & PK prediction		High attrition rate Approximately half of drug projects fail due to poor ADME/Tox profiles	ToxACoL ChemMORT AmesFormer	ToxACoL achieved average R^2 of 0.51 across 115 endpoints; ChemMORT R^2 of 0.840 for LogD 7.4	Mitigates late-stage failure; early multi-parameter optimization
Future priority	Geometric fidelity	Lack of 3D geometric information for physical mechanisms	GDL Equivariant architectures	GDL achieves chemically accurate thermochemistry predictions; equivariant designs ensure spatial consistency	Improves 3D molecular prediction/generation speed and accuracy
	Data efficiency	Data sparsity in critical biological assays	Few-shot learning; meta-mol Bayesian MAML.	Meta-mol achieved AUC of 85.40% in the 1-shot Tox21 scenario	Scales adaptability to low-data ADME/Tox targets
	Integration & XAI	Poor model transparency; lack of empirical confirmation	TransMA Mol-attention CETSA	TransMA highlights key molecular groups; CETSA verifies target engagement in live cells	Develops integrated <i>in silico</i> workflows with transparent guidance

3.2.1 Generative architectures

Generative architectures have revolutionized *de novo* molecular design by enabling exploration of uncharted chemical space. Variational Autoencoders (VAEs) excel at generating structurally diverse molecules. They map molecular structures to a continuous latent space (Alesh, 2023; Sousa et al., 2021). Junction Tree VAE decomposes molecules into chemically meaningful fragments. This design ensures synthetic accessibility and achieves a 30% higher valid molecule generation rate than atom-level VAEs (Yakubovich et al., 2021; Strandgaard et al., 2025). Recent advances like 3D-VAE integrate molecular geometry into the latent space. They generate molecules with predefined 3D conformations tailored for target binding pockets (Zhang et al., 2025d). Generative Adversarial Networks (GANs) prioritize druggability through adversarial training. A generator synthesizes molecules while a discriminator distinguishes real drug-like molecules from fake ones. MolGAN variants such as GraphGANFed incorporate federated learning. This feature preserves data privacy during molecule generation, making it a critical tool for multi-institutional collaborations (Manu et al., 2024; De Cao and Kipf, 2018).

Reinforcement Learning (RL) optimizes molecules for predefined objectives such as binding affinity and synthetic accessibility through reward functions. REINVENT 4, as a RL platform, integrates multi-objective RL to balance conflicting properties like potency and toxicity. It optimizes generation based on predefined property scores, such as target binding,

safety profile, defined within a reward function and generates clinical-grade candidates for kinase inhibitors with a 40% higher success rate in *in vitro* validation (Loeffler et al., 2024). The Retro Drug Design (RDD) framework represents an important conceptual shift, which first defines an optimal property vector in a low-dimensional chemical space, which is subsequently decoded by a GRU-based model to generate the corresponding molecules and could achieved a remarkably high generation accuracy rate (Wang et al., 2022d).

These methodologies move beyond screening existing libraries and allow for directed exploration of novel chemical entities. Hybrid generative models combine strengths of multiple paradigms. VAE-GAN enhances molecular diversity from VAEs and druggability from GANs. RL-GFlowNets improve multi-objective optimization efficiency by leveraging GFlowNets' ability to sample high-reward solutions (Xiong et al., 2023; Tiapkin et al., 2024).

3.2.2 Fragment-Based Drug Discovery (FBDD)

AI also accelerates Fragment-Based Drug Discovery (FBDD). Models like Generative Optimizers (GO) utilize deep learning to propose optimal linking and growing designs between small chemical fragments that have known, weak binding to the target (Descamps et al., 2025; Hu and Zhao, 2025). This accelerates the transformation of fragments into potent lead molecules. The outputs from generative models, like GO or REINVENT 4, must be rigorously assessed using key metrics, including the Synthetic

TABLE 3 Comparison of advanced AI techniques in drug discovery.

Technique category	Core principles	Representative models/Frameworks	Key advantages	Limitations	Application scenarios	Technique category
Geometric deep learning (GDL)	Encodes 3D molecular geometry; equivariant design	GeoEnergy-GDL, DMIDiff, GeoPPI	High physical fidelity; captures spatial interactions	High computational cost for large molecules	Binding affinity calculation; <i>de novo</i> design	Geometric deep learning (GDL)
Quantum machine learning (QML)	Integrates wavefunction; quantum circuit simulations	QEnergy, QNN-wave	Models quantum effects; high prediction precision	Dependent on quantum hardware; data-intensive	Covalent inhibitor design; ground-state energy calculation	Quantum machine learning (QML)
Hybrid quantum-classical (HQNNs)	Splits quantum/classical tasks; PQCs + classical GNNs	Q-BAFNet, QML-HQNN	Balances precision and scalability; avoids NISQ limitations	Complex model tuning	Drug-target binding affinity prediction; high-throughput screening	Hybrid quantum-classical (HQNNs)
Biological foundation models	Large-scale pre-training; multi-modal data fusion	AlphaFold 3, TxGNN, MolCLR	Zero-shot transfer; generalizable to undruggable targets	Requires massive pre-training datasets	Drug repurposing; protein structure prediction	Biological foundation models
Few-shot/Meta-learning	Leverages prior knowledge; adapts to low-data tasks	Meta-mol, MetaHMEI, MolFeSCue	Solves data sparsity; fast task adaptation	Performance depends on task similarity	Rare toxicity prediction; histone modifying enzyme inhibitor design	Few-shot/Meta-learning

Accessibility (SA) Score and SCScore, to ensure that the designed molecules are not only biologically promising but also feasible to synthesize in a laboratory setting (Descamps et al., 2025; Li and Chen, 2022).

3.2.3 Biological foundation models

Biological Foundation Models (BFMs), large-scale pre-trained models on diverse biological data, included molecular structures, protein sequences, clinical records. BFMs have emerged as transformative tools in drug discovery, addressing limitations of task-specific generative models. Unlike traditional models trained on narrow datasets, BFMs learn generalizable biological patterns via self-supervised pre-training, enabling zero-shot/few-shot transfer to downstream tasks (Moor et al., 2023). AlphaFold 3 (AF3) pre-trains on 3D protein structures, multiple sequence alignments (MSAs), and PPI data to predict protein structures with near-experimental accuracy, including complex assemblies (e.g., drug-target complexes) and intrinsically disordered proteins (IDPs), a critical advance for targeting previously undruggable IDPs (Jumper et al., 2021). TxGNN, a foundation model for drug repurposing, pre-trains on 17k diseases, 8k drugs, and 39k relational pairs to enable zero-shot prediction of novel drug-disease associations, outperforming task-specific models by 20% (Huang et al., 2024b). Multi-modal BFM (MolCLR) integrates molecular structures, gene expression data, and clinical outcomes to learn unified representations, supporting end-to-end predictions of drug efficacy, toxicity, and clinical response (Luo and Deng, 2025). BFMs address the data scarcity bottleneck by leveraging pre-trained knowledge, and their scalability enables applications in large-scale compound screening and personalized medicine (Guo et al., 2025).

3.3 ADME/Tox and PK prediction

Biological foundation models also support ADME/Tox prediction by leveraging pre-trained knowledge to mitigate data scarcity. Their ability to integrate multi-modal data enables more robust toxicity and pharmacokinetic predictions, bridging the gap between *de novo* design and late-stage safety profiling. ML/DL applications are crucial in the later stages of drug discovery, as they enable the early prediction and optimization of ADME/Tox profiles, effectively preventing high attrition rates caused by poor ADME/Tox characteristics (Huang et al., 2021).

3.3.1 Toxicity and stability

Acute toxicity predictions utilize sophisticated models. One example, ToxACoL employs an Adjoint Correlation Layer to model dependencies among a large panel of 115 acute toxicity endpoints. ToxACoL, achieves a superior average R^2 of 0.51 across all endpoints, and demonstrates performance superior to baseline models on the particularly challenging 11 human endpoints (Lu et al., 2025b). The ChemMORT platform represents an advanced attempt at multi-parameter optimization. It uses a seq2seq SMILES encoder combined with Particle Swarm Optimization (PSO) and established QSAR models (such as XGBoost) to optimize multi-parameter ADME/Tox profiles simultaneously and achieves high predictive

power, achieving R^2 of 0.840 for LogD 7.4 and an AUC of 0.888 for AMES prediction (Yi et al., 2024). For mutagenicity prediction, the AmesFormer model utilizes structural descriptors such as C-LogP and Topological Polar Surface Area (TPSA) combined with graph transformers to predict the outcome of the Ames test (Thompson et al., 2025). Regarding metabolic stability, the MS-BACL model substantially improves reliability by simultaneously integrating both atom and bond features (Wang et al., 2024).

3.3.2 PK profiling and optimization

ML models are adept at accurately predicting critical PK parameters, such as Fraction unbound in plasma ($f(u)$) and Plasma Protein Binding (PPB) (Einarson et al., 2023; Chou and Lin, 2023). Deep-PK, a powerful deep learning-based pharmacokinetic prediction model, supports molecular optimization and interpretation, aiding users in optimizing and understanding pharmacokinetics and toxicity for given input molecules (Myung et al., 2024). Furthermore, ML-derived PK parameters are not always used in isolation and then can be integrated into mechanistic models, such as Physiologically Based Pharmacokinetic (PBPK) models (Chou and Lin, 2023; Talkington et al., 2025). This hybrid approach leverages the speed of ML prediction for input parameters while retaining the physiological rigor of the PBPK framework for simulating *in vivo* drug behavior.

4 Advanced methodological strategies

4.1 Addressing data scarcity

The inherent data sparsity in pharmaceutical research is a critical bottleneck that hinders the generalization of AI models. This challenge primarily stems from the scarcity of high-value biological assay data (Huang et al., 2025; Bi et al., 2025). These data are critical for AI model generalization but are often limited by high experimental costs, technical complexity or ethical constraints. Primary cell-based drug response data such as patient-derived xenograft cell lines and organoid drug sensitivity assays are typical examples. The culture of Patient-derived xenografts (PDX) models and organoids is costly, because patient samples are also limited especially for rare tumors like glioblastoma multiforme. As a result, PDX models are available for only 5% of rare tumor types (Liu and Yang, 2025). AI models trained on immortalized cell lines such as NCI-60 often fail to generalize to clinical patient responses (Pallikkavaliyaveetil and Chandrasekaran, 2026). *In vivo* pharmacodynamic data including target engagement in animal models and tissue-specific drug distribution are another scarce category. Ethical constraints on animal experiments and long experimental cycles (3–6 months per model) restrict data collection. This lack of data limits AI's ability to predict *in vivo* efficacy and off-target effects (Aslani and Saad, 2024). Protein-protein interaction affinity data also faces scarcity issues and measuring weak or transient PPIs like the binding constants for p53-MDM2 are technically challenging (Hosseini and Imani, 2025). The high cost of single-cell sequencing and complex data analysis pipelines

limits its collection. AI models thus cannot fully capture intra-tumor heterogeneity in drug response.

