arXiv:2506.01302v1 [cs.LG] 2 Jun 2025

Recent Developments in GNNs for Drug Discovery

Zhengyu Fang[1], Xiaoge Zhang[1], Anyin Zhao[1], Xiao Li[1,2,3,4], Huiyuan Chen[1], and Jing Li[*1]

[1]Department of Computer and Data Sciences, Case Western Reserve University
[2]Department of Biochemistry, Case Western Reserve University
[3]Center for RNA Science and Therapeutics, Case Western Reserve University
[4]Department of Biomedical Engineering, Case Western Reserve University

Abstract

In this paper, we review recent developments and the role of Graph Neural Networks (GNNs) in computational drug discovery, including molecule generation, molecular property prediction, and drug-drug interaction prediction. By summarizing the most recent developments in this area, we underscore the capabilities of GNNs to comprehend intricate molecular patterns, while exploring both their current and prospective applications. We initiate our discussion by examining various molecular representations, followed by detailed discussions and categorization of existing GNN models based on their input types and downstream application tasks. We also collect a list of commonly used benchmark datasets for a variety of applications. We conclude the paper with brief discussions and summarize common trends in this important research area.

Keywords: Drug Discovery, Graph Neural Networks, Machine Learning

1 Introduction

It is well known that traditional drug discovery is costly, time-consuming, and with high failure rates [1]. To streamline the process of drug discovery and mitigate resource-intensive laboratory work, significant research has been dedicated to the development of computational methods. Existing literature provides some comprehensive reviews on deep learning approaches in drug discovery [2, 3, 4, 5]. In this review, we focus on the development and applications of Graph Neural Networks (GNNs) on three related areas of computational drug development, namely, *Molecule Generation*, *Molecular Property Prediction*, and *Drug-Drug Interaction Prediction*, which not only receive increasing attention but also show promising results. We will summarize some most recent developments in these research areas and focus on computational advances published since 2021.

*Corresponding author. jingli@case.edu

1.1 Molecule Generation

An earlier step in developing a drug for a certain disease usually involves the search of enormous amount of molecules to create a small subset of candidates for further laboratory testing. To reduce experimental costs and to produce drug candidates more efficiently, many computational approaches have been developed to generate novel molecules [6, 7, 8, 9, 10]. These approaches usually take some existing molecules as inputs and generate valid but different molecules with high variety. Validity basically means that the generated molecules are possible to be synthesized. At the same time, high variety will ensure that the generated molecules span over a reasonably large molecular space in hope that some of them will have desired properties. Computational approaches are usually evaluated based on these two metrics.

Generally speaking, existing computational approaches can be characterized as two different types based on whether they can incorporate certain constrains. For some approaches [6, 7], any molecules can be generated, *i.e.*, the generation of molecules has no constraints associated with them. For the second type of approaches [8, 9, 10], they only generate molecules satisfying certain constrains. For example, all the generated molecules must include some fixed substructures such as ring structures. In other cases, the constraints can be task specifics, for example, many approaches [11, 12, 13, 14, 15] have been developed to generate molecules that bind to certain protein binding sites.

Early computational approaches for molecule generation were mostly applications of generative deep learning models, such as variational autoencoder [16]. More recently, there was a shift from generations of whole molecules all at once, to iterative generations of atoms and bonds [6, 8]. Many of such approaches utilized GNN modules as backbones to generate molecules with specific substructures or with high binding affinity to specific proteins [8].

1.2 Molecular Property Prediction

Once drug candidates are selected, researchers need to study drugable properties of candidate molecules, such as toxicity and possible adverse effects, water solubility, and binding affinity to target proteins. Experimental studies are slow and costly. With quick accumulation of high-quality training data, a great number of computational methods for molecular property prediction have been developed [17, 18, 19, 20], the goals of which are to reduce the need for extensive experimental validation and to accelerate the drug discovery process.