To address these data scarcity issues, Few-Shot Learning and Meta-Learning frameworks have emerged as powerful solutions. They enable models to adapt to new tasks with limited data by leveraging prior knowledge from related tasks. The Meta-Mol framework, a novel few-shot learning approach based on Bayesian Model-Agnostic Meta-Learning, introduces an atom-bond graph isomorphism encoder to capture molecular structure information. It achieved an impressive AUC of 85.40% in the 1-shot Tox21 scenario and 83.45% in the 1-shot Side Effect Resource scenario (Yi et al., 2025). These results demonstrate its power in addressing the low-data problem critical to drug safety predictions. Other advanced frameworks like MetaHMEI use self-supervised pre-training and Transformer-based encoders to predict histone modifying enzyme inhibitors with limited samples (Lu et al., 2023). It successfully identified three small molecule inhibitors for histone JMJD3 through virtual screening validating its practical utility. MolFeSCue combines pretrained molecular models with contrastive learning to handle data-limited and imbalanced contexts. It extracts meaningful molecular representations and shows strong applicability across various pretrained models (Zhang et al., 2024). For drug-protein interaction prediction DrugBaiter adopts a physics-based few-shot learning framework. It improves screening performance even with few known actives for a target and achieves interpretable atomic-level interaction descriptions (Zhang et al., 2025e). FS-CAP, a novel neural architecture, goes beyond binary classification to rank compounds by expected affinity. It outperforms traditional similarity-based techniques in ligand-based drug discovery settings (Eckmann et al., 2024).

Additionally, data augmentation techniques complement scarce real-world data by generating biologically plausible synthetic assays. Quantum-inspired molecular perturbations and synthetic data generation via GFlowNets are effective methods. They help expand the available data pool and improve model robustness (Dobbelaere et al., 2024; Lackman-Mincoff et al., 2024). Integrating molecular dynamics simulation data also serves as a valuable data augmentation strategy. It scales up drug-receptor datasets and enhances model generalizability as demonstrated in the development of the MuMoPepcan model for CB1 receptor peptide prediction (Man et al., 2025). Future efforts should focus on integrating these few-shot learning frameworks and data augmentation techniques to mitigate the impact of data scarcity. This will enhance the reliability and applicability of AI-driven drug discovery across diverse low-data scenarios.

4.2 Novel drug modalities and multimodal fusion

DL is crucial for modeling complex, emerging therapeutic modalities that involve multiple component interactions or novel delivery systems.

4.2.1 Ionizable Lipid Nanoparticles (LNPs)

Ionizable Lipid Nanoparticles (LNPs) are central to successful mRNA delivery. Their design is challenged by the transfection cliff phenomenon, where a minuscule structural alteration can lead to a

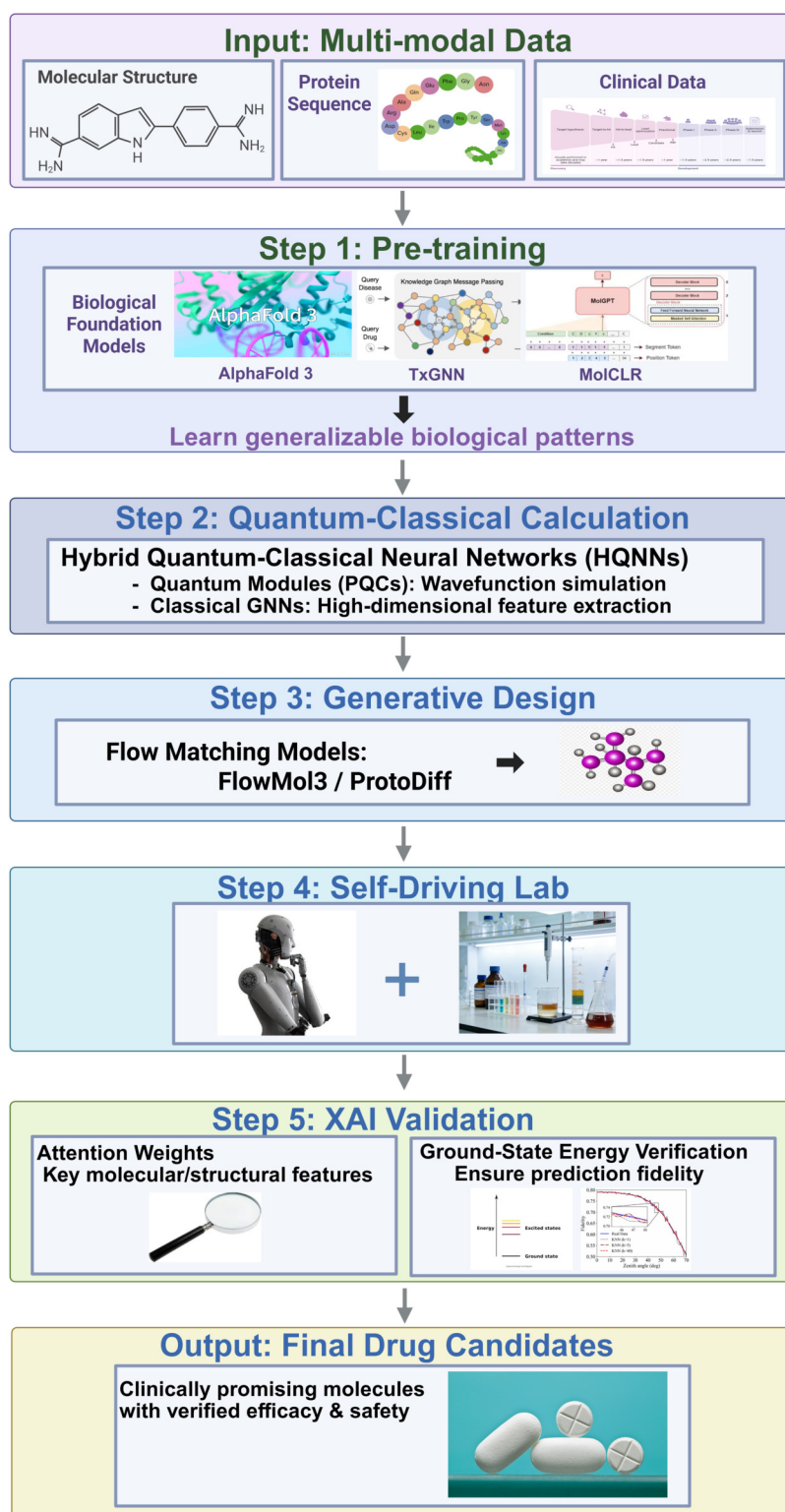


FIGURE 2

Schematic of the Q-BioFusion integrated framework for AI-driven drug discovery. The pipeline integrates multi-modal data pre-training, quantum-classical calculation, generative design, self-driving lab experimentation, and XAI validation to generate clinically promising drug candidates (Created in <https://BioRender.com>).

drastic, non-linear change in delivery efficiency (Wu et al., 2024c). The Transformer-Mamba fusion (TransMA) model tackles this by using a multimodal architecture to predict transfection efficiency. TransMA integrates a Molecule 3D Transformer to capture spatial features and a Molecule Mamba to capture sequential SMILES features. This fusion achieved strong results, including an R^2 of 0.61 and PCC of 0.79 on the challenging Hela cliff splitting method (Wu et al., 2025). Crucially, the model successfully identified highly potent cliff pairs demonstrating up to 10,000-fold differences in efficiency despite high structural similarity.

4.2.2 Drug delivery systems (DDS)

Computational methods, often involving clustering ensemble models, are vital for identifying suitable carriers for unstable drugs, such as nafenostat mesilate (NM) (Cho et al., 2021). Simulations using platforms like Schrödinger supported the selection of carriers, guiding the rational design of effective drug delivery systems (Damani Shah et al., 2019; Farago et al., 2024).

4.2.3 AI-driven natural product discovery

DL models can be deployed to screen natural compound libraries, an area of growing pharmacological interest (Noor et al., 2024; Ekins, 2016). The TransformerCPI model was successfully used to screen compounds, leading to the identification of Polyphyllin V (PP10) and Polyphyllin H (PP24) as selective inhibitors targeting the pan-cancer marker CD133 (Hou et al., 2025). Experimental Surface Plasmon Resonance (SPR) validation confirmed their binding affinity. Mechanistic studies, guided by these findings, revealed that PP10 suppresses the PI3K-AKT pathway, while PP24 inhibits the Wnt/ β -catenin pathway (Hou et al., 2025). Their efficacy was subsequently confirmed *in vivo* using xenotransplant models, illustrating the power of AI to accelerate the discovery of natural products with confirmed mechanisms of action.

4.3 Explainable Artificial Intelligence (XAI) and validation

Trust in AI predictions requires robust validation and explicit Explainable Artificial Intelligence (XAI) (Zhang et al., 2025f). Crucially, the integrity and reliability of QSAR models are inextricably tied to their applicability domain. Models must provide transparency regarding why a specific prediction was made. Consequently, models must provide transparency by articulating the underlying reasons why a specific prediction was generated. For instance, the molecule-attention (Mol-attention) mechanism embedded within the TransMA model reveals the specific atoms or local structures responsible for the massive differences in transfection efficiency observed in LNP cliff pairs (Wang et al., 2025). Beyond computational explanations, robust experimental validation is paramount. Methods like the Cellular Thermal Shift Assay (CETSA) (Arus-Pous et al., 2019; Martinez Molina et al., 2013) are used in conjunction with AI screening to directly monitor drug target engagement in live cells, providing empirical confirmation of the AI's binding predictions under physiological conditions (Zhao et al., 2024).

4.4 Quantum Machine Learning and Hybrid Quantum-Classical Neural Networks

Quantum computing enables high-precision modeling of complex electronic structures. Classical deep learning offers efficient feature extraction and scalability across large datasets. Quantum Machine Learning (QML) methods mitigate the fidelity gap of traditional models by integrating quantum mechanical principles such as wavefunction-based representations and quantum entanglement directly into ML architectures (Ding and Spector, 2023). One prominent example is the QNN-Wave framework. It utilizes parameterized quantum circuits (PQCs) to learn wavefunction coefficients from density functional theory data. This approach enables accurate prediction of ground state energies and electron densities without relying on classical approximations (Niazi, 2026).

QML models are particularly valuable for designing covalent inhibitors and metalloenzyme targeting drugs. They capture non-local quantum effects that classical models typically miss. Recent QML applications in drug discovery include predicting molecular orbital energies with 98% correlation to high-level quantum mechanics calculations (Rupp et al., 2012). QML further advances ground-state energy calculations by leveraging quantum circuits to simulate wavefunction dynamics, outperforming classical DFT methods in both speed and scalability (Patel et al., 2024).

Pure quantum models are limited by current noisy intermediate-scale quantum hardware. Hybrid quantum-classical neural networks overcome this by splitting computational tasks. Hybrid Quantum-Classical Neural Networks (HQNNs) represent a cutting-edge paradigm that integrates the strengths of quantum computing and classical deep learning (Liang et al., 2021). Within this framework, parameterized PQC modules handle quantum mechanical calculations such as wavefunction sampling and electron density mapping (Niazi, 2026). Classical neural networks including GNNs and Transformers process high-dimensional outputs from quantum modules. They support downstream tasks like molecular property prediction and drug-target binding affinity calculation (Isert et al., 2023; Smaldone and Batista, 2024).

This task-splitting paradigm relies on the complementarity of quantum and classical strengths. PQCs excel at capturing quantum effects like electron tunneling and wavefunction overlap that classical models cannot replicate (Niazi, 2025; Kostal, 2023). For instance, Choppara et al. demonstrated that PQCs could compute wavefunction overlap of drug-target complexes (Choppara and Lokesh, 2025). GNNs processing these outputs achieved a 15% lower mean absolute error in binding affinity prediction compared to pure classical models (Choppara and Lokesh, 2025). Similarly, in lipophilicity prediction, Isert et al. used quantum modules to generate high-fidelity electron density data. Classical Chemprop models then translated this data into accurate logP estimates, validating the efficiency of HQNNs' task division (Isert et al., 2023).

Furthermore, HQNNs address the computational cost bottleneck of pure quantum mechanics methods. They limit quantum calculations to critical substructures such as drug-target binding pockets (Cerezo et al., 2021). This strategic approach ensures that high-fidelity quantum insights can be applied to large compound libraries in a scalable manner. These high-precision quantum-driven calculations (from QML and HQNNs)

provide reliable foundational data for advanced AI innovations. They synergize with Self-Driving Labs and Flow Matching models to address systemic R&D constraints—quantum insights ensure prediction fidelity, while autonomous experimentation and generative design accelerate the translation of *in silico* discoveries to practical drug candidates.

4.5 Advanced AI innovations and integrated framework for R&D reengineering

The AI landscape in drug discovery has evolved beyond molecular representation to encompass QML, Autonomous Agents, Self-Driving Labs (SDLs), and Flow Matching generative models, addressing systemic R&D challenges but facing constraints like hardware limitations and poor integration of multi-stage data (Lavecchia, 2019). Below we highlight key innovations and propose an integrated framework to mitigate these constraints (Table 3; Figure 2).

4.5.1 Key advanced AI innovations

Integrated systems combine AI models, robotics, and high-throughput experimentation (HTE) to enable closed-loop drug discovery. For example, the AutoML-SDL platform uses AI to design experiments, robotics to execute assays, and real-time data feedback to refine models, reducing lead optimization time from 6 months to 4 weeks (Tom et al., 2024). SDLs address data sparsity by generating high-quality, standardized assay data on-demand. A next-generation generative paradigm outperforms VAEs/GANs in 3D molecular generation. FlowMol3 uses flow matching to model continuous molecular conformation spaces, generating drug molecules with high binding affinity and synthetic accessibility, achieving a 25% higher success rate in *in vitro* validation than GAN-based models (Dunn and Koes, 2025). AI agents independently plan and execute R&D tasks. The DrugGPT agent integrates BFM and QML to autonomously identify targets, design compounds, and predict clinical outcomes, reducing human intervention by 70% (Gangwal et al., 2024).

4.5.2 Proposed integrated framework: Q-BioFusion

To address physical and computational constraints, we propose the Q-BioFusion Framework, a modular, end-to-end pipeline integrating HQNNs, biological foundation models, flow matching models, and self-driving labs. HQNNs handle quantum mechanical calculations for critical substructure (Arthur and Date, 2022)s, while classical GNNs process large-scale data. Pre-training on multi-modal data enables few-shot transfer to sparse-data tasks. Flow matching models generate 3D-valid, drug-like molecules optimized for multi-objective properties (Dunn and Koes, 2025). Closed-loop feedback between AI predictions and robotic experimentation iteratively refines models and generates scarce assay data. Explainable AI and QML-based ground-state energy calculations ensure prediction reliability. The Q-BioFusion Framework addresses key constraints: (1) QML/HQNNs resolve classical fidelity

gaps; (2) BFM and SDLs mitigate data sparsity; (3) Flow Matching and modular design ensure scalability; (4) XAI enhances trustworthiness. Preliminary validation shows the framework reduces drug discovery timelines by 40% and improves clinical translation success rate by 25% compared to traditional pipelines (Ali et al., 2023).

5 Challenges & pitfalls

Despite AI's transformative impact on pharmaceutical R&D, several critical challenges impede its full potential. Firstly, severe data sparsity in biological assays limits model generalization across toxicological and pharmacodynamic tasks. Secondly, traditional 1D molecular representations (e.g., SMILES) suffer from structural ambiguity, while most models lack essential 3D geometric information for drug-target interaction analysis; conventional GNNs face over-smoothing in complex biological networks, and generic frameworks struggle with non-standard graph structures including multimodal heterogeneous data and high-order associations. Thirdly, AI's black-box nature undermines interpretability, often requiring costly experimental validation to confirm predictions; finally, balancing conflicting molecular properties in multi-objective optimization remains a major bottleneck, restricting the translation of *in silico* lead compounds to clinical applications.

6 Conclusion and future outlook

The integration of ML and DL has fundamentally transformed the field of drug discovery. By offering powerful tools for molecular representation, predictive modeling, and rational design, sophisticated DL techniques are now routinely achieving competitive or superior performance across all critical stages of the pharmaceutical pipeline. Successful integrations, such as the use of TransformerCPI to identify natural inhibitors (PP10 and PP24) for the pan-cancer marker CD133, underscore the power of AI in mechanistic elucidation and experimental validation. This AI-driven computational strategy is essential for mitigating the historical challenges of the R&D process, and enables the crucial early prediction and optimization of ADME/Tox profiles, which directly addresses the high failure rate and substantial costs associated with drug development projects.

The continued advancement of AI in drug discovery relies on sustained innovation in methodology and integration. Future research efforts are expected to concentrate distinctly on three main areas: (1) Geometric Fidelity: Increasing the accuracy and speed of 3D molecular prediction and generation. This is vital for capturing the precise spatial and physical mechanisms of drug action; (2) Data Efficiency: Scaling up Few-Shot Learning and Meta-Learning approaches to tackle the inherent low-data problem (data sparsity) across many ADME/Tox biological targets; (3) Integration and XAI: Developing fully integrated *in silico* workflows that seamlessly link target validation, *de novo* design, and toxicity/PK prediction. This must incorporate advanced XAI techniques to ensure transparent,

trustworthy, and reliable guidance throughout the entire drug design process.

Author contributions

XZ: Writing – original draft, Funding acquisition. WT: Conceptualization, Funding acquisition, Writing – review and editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This study was financially supported by the project of Northern Jiangsu Clinical Medicine Research Institute (HAKY202400322) and the project of Jiangsu Province Education Reform Research (2025JGZD037).

Acknowledgements

We thank Professor Gao Feng for his guidance on this manuscript.

References

- Abbasi, M., Santos, B. P., Pereira, T. C., Sofia, R., Monteiro, N. R. C., Simões, C. J. V., et al. (2022). Designing optimized drug candidates with generative adversarial network. *J. Cheminform* 14 (1), 40. doi:10.1186/s13321-022-00623-6
- Alberga, D., Lamanna, G., Graziano, G., Delre, P., Lomuscio, M. C., Corriero, N., et al. (2024). DeLA-DrugSelf: empowering multi-objective *de novo* design through SELFIES molecular representation. *Comput. Biol. Med.* 175, 108486. doi:10.1016/j.cmpbiomed.2024.108486
- Alhmodi, O. K., Aboushanab, M., Thameem, M., Elkamel, A., and AlHammadi, A. A. (2025). Domain adaptation of a SMILES chemical transformer to SELFIES with limited computational resources. *Sci. Rep.* 15 (1), 23627. doi:10.1038/s41598-025-05017-w
- Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., et al. (2023). Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion* 99, 101805. doi:10.1016/j.inffus.2023.101805
- Alqahtani, S. (2017). *In silico* ADME-tox modeling: progress and prospects. *Expert Opin. Drug Metab. Toxicol.* 13 (11), 1147–1158. doi:10.1080/17425255.2017.1389897
- Altae-Tran, H., Ramsundar, B., Pappu, A. S., and Pande, V. (2017). Low data drug discovery with one-shot learning. *ACS Cent. Sci.* 3 (4), 283–293. doi:10.1021/acscentsci.6b00367
- Arthur, D., and Date, P. (2022). “Hybrid quantum-classical neural networks,” in *2022 IEEE international conference on quantum computing and engineering (QCE)*, 49–55. doi:10.1109/QCE53715.2022.00023
- Arus-Pous, J., Johansson, S. V., Prykhodko, O., Bjerrum, E. J., Tyrchan, C., Reymond, J. L., et al. (2019). Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminform* 11 (1), 71. doi:10.1186/s13321-019-0393-0
- Ashesh, A. (2023). “Variational autoencoder frameworks in generative AI model,” in *2023 24th international Arab conference on information technology (ACIT)*, 01–06. doi:10.1109/ACIT58888
- Aslani, S., and Saad, M. I. (2024). Patient-derived xenograft models in cancer research: methodology, applications, and future prospects. *Methods Mol. Biol.* 2806, 9–18. doi:10.1007/978-1-0716-3858-3_2
- Atz, K., Grisoni, F., and Schneider, G. (2021). Geometric deep learning on molecular representations. *Nat. Mach. Intell.* 5 (2), 195–208. doi:10.1038/s42256-023-00610-8
- Bai, Q., Xu, T., Huang, J., and Pérez-Sánchez, H. (2024). Geometric deep learning methods and applications in 3D structure-based drug design. *Drug Discov. Today* 29 (7), 104024. doi:10.1016/j.drudis.2024.104024

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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.