Molecular properties can be classified into different categories based on their levels, from molecular level (e.g., quantum mechanics) to physiological level [21]. Computationally, the tasks of molecular property prediction can be categorized into two broad classes based on their inputs: predictions based on single molecules (i.e., drug candidates) [17, 22, 23], and predictions of relationships based on two or more molecules (e.g., drug-target or protein-ligand) [24, 25, 26]. The predictions based on single molecules can be either classification-based or regression-based. The goal of classification tasks is to predict the presence/absence of certain properties. The regression tasks focus on the prediction of the quantitative value of a particular property of a molecule. Similarly, relationships among molecules [27, 28] can be represented by a binary value or a numerical value. Examples of the former include whether a protein and a target interact. The affinity score between two molecules is an example of the latter.

Traditionally, drug property prediction heavily relied on classical machine learning techniques such as random forests or support vector machines [2], necessitating deep domain knowledge for feature engineering. With the emergence of deep learning, there was a significant shift towards neural network based models [2, 29]. This transition marked a move from one-dimensional string representations of drugs to richer two-dimensional graph and three-dimensional conformation models, broadening the scope and accuracy of drug property prediction. In particular, GNNs, which take in a molecular graph representation and directly

predict molecular properties, have attracted substantial attention in cheminformatics and bioinformatics (*e.g.*, [30]).

1.3 Drug-Drug Interaction Prediction

In the treatment of complex diseases such as cancer and neurological disorders, combinational drug therapies have shown promise in improving treatment efficacy by targeting multiple biological pathways simultaneously [4, 31, 32]. However, as the number of possible drug combinations grows with the increasing number of available drugs, the potential for undetected and unexpected drug-drug interaction (DDIs) also rises [33]. Such interactions can lead to reduction of therapeutic effectiveness, adverse side effects [34], and sometimes increased hospitalization [35], posing significant challenges to the screening and optimization of combinational drug therapies. Therefore, it is crucial to develop methods that can precisely predict the drug-drug synergies and adverse interactions to ensure safer and more effective treatments.

Strictly speaking, drug combination therapies, which rely on synergistic or additive interactions of drugs to increase efficacy, reduce toxicity, and/or prevent drug resistance, can be treated as one type of drug-drug interactions. In this review, we use the broad definition of drug-drug interactions, which include both synergistic/additive as well as antagonistic/adverse interactions. To predict drug-drug interactions computationally, the inputs usually include representations of drugs and types of interactions. In the simplest case, methods may only focus on the prediction of the existence of a particular type of interaction [36]. The problem is simply formulated as a binary classification problem. In other cases (*e.g.*, data from cell-based assays), the relationships can be quantified based on one or multiple measures of individual drugs and/or drug pairs (e.g., half maximal inhibitory concentration or IC50), which can be viewed as regression tasks. Yet in some other cases (data from patient-based studies), different criteria or biomarkers (*e.g.*, blood pressure, blood sugar) are considered and only a few categorical values (*e.g.*, increasing, no change, decreasing) are considered for each criterion/biomarker. The problem can be formulated as a multi-class classification problem [37].

Over the years, many computational approaches have been developed for drug-drug interaction predictions [37, 36, 38, 39]. Earlier approaches were mostly based on traditional machine learning algorithms and/or matrix decomposition framework [38]. Some methods [40, 41, 42] can even take into considerations of patients' medical history and recommend drugs that are safe to use for specific patients. More recently, more and more approaches based on deep learning models have drawn increasing attention, and have achieved better performance [38]. In particular, with advancements in the development of GNNs, methods that adapt GNNs as backbones to incorporate interactions as networks have achieved state-of-the-art results [42].

2 Representations of Molecules

In general, molecules can be represented via fingerprints, the Simplified Molecular Input Line Entry System (SMILES) strings, or 2D-/3D-graphs in Fig. 1. Binary fingerprints representing molecular substructure or topology allows efficient computation and database search [43]. However, they cannot easily encode global features of molecules such as size and shape. SMILES is the most widely used linear representation for describing chemical structures since its invention [44], and is superior to other one-dimensional representation schemes such as binary fingerprints. However, there are innate limitations associated with the internal structure of SMILES representations when used in Natural Language Processing (NLP) algorithms.

Molecules can be represented as 2D graphs, where nodes represent atoms and edges represent chemical bonds, or as 3D graphs that also incorporate 3D coordinates, providing detailed spatial information. In

both cases, both nodes and edges can have their own unique properties or features. While the 2D graph representation is simpler, 3D representation can better capture interactions based on distances and angles between atoms [45, 46], thus providing a more comprehensive view that is crucial for modeling molecular dynamics and bindings.

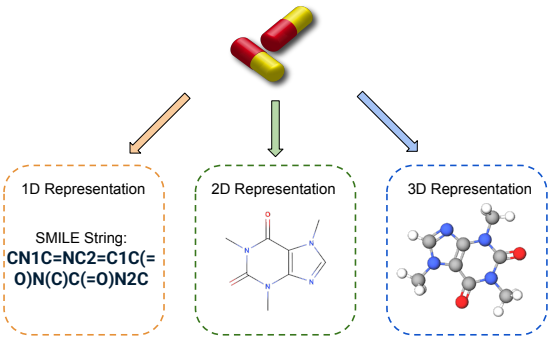

Figure 1: Various molecular representations commonly employed in computational drug discovery models: 1D SMILES strings, 2D molecular graphs, and 3D molecular graphs.

With molecules intuitively represented as graphs, GNNs offer a natural framework for handling and analyzing molecular data. This synergy has sparked extensive research, positioning GNNs at the forefront of innovation in molecular property and interaction prediction. GNNs allow nodes to aggregate information through their edges, creating comprehensive graph representations. Furthermore, by combining graph structures with neural networks, GNNs can readily handle both classification and regression tasks [47, 48, 49]. Therefore, many innovations in GNNs focus on graph representation learning, rather than specific prediction tasks. Most of the approaches to be discussed in this survey thus can be utilized in many downstream applications.

3 Molecule Generation

One of the initial steps in drug discovery involves selecting a set of candidate molecules for early-stage screening. Traditionally, this process has been conducted in laboratories. However, recent advances in computational methods have expanded the toolkit available to researchers, including the use of GNN models for molecule generation. These models leverage 2D and 3D graph representations of molecules to efficiently generate novel compounds within defined constraints. The development of GNN-based molecule generation methods can be broadly categorized into three types, as illustrated in Fig. 2: unconstrained generation [6, 7, 50, 51], constrained generation with targeted substructures [8, 10, 9], and ligand-protein-based generation [11, 12, 13, 14, 15]. Unconstrained methods prioritize structural diversity, constrained approaches focus on generating molecules containing specific functional groups or motifs relevant to desired chemical or biological properties, and ligand–protein-based strategies are designed to produce molecules that interact with specific protein targets. These advances demonstrate the versatility and promise of GNNs in accelerating drug discovery and highlight their pivotal role in identifying novel therapeutic candidates [52, 53, 54, 55].