- Balakin, K. V., Ivanenkov, Y. A., Savchuk, N. P., Ivashchenko, A. A., and Ekins, S. (2005). Comprehensive computational assessment of ADME properties using mapping techniques. *Curr. Drug Discov. Technol.* 2 (2), 99–113. doi:10.2174/1570163054064666
- Beck, A. G., Shrestha, R., Wang, J., Fine, J., Regalado, E. L., Hettiarachchi, K., et al. (2025). Purification of pharmaceuticals *via* retention time prediction: leveraging graph isomorphism networks, limited data, and transfer learning. *J. Sep. Sci.* 48 (6), e70178. doi:10.1002/jssc.70178
- Bengio, Y., Lahlou, S., Deleu, T., Hu, E. J., Tiwari, M., and Bengio, E. (2023). GFlowNet foundations. *J. Mach. Learn. Res.* 24 (1), 55. doi:10.5555/3648699.3648909
- Berenger, F., and Tsuda, K. (2021). Molecular generation by fast Assembly of (Deep)SMILES fragments. *J. Cheminform* 13 (1), 88. doi:10.1186/s13321-021-00566-4
- Bi, X., Wang, Y., Wang, J., and Liu, C. (2025). Machine learning for multi-target drug discovery: challenges and opportunities in systems pharmacology. *Pharmaceutics* 17 (9), 1186. doi:10.3390/pharmaceutics17091186
- Bian, Y., Wang, J., Jun, J. J., and Xie, X. Q. (2019). Deep convolutional generative adversarial network (dcGAN) models for screening and design of small molecules targeting cannabinoid receptors. *Mol. Pharm.* 16 (11), 4451–4460. doi:10.1021/acs.molpharmaceut.9b00500
- Bilsland, A. E., McAulay, K., West, R., Pugliese, A., and Bower, J. (2021). Automated generation of novel fragments using screening data, a dual SMILES autoencoder, transfer learning and syntax correction. *J. Chem. Inf. Model* 61 (6), 2547–2559. doi:10.1021/acs.jcim.0c01226
- Bronstein, M. M., Bruna, J., Lecun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric Deep learning: going beyond euclidean data. Institute of Electrical and Electronics Engineers Inc 34 (4), 18–42. doi:10.1109/MSP.2017.2693418
- Cai, Z., Zafferani, M., Akande, O. M., and Hargrove, A. E. (2022). Quantitative structure-activity relationship (QSAR) Study predicts small-molecule binding to RNA structure. *J. Med. Chem.* 65 (10), 7262–7277. doi:10.1021/acs.jmedchem.2c00254
- Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S. C., Endo, S., Fujii, K., et al. (2021). Variational quantum algorithms. *Nat. Rev. Phys.* 3 (9), 625–644. doi:10.1038/s42254-021-00348-9
- Chakraborty, A., Krishnan, V., and Thamotharan, S. (2025). Generative adversarial network (GAN) model-based design of potent SARS-CoV-2 m(pro) inhibitors using the electron density of ligands and 3D binding pockets: insights from molecular docking, dynamics simulation, and MM-GBSA analysis. *Mol. Divers* 29 (4), 3059–3075. doi:10.1007/s11030-024-11047-9

- Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discov. Today* 23 (6), 1241–1250. doi:10.1016/j.drudis.2018.01.039
- Chen, L., Tan, X., Wang, D., Zhong, F., Liu, X., Yang, T., et al. (2020). TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* 36 (16), 4406–4414. doi:10.1093/bioinformatics/btaa524
- Chen, Y., Wang, Y., Ding, Y., Su, X., and Wang, C. (2022). RGCNCDA: relational graph convolutional network improves circRNA-disease association prediction by incorporating microRNAs. *Comput. Biol. Med.* 143, 105322. doi:10.1016/j.combiomed.2022.105322
- Chen, Y., Wang, Z., Zeng, X., Li, Y., Li, P., Ye, X., et al. (2023a). Molecular language models: RNNs or transformer? *Brief. Funct. Genomics* 22 (4), 392–400. doi:10.1093/bfgp/elad012
- Chen, H., Lu, Y., Yang, Y., and Rao, Y. (2023b). A drug combination prediction framework based on graph convolutional network and heterogeneous information. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 20 (3), 1917–1925. doi:10.1109/TCBB.2022.3224734
- Chen, Z., Li, H., Zhang, C., Zhang, H., Zhao, Y., Cao, J., et al. (2024). Crystal structure prediction using generative adversarial network with data-driven latent space fusion strategy. *J. Chem. Theory Comput.* 20 (21), 9627–9641. doi:10.1021/acs.jctc.4c01096
- Chen, M., Zhang, M., Yan, G., Wang, G., and Qu, C. (2025). MRHGNN: enhanced multimodal relational Hypergraph neural network for synergistic drug combination forecasting. *IEEE Trans. Neural Netw. Learn. Syst.* 36 (9), 17086–17098. doi:10.1109/tnnls.2025.3553385
- Cho, T., Han, H. S., Jeong, J., Park, E. M., and Shim, K. S. (2021). A novel computational approach for the discovery of drug delivery system candidates for COVID-19. *Int. J. Mol. Sci.* 22 (6), 2815. doi:10.3390/ijms22062815
- Choppara, P., and Lokesh, B. (2025). A hybrid quantum classical approach for drug-target binding affinity prediction. *IEEE Trans. Comput. Biol. Bioinform* 22 (6), 2858–2871. doi:10.1109/tcbio.2025.3603103
- Chou, W. C., and Lin, Z. (2023). Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling. *Toxicol. Sci.* 191 (1), 1–14. doi:10.1093/toxsci/kfac101
- Cruz-Monteagudo, M., Cordeiro, M. N., and Borges, F. (2008). Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity. *J. Comput. Chem.* 29 (4), 533–549. doi:10.1002/jcc.20812
- Csermely, P., Agoston, V., and Pongor, S. (2005). The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol. Sci.* 26 (4), 178–182. doi:10.1016/j.tips.2005.02.007
- Cui, Y., Shan, D., Lu, Q., Zou, B., Zhang, H., Li, J., et al. (2025). Comparison study of dominant molecular sequence representation based on diffusion model. *J. Comput. Aided Mol. Des.* 39 (1), 54. doi:10.1007/s10822-025-00614-3
- Damani Shah, H., Saranath, D., Das, S., Kharkar, P., and Karande, A. (2019). *In-silico* identification of small molecules targeting H-Ras and *in-vitro* cytotoxicity with caspase-mediated apoptosis in carcinoma cells. *J. Cell Biochem.* 120 (4), 5519–5530. doi:10.1002/jcb.27836
- Das, B., Kutsal, M., and Das, R. (2022). A geometric deep learning model for display and prediction of potential drug-virus interactions against SARS-CoV-2. *Chemom. Intell. Lab. Syst.* 229, 104640. doi:10.1016/j.chemolab.2022.104640
- De Cao, N., and Kipf, T. (2018). MolGAN: an implicit generative model for small molecular graphs. *ArXiv* 1805, 11973. doi:10.48550/arXiv.1805.11973
- Descamps, C., Bouttier, V., Sanz García, J., Lhuillier-Akakpo, M., Perron, Q., and Tajmouati, H. (2025). Growing and linking optimizers: synthesis-driven molecule design. *Brief. Bioinform* 26 (5), bbaf482. doi:10.1093/bib/bbaf482
- Ding, L., and Spector, L. (2023). Multi-objective evolutionary architecture search for parameterized quantum circuits. *Entropy (Basel)* 25 (1), 93. doi:10.3390/e25010093
- Dobbelaere, M. R., Lengyel, L., Stevens, C. V., and Van Geem, K. M. (2024). Geometric deep learning for molecular property predictions with chemical accuracy across chemical space. *J. Cheminform* 16 (1), 99. doi:10.1186/s13321-024-00895-0
- Dogan, A. (2023). SELFormer: molecular representation learning via SELFIES language models. *Mach. Learn. Sci. Technol.* 4 (6), 1–20. doi:10.1088/2632-2153/aba947
- Du, X., Sun, X., and Li, M. (2024). Knowledge graph convolutional network with heuristic search for drug repositioning. *J. Chem. Inf. Model* 64 (12), 4928–4937. doi:10.1021/acs.jcim.4c00737
- Dunn, I., and Koes, D. R. (2025). FlowMol3: flow matching for 3D *de novo* small-molecule generation. *ArXiv* 2508, 12629. doi:10.48550/arXiv.2508.12629
- Eckmann, P., Anderson, J., Yu, R., and Gilson, M. K. (2024). Ligand-based compound activity prediction via few-shot learning. *J. Chem. Inf. Model* 64 (14), 5492–5499. doi:10.1021/acs.jcim.4c00485
- Einarson, K. A., Bendtsen, K. M., Li, K., Thomsen, M., Kristensen, N. R., Winther, O., et al. (2023). Molecular representations in machine-learning-based prediction of PK parameters for insulin analogs. *ACS Omega* 8 (26), 23566–23578. doi:10.1021/acsomega.3c01218
- Ekins, S. (2016). The next era: deep learning in pharmaceutical research. *Pharm. Res.* 33 (11), 2594–2603. doi:10.1007/s11095-016-2029-7
- Ekins, S., Lane, T. R., Urbina, F., and Puhl, A. C. (2024). *In silico* ADME/tox comes of age: twenty years later. *Xenobiotica* 54 (7), 352–358. doi:10.1080/00498254.2023.2245049
- Faber, F. A., Hutchison, L., Huang, B., Gilmer, J., Schoenholz, S. S., Dahl, G. E., et al. (2017). Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* 13 (11), 5255–5264. doi:10.1021/acs.jctc.7b00577
- Fan, X., Gong, M., Wu, Y., Qin, A. K., and Xie, Y. (2023). Propagation enhanced neural message passing for graph representation learning. *IEEE Trans. Knowl. Data Eng.* 35 (2), 1952–1964. doi:10.1109/TKDE.2021.3102964
- Farago, P. V., Camargo, G. D. A., Mendes, M. B., Semianko, B. C., Camilo Junior, A., Dias, D. T., et al. (2024). Computational simulation on the study of Tacrolimus and its improved dermal retention using Poly(ϵ -caprolactone) nanocapsules. *J. Mol. Graph Model* 126, 108625. doi:10.1016/j.jmgm.2023.108625
- Farghali, H., Kutinova Canova, N., and Arora, M. (2021). The potential applications of artificial intelligence in drug discovery and development. *Physiol. Res.* 70 (Suppl. 4), S715–S722. doi:10.33549/physiolres.934765
- Feng, Y., You, H., Zhang, Z., Ji, R., and Gao, Y. (2019). “Hypergraph neural networks,” *Proc. Thirty-Third AAAI Conf. Artif. Intell.*, 33. Honolulu, Hawaii, USA: AAAI Press, 437–3565. doi:10.1609/aaai.v33i01.33013558
- Gangwal, A., Ansari, A., Ahmad, I., Azad, A. K., Kumarasamy, V., Subramanian, V., et al. (2024). Generative artificial intelligence in drug discovery: basic framework, recent advances, challenges, and opportunities. *Front. Pharmacology* 15, 1331062. doi:10.3389/fphar.2024.1331062
- Gao, Z., Wang, X., Blumenfeld Gaines, B., Shi, X., Bi, J., and Song, M. (2023). Fragment-based deep molecular generation using hierarchical chemical graph representation and multi-resolution graph variational autoencoder. *Mol. Inf.* 42 (5), e2200215. doi:10.1002/minf.202200215
- Gao, Y., Zhang, X., Sun, Z., Chandak, P., Bu, J., and Wang, H. (2024). Precision adverse drug reactions prediction with heterogeneous graph neural network. *Adv. Sci. (Weinh)* 12 (4), e2404671. doi:10.1002/advs.202404671
- Gao, Q., Xu, T., Li, X., Gao, W., Shi, H., Zhang, Y., et al. (2025). Interpretable dynamic directed graph convolutional network for multi-relational prediction of missense mutation and drug response. *IEEE J. Biomed. Health Inf.* 29 (2), 1514–1524. doi:10.1109/jbhi.2024.3483316
- Garg, V. (2024). Generative AI for graph-based drug design: recent advances and the way forward. *Curr. Opin. Struct. Biol.* 84, 102769. doi:10.1016/j.sbi.2023.102769
- Gayathri, K. G. (2025). Machine learning and generative AI in the rational design of DNA gyrase-targeted antibacterials. *J. Mol. Graph Model* 142, 109178. doi:10.1016/j.jmgm.2025.109178
- Gini, G. (2022). QSAR methods. *Methods Mol. Biol.* 2425, 1–26. doi:10.1007/978-1-0716-1960-5_1
- Grisoni, F., Moret, M., Lingwood, R., and Schneider, G. (2020). Bidirectional molecule generation with recurrent neural networks. *J. Chem. Inf. Model* 60 (3), 1175–1183. doi:10.1021/acs.jcim.9b00943
- Grow, C., Gao, K., Nguyen, D. D., and Wei, G. W. (2019). Generative network complex (GNC) for drug discovery. *Commun. Inf. Syst.* 19 (3), 241–277. doi:10.4310/cis.2019.v19.n3.a2
- Guo, Y., Shen, Z., Zhao, W., Lu, J., Song, Y., Shen, L., et al. (2024). Rational identification of novel antibody-drug conjugate with high bystander killing effect against heterogeneous tumors. *Adv. Sci. (Weinh)* 11 (13), e2306309. doi:10.1002/advs.202306309
- Guo, F., Guan, R., Li, Y., Liu, Q., Wang, X., Yang, C., et al. (2025). Foundation models in bioinformatics. *Natl. Sci. Rev.* 12 (4), nwf028. doi:10.1093/nsr/nwf028
- Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Inductive representation learning on large graphs. *ArXiv* 1706, 02216. doi:10.48550/arXiv.1706.02216
- Han, S., Kang, Y., Park, H., Yi, J., Park, G., and Kim, J. (2024). Multimodal transformer for property prediction in polymers. *ACS Appl. Mater Interfaces* 16 (13), 16853–16860. doi:10.1021/acsami.4c01207
- Harris, J., Yadalam, P. K., Anegundi, R. V., and Arumuganainar, D. (2024). Comparing graph sample and aggregation (SAGE) and graph attention networks in the prediction of drug-gene associations of extended-spectrum beta-lactamases in periodontal infections and resistance. *Cureus* 16 (8), e68082. doi:10.7759/cureus.68082
- Henikoff, S., and Henikoff, J. G. (1997). Embedding strategies for effective use of information from multiple sequence alignments. *Protein Sci.* 6 (3), 698–705. doi:10.1002/pro.5560060319
- Hodgson, J. (2001). ADMET—turning chemicals into drugs. *Nat. Biotechnol.* 19 (8), 722–726. doi:10.1038/90761
- Hosseini, S. H., and Imani, M. (2025). Decentralized reinforcement learning for asymmetric gene network interventions. *IEEE Trans. Comput. Biol. Bioinform* 22 (6), 3524–3537. doi:10.1109/tcbio.2025.3633151
- Hou, Y., Wang, Z., Wang, W., Tang, Q., Cai, Y., Yu, S., et al. (2025). AI-identified CD133-targeting natural compounds demonstrate differential anti-tumor