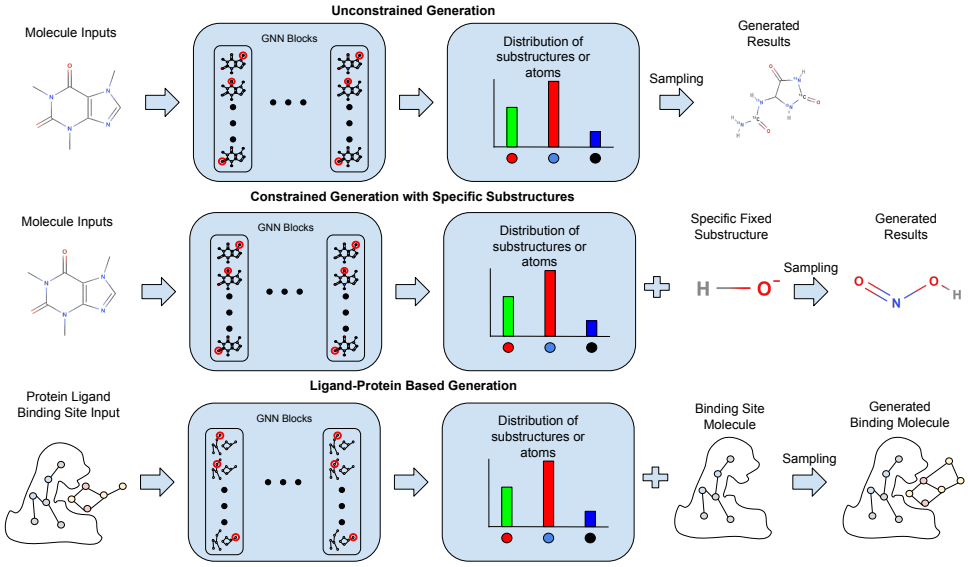

Figure 2: The general framework of three different types of molecule generation processes. Molecular graphs and protein-ligand complexes are fed into GNN backbone models, which output the probability distribution of molecular substructures to be sampled, based on which the models select substructures and assemble the resulting molecules.

3.1 Unconstrained Generation

For computational methods to generate promising drug candidates, a natural approach is to generate chemically valid molecules in the vicinity of existing drugs and drug candidates (*i.e.*, known training data). In other words, computational models are typically trained on "hit molecules"–compounds that already exhibit desirable properties for drug development. Because there are no additional constraints on molecular substructures or chemical properties (*e.g.*, binding affinity) imposed explicitly, such methods are classified as unconstrained. The primary goal of such approaches is to generate structurally valid and chemically plausible molecules with high diversity, guided by the distribution of the training data.

GraphINVENT [56] is one of the initial explorations into unconstrained molecular graph generation, which explored various GNN architectures–such as gated GNNs and attention-based GNNs–to learn an "action" probability distribution from training molecules. The model then iteratively sampled one action at a time, *e.g.*, adding a bond or an atom, until a "terminate generation" action was reached. Doing so allowed the approach to construct a molecule step by step. GraphINVENT is a representative among a broader class of models [57, 58, 59, 60] that all adopt such an iterative sampling strategy that build new graphs by repeatedly sampling components or actions from learned distributions.

Beyond this one-piece-at-a-time paradigm, another type of method adopted generative frameworks that produce complete molecules (graphs) in a single pass. For instance, ConfVAE [61] integrated both 2D molecular graphs and 3D conformations, ensuring rotational and translational invariance. It employed a Conditional Variational Autoencoder (CVAE) framework, with Message Passing Neural Networks (MPNNs) for graph encoding, enabling end-to-end conditional molecule generation. While ConfVAE leveraged GNNs within a generative framework, VonMisesNet [7] took a different approach by focusing on capturing the re-

alistic distribution of torsional angles in molecules. It introduced a novel GNN architecture that sampled torsion angles from the Von Mises distribution, which better reflected the physical constraints of molecular geometry. Moreover, VonMisesNet addressed key challenges such as chirality inversion of atoms and supports molecules with a large number of rotatable bonds-enhancing the chemical accuracy and diversity of its outputs.

Overall, unconstrained generation models are increasingly integrating with generative architectures such as VAEs and GANs, aiming to better approximate the true data distribution. Nonetheless, as the need grows to reduce the cost of candidate selection, improve the quality of generated molecules, and generate molecules that could be synthesized, research is gradually shifting toward generation models that incorporate explicit constraints on substructures and target properties.

3.2 Constrained Generation with Specific Substructures

In drug development, generating molecules with specific substructures and targeted chemical properties is often more desirable than unconstrained molecule generation. MoLeR, introduced by Maziarz et al. [8], demonstrated the capability to perform both constrained and unconstrained molecule generation. It utilized motifs–common chemical substructures–within an encoder-decoder framework. By combining GNN and Multilayer Perceptron (MLP) blocks, MoLeR carefully constructed molecules one motif at a time, sequentially selecting motifs or atoms, determining attachment points, and assigning bond types, each step optimized for molecular validity and functionality.

Building on the success of fragment-based approaches like MoLeR, newer models have further advanced constrained molecule generation. One such model, GEAM [10], introduced the Graph Information Bottleneck (GIB) principle to identify substructures most relevant to specific drug properties. GEAM first extracted a vocabulary of meaningful substructures and then assembled molecules from this learned vocabulary. A soft actor-critic (SAC) reinforcement learning algorithm was used to identify high-quality samples, which were subsequently mutated through a genetic algorithm (GA) to produce final molecules that were chemically valid and aligned with the desired drug properties.

While models like MoLeR and GEAM focused primarily on the generation process, MiCam [9] proposed a novel strategy for building a chemically "reasonable" motif vocabulary. MiCam addressed the limitation of previous fragment-based methods, which often failed to identify appropriate motifs for molecule generation. Its vocabulary construction involved two phases: in the merging-operation learning phase, the model iteratively merged the most frequent atomic patterns found across molecules to form a preliminary set of motifs. In the motif-vocabulary construction phase, the model disconnected fragments at learned attachment points, marking these connection sites with special tokens to preserve the information necessary for molecule assembly. This approach allowed MiCam to flexibly generate molecules either by adding known motifs or by extending partially generated structures based on the connection history.

Overall, these models share a common strategy of using substructures as modular building blocks, aligning generation objectives with the training loss, and constructing molecules in a stepwise manner. Among the three, GEAM and MiCam offer greater flexibility in incorporating specific constraints on both substructures and chemical properties, as they allow the use of both atoms and motifs during generation. In contrast, MoLeR primarily relies on starting from a predefined scaffold.

3.3 Protein-Ligand based Generation

In addition to generating molecules based solely on training data of individual compounds, researchers have developed models that focus on protein binding sites and their associated ligands, addressing a new

set of challenges in drug discovery. Recent advances in GNN-based molecule generation have enabled the creation of molecules specifically tailored for target proteins. These models employ GNN blocks to maintain structural consistency—ensuring robustness against flips, shifts, and rotations—while processing attributes and 3D coordinates of protein binding sites. Approaches such as the AR model [11] and GraphBP [12] introduced distinct strategies for representing atoms and binding environments.

These models adopted various techniques to prioritize contextual representation and resilience to rigid transformations. For instance, AR combined MLP blocks with an auxiliary network to guide atom generation and bonding decisions, whereas GraphBP utilized spherical coordinates alongside MLPs for sequential atom-by-atom construction. Other notable methods, including Pocket2Mol [13] and FLAG [14], incorporated auxiliary MLP classifiers and predictors to optimize atom positioning and motif attachment. Collectively, these strategies significantly improved model robustness and adaptability, representing critical progress toward customizing molecules for specific protein targets—a key advancement in drug discovery.

Each model has distinct strengths: the AR model emphasized specificity and binding affinity optimization for particular protein site structures; GraphBP introduced a dual-diffusion architecture to enhance flexibility; Pocket2Mol achieved greater computational efficiency through conditional 3D coordinate sampling; and FLAG leveraged motif-based generation to improve structural realism and diversity.

Beyond generating molecules that bind to specific protein sites, there is also a growing need to generate molecules constrained by a desired 3D binding conformation. In many cases, experimental data reveal that certain binding postures are particularly effective for interacting with specific proteins, and generating candidate molecules that adopt these conformations can greatly accelerate screening. SQUID [15] was the first model designed to address this challenge. Given a target 3D shape, SQUID encodes the input conformation—treated as an unordered point cloud—into hidden features using GNN layers, and then iteratively generates 3D molecular fragments that reconstruct the desired shape fragment-by-fragment.