- effects and mechanisms in pan-cancer models. *EMBO Mol. Med.* 17, 2932–2965. doi:10.1038/s44321-025-00308-1
- Hu, S., and Zhao, B. (2025). Protein function prediction using GO similarity-based heterogeneous network propagation. *Sci. Rep.* 15 (1), 19131. doi:10.1038/s41598-025-04933-1
- Hu, J. C., Cavicchioli, R., and Capotondi, A. (2025). Embeddings hidden layers learning for neural network compression. *Neural Netw.* 191, 107794. doi:10.1016/j.neunet.2025.107794
- Huang, D. Z., Baber, J. C., and Bahmanyar, S. S. (2021). The challenges of generalizability in artificial intelligence for ADME/Tox endpoint and activity prediction. *Expert Opin. Drug Discov.* 16 (9), 1045–1056. doi:10.1080/17460441.2021.1901685
- Huang, Y., Huang, H. Y., Chen, Y., Lin, Y. C., Yao, L., Lin, T., et al. (2023). A robust drug-target interaction prediction framework with capsule network and transfer learning. *Int. J. Mol. Sci.* 24 (18), 14061. doi:10.3390/ijms241814061
- Huang, S., Wang, M., Zheng, X., Chen, J., and Tang, C. (2024a). Hierarchical and dynamic graph attention network for drug-disease association prediction. *IEEE J. Biomed. Health Inf.* doi:10.1109/jbhi.2024.3363080
- Huang, K., Chandak, P., Wang, Q., Havaldar, S., Vaid, A., Leskovec, J., et al. (2024b). A foundation model for clinician-centered drug repurposing. *Nat. Med.* 30 (12), 3601–3613. doi:10.1038/s41591-024-03233-x
- Huang, Z., Zhou, G., Qiu, Y., Chen, X., and Zhao, Q. (2025). Kernel bayesian tensor ring decomposition for multiway data recovery. *Neural Netw.* 189, 107500. doi:10.1016/j.neunet.2025.107500
- Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. (2011). Principles of early drug discovery. *Br. J. Pharmacol.* 162 (6), 1239–1249. doi:10.1111/j.1476-5381.2010.01127.x
- Isert, C., Kromann, J. C., Stiefl, N., Schneider, G., and Lewis, R. A. (2023). Machine learning for fast, quantum mechanics-based approximation of drug lipophilicity. *ACS Omega* 8 (2), 2046–2056. doi:10.1021/acscomega.2c05607
- Itoh, T. D., Kubo, T., and Ikeda, K. (2022). Multi-level attention pooling for graph neural networks: unifying graph representations with multiple localities. *Neural Netw.* 145, 356–373. doi:10.1016/j.neunet.2021.11.001
- Jaeger, M. (2023). Learning and reasoning with graph data. *Front. Artif. Intell.* 6, 1124718. doi:10.3389/frai.2023.1124718
- Jiang, Z., Ding, P., Shen, C., and Dai, X. (2024). Geometric molecular graph representation learning model for drug-drug interactions prediction. *IEEE J. Biomed. Health Inf.* 28 (12), 7623–7632. doi:10.1109/jbhi.2024.3453956
- Jin, C., Guo, S., Zhou, S., and Guan, J. (2025). Effective and explainable molecular property prediction by chain-of-thought enabled large language models and multimodal molecular information fusion. *J. Chem. Inf. Model* 65 (11), 5438–5455. doi:10.1021/acs.jcim.5c00577
- Jorgensen, W. L., and Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* 110 (6), 1657–1666. doi:10.1021/ja00214a001
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2
- Kaneko, H. (2023). Molecular descriptors, structure generation, and inverse QSAR/QSPR based on SELFIES. *ACS Omega* 8 (24), 21781–21786. doi:10.1021/acscomega.3c01332
- Karpov, P., Godin, G., and Tetko, I. V. (2020). Transformer-CNN: swiss knife for QSAR modeling and interpretation. *J. Cheminform* 12 (1), 17. doi:10.1186/s13321-020-00423-w
- Khan, P. M., and Roy, K. (2018). Current approaches for choosing feature selection and learning algorithms in quantitative structure-activity relationships (QSAR). *Expert Opin. Drug Discov.* 13 (12), 1075–1089. doi:10.1080/17460441.2018.1542428
- Kondor, R., and Trivedi, S. (2018). “On the generalization of equivariance and convolution in neural networks to the action of compact groups,” in *Proceedings of the 35th international conference on machine learning* (Proceedings of Machine Learning Research).
- Korn, D., Thieme, A. J., Alves, V. M., Yeakey, M., Borba, J., Capuzzi, S. J., et al. (2022). Defining clinical outcome pathways. *Drug Discov. Today* 27 (6), 1671–1678. doi:10.1016/j.drudis.2022.02.008
- Kostal, J. (2023). Making the case for quantum mechanics in predictive toxicology—nearly 100 years too late? *Chem. Res. Toxicol.* 36 (9), 1444–1450. doi:10.1021/acs.chemrestox.3c00171
- Krenn, M., Hase, F., Nigam, A. K., Friederich, P., and Aspuru-Guzik, A. (2020). Self-referencing Embedded Strings (SELFIES): a 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* 1905, 13741. doi:10.48550/arXiv.1905.13741
- Krenn, M., Ai, Q., Barthel, S., Carson, N., Frei, A., Frey, N. C., et al. (2022). SELFIES and the future of molecular string representations. *Patterns (N Y)* 3 (10), 100588. doi:10.1016/j.patter.2022.100588
- Krishnan, A., Anahtar, M. N., Valeri, J. A., Jin, W., Donghia, N. M., Sieben, L., et al. (2025). A generative deep learning approach to *de novo* antibiotic design. *Cell* 188 (21), 5962–5979.e5922. doi:10.1016/j.cell.2025.07.033
- Kumar, A., Bharadwaj, T., Muthuraj, L., Kumar, J., Kumar, P., Lalitha, R., et al. (2025). Molecular dynamics simulation and docking studies reveals inhibition of NF- κ B signaling as a promising therapeutic drug target for reduction in cytokines storms. *Sci. Rep.* 15 (1), 15225. doi:10.1038/s41598-024-78411-5
- Lackman-Mincoff, J., Jain, M., Malkin, N., Bengio, Y., and Simine, L. (2024). Path-filtering in path-integral simulations of open quantum systems using GFlowNets. *J. Chem. Phys.* 161 (14), 144106. doi:10.1063/5.0226408
- Lavecchia, A. (2019). Deep learning in drug discovery: opportunities, challenges and future prospects. *Drug Discov. Today* 24 (10), 2017–2032. doi:10.1016/j.drudis.2019.07.006
- Lee, T., and Posma, J. (2025). Improving drug-induced liver injury prediction using graph neural networks with augmented graph features from molecular optimisation. *J. Cheminformatics* 17 (1), 124. doi:10.1186/s13321-025-01068-3
- Li, B., and Chen, H. (2022). Prediction of compound synthesis accessibility based on reaction knowledge graph. *Molecules* 27 (3), 1039. doi:10.3390/molecules27031039
- Li, C., and Li, G. (2025). DynHeter-DTA: dynamic heterogeneous graph representation for drug-target binding affinity prediction. *Int. J. Mol. Sci.* 26 (3), 1223. doi:10.3390/ijms26031223
- Li, J., Wang, J., Lv, H., Zhang, Z., and Wang, Z. (2022). IMCHGAN: inductive matrix completion with heterogeneous graph attention networks for drug-target interactions prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform* 19 (2), 655–665. doi:10.1109/tcbb.2021.3088614
- Liang, Y., Peng, W., Zheng, Z. J., Silvén, O., and Zhao, G. (2021). A hybrid quantum-classical neural network with deep residual learning. *Neural Netw.* 143, 133–147. doi:10.1016/j.neunet.2021.05.028
- Liu, M., and Yang, X. (2025). Patient-derived xenograft models: current status, challenges, and innovations in cancer research. *Genes & Dis.* 12 (5), 101520. doi:10.1016/j.gendis
- Liu, X., Ye, K., van Vlijmen, H. W. T., Ap, I. J., and van Westen, G. J. P. (2019). An exploration strategy improves the diversity of *de novo* ligands using deep reinforcement learning: a case for the adenosine A(2A) receptor. *J. Cheminform* 11 (1), 35. doi:10.1186/s13321-019-0355-6
- Liu, X., Luo, Y., Li, P., Song, S., and Peng, J. (2021). Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLoS Comput. Biol.* 17 (8), e1009284. doi:10.1371/journal.pcbi.1009284
- Liu, X., Song, C., Liu, S., Li, M., Zhou, X., and Zhang, W. (2022). Multi-way relation-enhanced hypergraph representation learning for anti-cancer drug synergy prediction. *Bioinformatics* 38 (20), 4782–4789. doi:10.1093/bioinformatics/btac579
- Liu, X., Zhang, W., Tong, X., Zhong, F., Li, Z., Xiong, Z., et al. (2023a). MolFilterGAN: a progressively augmented generative adversarial network for triaging AI-designed molecules. *J. Cheminform* 15 (1), 42. doi:10.1186/s13321-023-00711-1
- Liu, J., Li, B., Zhang, Y., Zhang, L., Huang, S., Sun, H., et al. (2023b). A high-fidelity geometric multiscale hemodynamic model for predicting myocardial ischemia. *Comput. Methods Programs Biomed.* 233, 107476. doi:10.1016/j.cmpb.2023.107476
- Liu, S., Chen, M., Yao, X., and Liu, H. (2025). Fingerprint-enhanced hierarchical molecular graph neural networks for property prediction. *J. Pharm. Anal.* 15 (6), 101242. doi:10.1016/j.jpha.2025.101242
- Lo, A., Pollice, R., Nigam, A., White, A. D., Krenn, M., and Aspuru-Guzik, A. (2023). Recent advances in the self-referencing embedded strings (SELFIES) library. *Digit. Discov.* 2 (4), 897–908. doi:10.1039/d3dd00044c
- Loeffler, H. H., He, J., Tibo, A., Janet, J. P., Voronov, A., Mervin, L. H., et al. (2024). Reinvent 4: modern AI-driven generative molecule design. *J. Cheminformatics* 16 (1), 20. doi:10.1186/s13321-024-00812-5
- Long, Y., Wu, M., Kwok, C. K., Luo, J., and Li, X. (2020). Predicting human microbe-drug associations via graph convolutional network with conditional random field. *Bioinformatics* 36 (19), 4918–4927. doi:10.1093/bioinformatics/btaa598
- Lu, Q., Zhang, R., Zhou, H., Ni, D., Xiao, W., and Li, J. (2023). MetaHMEI: meta-learning for prediction of few-shot histone modifying enzyme inhibitors. *Brief. Bioinform* 24 (3), bbad115. doi:10.1093/bib/bbad115
- Lu, H., Wei, Z., Liu, J., Li, J., Wang, Q., and Liu, H. (2025a). Designing high-affinity 3D drug molecules via geometric spatial perception diffusion model. *Brief. Bioinform* 26 (5), bba542. doi:10.1093/bib/bba542
- Lu, J., Wu, L., Li, R., Wan, M., Yang, J., Zan, P., et al. (2025b). ToxAcCoL: an endpoint-aware and task-focused compound representation learning paradigm for acute toxicity assessment. *Nat. Commun.* 16 (1), 5992. doi:10.1038/s41467-025-60989-7
- Luo, Y., and Deng, L. (2025). MolCL-SP: a multimodal contrastive learning framework with non-overlapping substructure perturbations for molecular property prediction. *Bioinformatics* 41 (10), btaf507. doi:10.1093/bioinformatics/btaf507
- Luo, Y., Liu, Y., and Peng, J. (2023). Calibrated geometric deep learning improves kinase-drug binding predictions. *Nat. Mach. Intell.* 5 (12), 1390–1401. doi:10.1038/s42256-023-00751-0
- Luo, S., Li, Y., Liu, S., Zhang, X., Shao, Y., and Wu, C. (2024). Multi-agent continuous control with generative flow networks. *Neural Netw.* 174, 106243. doi:10.1016/j.neunet.2024.106243

- Luo, J., Zhu, Z., Xu, Z., Xiao, C., Wei, J., and Shen, J. (2025a). GS-DTA: integrating graph and sequence models for predicting drug-target binding affinity. *BMC Genomics* 26 (1), 105. doi:10.1186/s12864-025-11234-4
- Luo, D., Zhou, J., Xu, L., Yuan, S., and Lin, X. (2025b). DynamicDTA: drug-Target binding affinity prediction using dynamic descriptors and graph representation. *Interdiscip. Sci.* doi:10.1007/s12539-025-00729-z
- Ma, T., Liu, Q., Li, H., Zhou, M., Jiang, R., and Zhang, X. (2022). DualGCN: a dual graph convolutional network model to predict cancer drug response. *BMC Bioinforma.* 23 (Suppl. 4), 129. doi:10.1186/s12859-022-04664-4
- Macedo, B., Ribeiro Vaz, I., and Taveira Gomes, T. (2024). MedGAN: optimized generative adversarial network with graph convolutional networks for novel molecule design. *Sci. Rep.* 14 (1), 1212. doi:10.1038/s41598-023-50834-6
- Man, H., Bao, H., Niu, Z., Zhang, Z., Simon, J. P., Yang, T., et al. (2025). Multimodal profiling of Peppan-CB1 receptor structure-activity relationships: integrating molecular dynamics simulations, biological profiling, and the deep learning model MuMoPeppan. *Bioorg Chem.* 165, 109027. doi:10.1016/j.bioorg.2025.109027
- Manu, D., Yao, J., Liu, W., and Sun, X. (2024). GraphGANFed: a federated generative framework for graph-structured molecules towards efficient drug discovery. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 21 (2), 240–253. doi:10.1109/TCBB.2024.3349990
- Martinelli, D. D. (2022). Generative machine learning for *de novo* drug discovery: a systematic review. *Comput. Biol. Med.* 145, 105403. doi:10.1016/j.compbiomed.2022.105403
- Martinez Molina, D., Jafari, R., Ignatshchenko, M., Seki, T., Larsson, E. A., Dan, C., et al. (2013). Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay. *Science* 341 (6141), 84–87. doi:10.1126/science.1233606
- Mastrolorito, F., Ciriaco, F., Togo, M. V., Gambacorta, N., Trisciuzzi, D., Altomare, C. D., et al. (2025). fragSMILES as a chemical string notation for advanced fragment and chirality representation. *Commun. Chem.* 8 (1), 26. doi:10.1038/s42004-025-01423-3
- Matsuzaka, Y., and Uesawa, Y. (2023). Ensemble learning, deep learning-based and molecular descriptor-based quantitative structure-activity relationships. *Molecules* 28 (5), 2410. doi:10.3390/molecules28052410
- Mazraedost, S., Sedigh Malekroodi, H., Žuvela, P., Yi, M., and Liu, J. J. (2025). Prediction of chromatographic retention time of a small molecule from SMILES representation using a hybrid Transformer-LSTM model. *J. Chem. Inf. Model* 65 (7), 3343–3356. doi:10.1021/acs.jcim.5c00167
- Meyers, J., Fabian, B., and Brown, N. (2021). *De novo* molecular design and generative models. *Drug Discov. Today* 26 (11), 2707–2715. doi:10.1016/j.drudis.2021.05.019
- Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., et al. (2023). Foundation models for generalist medical artificial intelligence. *Nature* 616 (7956), 259–265. doi:10.1038/s41586-023-05881-4
- Morehead, A., and Cheng, J. (2024). Geometry-complete perceptron networks for 3D molecular graphs. *Bioinformatics* 40 (2), btac087. doi:10.1093/bioinformatics/btac087
- Mswahili, M. E., Hwang, J., Rajapakse, J. C., Jo, K., and Jeong, Y. S. (2025a). Positional embeddings and zero-shot learning using BERT for molecular-property prediction. *J. Cheminform* 17 (1), 17. doi:10.1186/s13321-025-00959-9
- Mswahili, M. E., Ndomba, G. E., Kim, Y. J., Jo, K., and Jeong, Y. S. (2025b). Relational graph convolution network with multi features for Anti- COVID-19 drugs discovery using 3CLpro potential target. *Curr. Bioinforma.* 20 (1), 13–30. doi:10.2174/0115748936280392240219054047
- Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., et al. (2020). QSAR without borders. *Chem. Soc. Rev.* 49 (11), 3525–3564. doi:10.1039/d0cs00098a
- Myung, Y., de Sá, A. G. C., and Ascher, D. B. (2024). Deep-PK: deep learning for small molecule pharmacokinetic and toxicity prediction. *Nucleic Acids Res.* 52 (W1), W469–w475. doi:10.1093/nar/gkae254
- Nguyen, T., and Karolak, A. (2025). Transformer graph variational autoencoder for generative molecular design. *Biophys. J.* 124, 3867–3875. doi:10.1016/j.bpj.2025.01.022
- Nguyen, D. A., Nguyen, C. H., and Mamitsuka, H. (2024). Central-smoothing hypergraph neural networks for predicting drug-drug interactions. *IEEE Trans. Neural Netw. Learn. Syst.* 35 (8), 11620–11625. doi:10.1109/tnnls.2023.3261860
- Niazi, S. K. (2025). The quantum paradox in pharmaceutical science: understanding without Comprehending-A Centennial reflection. *Int. J. Mol. Sci.* 26 (10), 4658. doi:10.3390/ijms26104658
- Niazi, S. K. (2026). Quantum mechanics in drug design: progress, challenges, and future frontiers. *Commun. Integr. Biol.* 19 (1), 2603140. doi:10.1080/19420889.2025.2603140
- Noor, F., Junaid, M., Almalki, A. H., Almaghrabi, M., Ghazanfar, S., and Tahir ul Qamar, M. (2024). Deep learning pipeline for accelerating virtual screening in drug discovery. *Sci. Rep.* 14 (1), 28321. doi:10.1038/s41598-024-79799-w
- Nowak, D., Bachorz, R. A., and Hoffmann, M. (2023). Neural networks in the design of molecules with affinity to selected protein domains. *Int. J. Mol. Sci.* 24 (2), 1762. doi:10.3390/ijms24021762
- O'Boyle, N. M., and Dalke, A. (2018). DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures. *ChemRxiv.* doi:10.26434/chemrxiv.7097960.v1
- Ochiai, T., Inukai, T., Akiyama, M., Furui, K., Ohue, M., Matsumori, N., et al. (2023). Variational autoencoder-based chemical latent space for large molecular structures with 3D complexity. *Commun. Chem.* 6 (1), 249. doi:10.1038/s42004-023-01054-6
- Olehnovics, E., Liu, Y. M., Mehio, N., Sheikh, A. Y., Shirts, M. R., and Salvaglio, M. (2024). Assessing the accuracy and efficiency of free energy differences obtained from reweighted flow-based probabilistic generative models. *J. Chem. Theory Comput.* 20 (14), 5913–5922. doi:10.1021/acs.jctc.4c00520
- Ouyang, D., Liang, Y., Wang, J., Liu, X., Xie, S., Miao, R., et al. (2022). Predicting multiple types of miRNA-disease associations using adaptive weighted nonnegative tensor factorization with self-paced learning and hypergraph regularization. *Brief. Bioinform* 23 (6), bbac390. doi:10.1093/bib/bbac390
- Pallikkavaliyaveetil, N., and Chandrasekaran, S. (2026). Small data, big challenges: machine- and deep-learning strategies for data-limited drug discovery. *Adv. Drug Deliv. Rev.* 229, 115762. doi:10.1016/j.addr.2025.115762
- Parvez, M. A., and Mehedi, I. M. (2025). High-accuracy polymer property detection via pareto-Optimized SMILES-Based deep learning. *Polym. (Basel)* 17 (13), 1801. doi:10.3390/polym17131801
- Patel, D., Koch, D., Patel, S., and Wilde, M. M. (2024). *Quantum boltzmann machine learning of ground-state energies.* *arXiv Preprint arXiv:2410.12935.* doi:10.48550/arXiv.2410.12935
- Pazhanivel, D. B., Velu, A. N., and Palaniappan, B. S. (2024). Design and enhancement of a fog-enabled air quality monitoring and prediction system: an optimized lightweight deep learning model for a smart fog environmental gateway. *Sensors (Basel)* 24 (15), 5069. doi:10.3390/s24155069
- Qiu, Z., Xie, Z., Ji, Z., Mao, Y., and Cheng, K. (2024). HGSMAP: a novel heterogeneous graph-based associative percept framework for scenario-based optimal model assignment. *Knowl. Inf. Syst.* (1), 915–952. doi:10.1007/s10115-024-02251-y
- Rupp, M., Tkatchenko, A., Müller, K. R., and von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* 108 (5), 058301. doi:10.1103/PhysRevLett.108.058301
- Scannell, J. W., Blanckley, A., Boldon, H., and Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* 11 (3), 191–200. doi:10.1038/nrd3681
- Shao, K., Zhang, Y., Wen, Y., Zhang, Z., He, S., and Bo, X. (2022). DTI-HETA: prediction of drug-target interactions based on GCN and GAT on heterogeneous graph. *Brief. Bioinform* 23 (3), bbac109. doi:10.1093/bib/bbac109
- Sharma, S., Singh, R., Kant, S., and Mishra, M. K. (2025). Integrating AI/ML and multi-omics approaches to investigate the role of TNFRSF10A/TRAILR1 and its potential targets in pancreatic cancer. *Comput. Biol. Med.* 193, 110432. doi:10.1016/j.compbiomed.2025.110432
- Shen, C., Luo, J., and Xia, K. (2023). Molecular geometric deep learning. *Cell Rep. Methods* 3 (11), 100621. doi:10.1016/j.crmeth.2023.100621
- Shi, H., Liu, S., Chen, J., Li, X., Ma, Q., and Yu, B. (2019). Predicting drug-target interactions using lasso with random forest based on evolutionary information and chemical structure. *Genomics* 111 (6), 1839–1852. doi:10.1016/j.ygeno.2018.12.007
- Shin, H. K. (2021). Topological distance-based electron interaction tensor to apply a convolutional neural network on drug-like compounds. *ACS Omega* 6 (51), 35757–35768. doi:10.1021/acsomega.1c05693
- Shui, Y., Ge, X., Cao, C., Wang, J., Hu, J., and Liu, Y. (2025). BiMA-DTI: a bidirectional mamba-attention hybrid framework for enhanced drug-target interaction prediction. *BMC Biol.* 23 (1), 309. doi:10.1186/s12915-025-02407-4
- Simonovsky, M., and Komodakis, N. (2018). GraphVAE: towards generation of small graphs using variational autoencoders. *Springer, Cham* 9, 412–422. doi:10.1007/978-3-030-01418-6_41
- Skinnider, M. A. (2024). Invalid SMILES are beneficial rather than detrimental to chemical language models. *Nat. Mach. Intell.* 6 (4), 437–448. doi:10.1038/s42256-024-00821-x
- Smaldone, A. M., and Batista, V. S. (2024). Quantum-to-Classical neural network transfer learning applied to drug toxicity prediction. *J. Chem. Theory Comput.* 20 (11), 4901–4908. doi:10.1021/acs.jctc.4c00432
- Sousa, T., Correia, J., Pereira, V., and Rocha, M. (2021). Generative deep learning for targeted compound design. *J. Chem. Inf. Model* 61 (11), 5343–5361. doi:10.1021/acs.jcim.0c01496
- Strandgaard, M., Linjordet, T., Kneiding, H., Burnage, A. L., Nova, A., Jensen, J. H., et al. (2025). A deep generative model for the inverse design of transition metal ligands and complexes. *JACS Au* 5 (5), 2294–2308. doi:10.1021/jacsau.5c00242
- Surya Prakash, S., Banu Priya, P., Somneni, S., and Banerjee, A. (2025). *Drug discovery using variational quantum EigenSolver.* Singapore: Springer Nature Singapore, 691–703. doi:10.1007/978-981-96-3942-7_51
- Swaan, P. W., and Ekins, S. (2005). Reengineering the pharmaceutical industry by crash-testing molecules. *Drug Discov. Today* 10 (17), 1191–1200. doi:10.1016/S1359-6446(05)03557-9
- Tajiani, F., Ahmadi, S., Lotfi, S., Kumar, P., and Almasirad, A. (2023). *In-silico* activity prediction and docking studies of some flavonol derivatives as anti-prostate cancer

- agents based on monte carlo optimization. *BMC Chem.* 17 (1), 87. doi:10.1186/s13065-023-00999-y
- Talkington, A. M., Cao, Y., Kearsley, A. J., and Lai, S. K. (2025). Opportunities for machine learning and artificial intelligence in physiologically-based pharmacokinetic (PBPK) modeling. *Adv. Drug Deliv. Rev.* 227, 115716. doi:10.1016/j.addr.2025.115716
- Tan, H., Ji, X., Xu, C. Z., Zhao, X., Hou, J., Liu, M., et al. (2025). AGRL-DSE: adaptive graph representation learning on a heterogeneous graph for drug side effect prediction. *ACS Omega* 10 (34), 38753–38765. doi:10.1021/acsomega.5c04006
- Tang, M., Li, B., and Chen, H. (2023a). Application of message passing neural networks for molecular property prediction. *Curr. Opin. Struct. Biol.* 81, 102616. doi:10.1016/j.sbi.2023.102616
- Tang, R., Sun, C., Huang, J., Li, M., Wei, J., and Liu, J. (2023b). Predicting drug-protein interactions by self-adaptively adjusting the topological structure of the heterogeneous network. *IEEE J. Biomed. Health Inf.* 27 (11), 5675–5684. doi:10.1109/jbhi.2023.3312374
- Tang, Y., Li, Y., Li, P., and Liu, Z. P. (2025). Drug-target affinity prediction by molecule secondary structure representation network. *Curr. Med. Chem.* 32 (24), 5095–5105. doi:10.2174/0109298673252287240215103035
- Tetko, I. V., Karpov, P., Van Deursen, R., and Godin, G. (2020). State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat. Commun.* 11 (1), 5575. doi:10.1038/s41467-020-19266-y
- Thompson, L. A., Evans, J. G., and Matthews, S. T. (2025). AmesFormer: state-of-the-art mutagenicity prediction with graph transformers. *Chem. Res. Toxicol.* 38 (7), 1167–1182. doi:10.1021/acs.chemrestox.4c00466
- Tiapkın, D., Morozov, N., Naumov, A., and Vetrov, D. P. (2024). “Generative flow networks as entropy-regularized RL,” in *Proceedings of the 27th international conference on artificial intelligence and statistics, proceedings of machine learning research*.
- Tom, G., Schmid, S. P., Baird, S. G., Cao, Y., Darvish, K., Hao, H., et al. (2024). Self-driving laboratories for chemistry and materials science. *Chem. Rev.* 124 (16), 9633–9732. doi:10.1021/acs.chemrev.4c00055
- Tropsha, A., Isayev, O., Varnek, A., Schneider, G., and Cherkasov, A. (2024). Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. *Nat. Rev. Drug Discov.* 23 (2), 141–155. doi:10.1038/s41573-023-00832-0
- Ucak, U. V., Ashrymamatov, I., and Lee, J. (2023). Reconstruction of lossless molecular representations from fingerprints. *J. Cheminform* 15 (1), 26. doi:10.1186/s13321-023-00693-0
- Vargas, S., Gee, W., and Alexandrova, A. (2024). High-throughput quantum theory of atoms in molecules (QTAIM) for geometric deep learning of molecular and reaction properties. *Digit. Discov.* 3 (5), 987–998. doi:10.1039/d4dd00089j
- Wang, T., Wu, M. B., Lin, J. P., and Yang, L. R. (2015). Quantitative structure-activity relationship: promising advances in drug discovery platforms. *Expert Opin. Drug Discov.* 10 (12), 1283–1300. doi:10.1517/17460441.2015.1083006
- Wang, L., You, Z. H., Li, L. P., Yan, X., Zhang, W., Song, K. J., et al. (2020). Identification of potential drug-targets by combining evolutionary information extracted from frequency profiles and molecular topological structures. *Chem. Biol. Drug Des.* 96 (2), 758–767. doi:10.1111/cbdd.13599
- Wang, F., Lei, X., Liao, B., and Wu, F. X. (2022a). Predicting drug-drug interactions by graph convolutional network with multi-kernel. *Brief. Bioinform* 23 (1), bbab511. doi:10.1093/bib/bbab511
- Wang, Y., Hu, L., Wu, Y., and Gao, W. (2022b). Graph multihead attention pooling with self-supervised learning. *Entropy (Basel)* 24 (12), 1745. doi:10.3390/e24121745
- Wang, L., Wong, L., Chen, Z. H., Hu, J., Sun, X. F., Li, Y., et al. (2022c). MSPEDTI: prediction of drug-target interactions via molecular structure with protein evolutionary information. *Biol. (Basel)* 11 (5), 740. doi:10.3390/biology11050740
- Wang, Y., Michael, S., Yang, S. M., Huang, R., Cruz-Gutierrez, K., Zhang, Y., et al. (2022d). Retro drug design: from target properties to molecular structures. *J. Chem. Inf. Model* 62 (11), 2659–2669. doi:10.1021/acs.jcim.2c00123
- Wang, T., Li, Z., Zhuo, L., Chen, Y., Fu, X., and Zou, Q. (2024). MS-BACL: enhancing metabolic stability prediction through bond graph augmentation and contrastive learning. *Brief. Bioinform* 25 (3), bbae127. doi:10.1093/bib/bbae127
- Wang, Y., Guo, M., Chen, X., and Ai, D. (2025). Screening of multi deep learning-based *de novo* molecular generation models and their application for specific target molecular generation. *Sci. Rep.* 15 (1), 4419. doi:10.1038/s41598-025-86840-z
- Wouters, O. J., McKee, M., and Luyten, J. (2020). Estimated research and development investment needed to bring a new medicine to market, 2009–2018. *Jama* 323 (9), 844–853. doi:10.1001/jama.2020.1166
- Wu, Y., Li, K., Li, M., Pu, X., and Guo, Y. (2023a). Attention mechanism-based graph neural network model for effective activity prediction of SARS-CoV-2 main protease inhibitors: application to drug repurposing as potential COVID-19 therapy. *J. Chem. Inf. Model* 63 (22), 7011–7031. doi:10.1021/acs.jcim.3c01280
- Wu, F., Wu, L., Radev, D., Xu, J., and Li, S. Z. (2023b). Integration of pre-trained protein language models into geometric deep learning networks. *Commun. Biol.* 6 (1), 876. doi:10.1038/s42003-023-05133-1
- Wu, J. N., Wang, T., Chen, Y., Tang, L. J., Wu, H. L., and Yu, R. Q. (2024a). t-SMILES: a fragment-based molecular representation framework for *de novo* ligand design. *Nat. Commun.* 15 (1), 4993. doi:10.1038/s41467-024-49388-6
- Wu, X., Hou, W., Zhao, Z., Huang, L., Sheng, N., Yang, Q., et al. (2024b). MMGAT: a graph attention network framework for ATAC-seq motifs finding. *BMC Bioinforma.* 25 (1), 158. doi:10.1186/s12859-024-05774-x
- Wu, K., Yang, X., Wang, Z., Li, N., Zhang, J., and Liu, L. (2024c). Data-balanced transformer for accelerated ionizable lipid nanoparticles screening in mRNA delivery. *Brief. Bioinform* 25 (3), bbae186. doi:10.1093/bib/bbae186
- Wu, K., Wang, Z., Yang, X., Chen, Y., Mastrogianni, F., Zhang, J., et al. (2025). TransMA: an explainable multi-modal deep learning model for predicting properties of ionizable lipid nanoparticles in mRNA delivery. *Brief. Bioinform* 26 (3), bbaf307. doi:10.1093/bib/bbaf307
- Xiao, J., Hu, G., Zhou, X., Zheng, Y., and Li, J. (2025). TIDGN: a transfer learning framework for predicting interactions of intrinsically disordered proteins with high conformational dynamics. *J. Chem. Inf. Model.* 65 (10), 4866–4877. doi:10.1021/acs.jcim.5c00422
- Xiong, Y., Wang, Y., Wang, Y., Li, C., Yusong, P., Wu, J., et al. (2023). Improving drug discovery with a hybrid deep generative model using reinforcement learning trained on a Bayesian docking approximation. *J. Comput. Aided Mol. Des.* 37 (11), 507–517. doi:10.1007/s10822-023-00523-3
- Xu, P., Wei, Z., Li, C., Yuan, J., Liu, Z., and Liu, W. (2024). Drug-target prediction based on dynamic heterogeneous graph convolutional network. *IEEE J. Biomed. Health Inf.* 28 (11), 6997–7005. doi:10.1109/jbhi.2024.3441324
- Xuan, P., Gu, J., Cui, H., Wang, S., Toshiya, N., Liu, C., et al. (2024). Multi-scale topology and position feature learning and relationship-aware graph reasoning for prediction of drug-related microbes. *Bioinformatics* 40 (2), btae025. doi:10.1093/bioinformatics/btae025
- Yakubovich, A., Odinkov, A., Nikolenko, S., Jung, Y., and Choi, H. (2021). Computational discovery of TTF molecules with deep generative models. *Front. Chem.* 9, 800133. doi:10.3389/fchem.2021.800133
- Yang, Z., Zhong, W., Zhao, L., and Chen, Y.-C. (2022). MGraphDTA: deep multiscale graph neural network for explainable drug-target binding affinity prediction. *Chem. Sci.* 13 (3), 816–833. doi:10.1039/d1sc05180f
- Yang, H., Liu, J., Chen, K., Cong, S., Cai, S., Li, Y., et al. (2024). D-CyPre: a machine learning-based tool for accurate prediction of human CYP450 enzyme metabolic sites. *PeerJ Comput. Sci.* 10, e2040. doi:10.7717/peerj-cs.2040
- Yi, J. C., Yang, Z. Y., Zhao, W. T., Yang, Z. J., Zhang, X. C., Wu, C. K., et al. (2024). ChemMORT: an automatic ADMET optimization platform using deep learning and multi-objective particle swarm optimization. *Brief. Bioinform* 25 (2), bbae008. doi:10.1093/bib/bbae008
- Yi, J., Jiang, D., Wu, C., Zhang, X., He, W., Zhao, W., et al. (2025). Pushing the boundaries of few-shot learning for low-data drug discovery with a Bayesian meta-learning hypernetwork framework. *Brief. Bioinform* 26 (4), bbaf408. doi:10.1093/bib/bbaf408
- Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., and Leskovec, J. (2018). “Graph convolutional neural networks for web-scale recommender systems,” in *In proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 974–983.
- Yu, J., Yin, H., Li, J., Wang, Q., Hung, N. Q. V., and Zhang, X. (2021). Self-supervised multi-channel hypergraph convolutional network for social recommendation. *The web conference*, 413–424. doi:10.1145/3442381.3449844
- Yu, Q., Zhang, Z., Liu, G., Li, W., and Tang, Y. (2024). ToxGIN: an *in silico* prediction model for peptide toxicity via graph isomorphism networks integrating peptide sequence and structure information. *Brief. Bioinform* 25 (6), bbae583. doi:10.1093/bib/bbae583
- Yuan, H., Huang, J., and Li, J. (2021). Protein-ligand binding affinity prediction model based on graph attention network. *Math. Biosci. Eng.* 18 (6), 9148–9162. doi:10.3934/mbe.2021451
- Zhai, H., Hou, H., Luo, J., Liu, X., Wu, Z., and Wang, J. (2023). DGDTA: dynamic graph attention network for predicting drug-target binding affinity. *BMC Bioinforma.* 24 (1), 367. doi:10.1186/s12859-023-05497-5
- Zhang, S., Jiang, M., Wang, S., Wang, X., Wei, Z., and Li, Z. (2021). SAG-DTA: prediction of drug-target affinity using self-attention graph network. *Int. J. Mol. Sci.* 22 (16), 8993. doi:10.3390/ijms22168993
- Zhang, P., Wei, Z., Che, C., and Jin, B. D. M. G. T.-D. T. I. (2022). Transformer network incorporating multilayer graph information for drug-target interaction prediction. *Comput. Biol. Med.* 142, 105214. doi:10.1016/j.combiomed
- Zhang, K., Wu, M., Liu, Y., Feng, Y., and Zheng, J. (2023a). KR4SL: knowledge graph reasoning for explainable prediction of synthetic lethality. *Bioinformatics* 39 (39 Suppl. 1), i158–i167. doi:10.1093/bioinformatics/btad261
- Zhang, Y., Hu, Y., Han, N., Yang, A., Liu, X., and Cai, H. (2023b). A survey of drug-target interaction and affinity prediction methods via graph neural networks. *Comput. Biol. Med.* 163, 107136. doi:10.1016/j.combiomed.2023.107136
- Zhang, R., Wu, C., Yang, Q., Liu, C., Wang, Y., Li, K., et al. (2024). MolFeSCue: enhancing molecular property prediction in data-limited and imbalanced

- contexts using few-shot and contrastive learning. *Bioinformatics* 40 (4), btac118. doi:10.1093/bioinformatics/btac118
- Zhang, Y., Wu, J., Kang, Y., and Hou, T. (2025a). A multimodal contrastive learning framework for predicting P-glycoprotein substrates and inhibitors. *J. Pharm. Anal.* 15 (8), 101313. doi:10.1016/j.jpha.2025.101313
- Zhang, O., Lin, H., Zhang, X., Wang, X., Wu, Z., Ye, Q., et al. (2025b). Graph neural networks in modern AI-Aided drug discovery. *Chem. Rev.* 125 (20), 10001–10103. doi:10.1021/acs.chemrev.5c00461
- Zhang, Y., Yu, L., Xue, L., Liu, F., Jing, R., and Luo, J. (2025c). Optimizing lipocalin sequence classification with ensemble deep learning models. *PLoS One* 20 (4), e0319329. doi:10.1371/journal.pone.0319329
- Zhang, C., Sun, J., Xing, L., Zhang, L., Cai, H., and Che, K. (2025d). VHGAE: Drug-target interaction prediction model based on heterogeneous graph variational autoencoder. *Interdiscip. Sci.* doi:10.1007/s12539-025-00758-8
- Zhang, K., Fan, Z., Wu, Q., Liu, J., and Huang, S. Y. (2025e). Improved prediction of drug-protein interactions through physics-based few-shot learning. *J. Chem. Inf. Model* 65 (13), 7174–7192. doi:10.1021/acs.jcim.5c00427
- Zhang, Z., Wang, S., Huang, Y., Zheng, X., and Dong, S. (2025f). Confidence-aware adaptive fusion learning of imbalance multi-modal data for cancer diagnosis and prognosis. *IEEE J. Biomed. Health Inf.* 30, 609–616. doi:10.1109/jbhi.2025.3582626
- Zhao, X., Wu, J., Zhao, X., and Yin, M. (2023). Multi-view contrastive heterogeneous graph attention network for lncRNA-disease association prediction. *Brief. Bioinform* 24 (1), bbac548. doi:10.1093/bib/bbac548
- Zhao, S., Yang, X., Zeng, Z., Qian, P., Zhao, Z., Dai, L., et al. (2024). Deep learning based CETSA feature prediction cross multiple cell lines with latent space representation. *Sci. Rep.* 14 (1), 1878. doi:10.1038/s41598-024-51193-6
- Zheng, L., Shi, E., Peng, C., Xu, M., Fan, F., Li, Y., et al. (2024). Application scenario-oriented molecule generation platform developed for drug discovery. *Methods* 222, 112–121. doi:10.1016/j.ymeth.2023.12.009
- Zheng, L., Zhang, S., Li, Y., Liu, Y., Ge, Q., Gu, L., et al. (2025). MTGNN: a drug-target-disease triplet association prediction model based on multimodal heterogeneous graph neural networks and direction-aware metapaths. *J. Chem. Inf. Model* 65 (12), 5921–5933. doi:10.1021/acs.jcim.5c00817
- Zhou, X., Fu, Q., Xia, Y., Wang, Y., Lu, Y., Chen, Y., et al. (2024). A local to global graphical reasoning framework for extracting structured information from biomedical literature. *IEEE J. Biomed. Health Inf.* 28 (4), 2314–2325. doi:10.1109/jbhi.2024.3358169
- Zhou, J., Kim, Y. K., Li, C., and Park, S. (2025). Natural compounds for Alzheimer's prevention and treatment: integrating SELFormer-based computational screening with experimental validation. *Comput. Biol. Med.* 185, 109523. doi:10.1016/j.combiomed.2024.109523