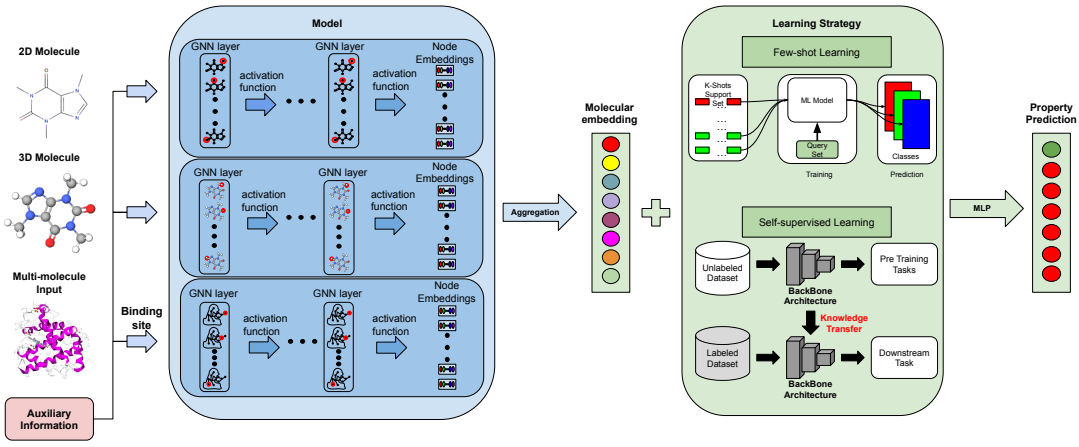

Figure 3: The general framework for GNN-based drug property and interaction prediction. Three common types of inputs are used individually or jointly: 2D molecule graphs, 3D molecule graphs, multi-molecule interaction graphs such as protein-ligand complexes. Additional auxiliary information can also be incorporated by some approaches. These inputs are then fed into GNN models, which aggregate information from neighboring nodes and produced final latent node representations. To alleviate the label sparsity issue, various learning strategies, such as few-shot learning or self-supervised learning, are widely adopted.

4 Prediction of Molecular Properties and Interactions

In this section, we review various GNN-based approaches for predicting molecular properties and interactions. Different algorithms have been developed for different applications, each utilizing varying types of inputs. Common types of input data include 2D molecular graphs, 3D molecular conformations, multi-molecule complexes, and potentially additional auxiliary information (Fig. 3).

A key component shared by all approaches is the learning of a latent molecular representation through GNNs, which is subsequently used for prediction tasks. To address challenges such as label sparsity, many methods also incorporate advanced learning strategies, including self-supervised (pre-training) and few-shot learning. In this section, we categorize and discuss these approaches based on their specific prediction tasks and input data types, highlighting their novel contributions and offering insights into how they advance the field.

4.1 Property Prediction and Molecule Representation Learning Based on 2D Graphs

Learning effective molecular representations is a fundamental step in property prediction. Generally, the goal of molecular representation learning is to embed molecules into numerical vectors in a latent space, enabling a variety of downstream tasks [17]. Therefore, much of the innovation in property prediction from 2D graphs lies in the strategies developed for learning molecular representations, which is the focus of this subsection.

When molecules are represented as 2D graphs, early studies typically employed MPNNs [62], wherein nodes (atoms) exchange messages with neighboring nodes, aggregating this information to update their respective states. However, obtaining class labels or quantitative measurements often requires costly laboratory experiments or human annotations. As a result, many datasets contain large numbers of unlabeled or imbalanced samples. To address this, newer approaches often incorporate pre-training strategies or rely on few-shot learning.

4.1.1 Pre-training

Pre-training[1] has emerged as a promising technique to mitigate label scarcity and is widely adopted in modern GNN models for molecular property prediction. For example, MGSSL [63] introduced a motif-based graph self-supervised learning framework that exploited rich information in subgraphs often overlooked at the node level. In this setting, the choice of molecular fragmentation method is critical, as poor fragmentation can lead to suboptimal motif generation and degraded model performance.

Many other methods adopt contrastive graph learning as their pre-training strategy. MoCL [18], for instance, utilized knowledge-aware contrastive learning informed by local and global domain knowledge. Local domain knowledge ensured semantic invariance during augmentation, while global domain knowledge infused structural similarity across the learning process. Nevertheless, designing augmentation schemes that generalize across diverse molecular structures remains challenging. KCL [64] combined contrastive learning with domain-specific knowledge graphs, offering tailored augmentations at the cost of generalizability. MCHNN [65] applied multi-view contrastive learning with task-specific augmentations, enhancing the expressiveness of molecular representations. HiMol [66] advanced self-supervised learning by proposing a hierarchical GNN that captured node, motif, and graph-level representations. However, constructing motif dictionaries can be computationally intensive, especially for large molecular databases.

[1]Here, pre-training refers to self-supervised learning on the input data itself or its augmented versions, differing slightly from the conventional notion in transfer learning where models are first trained on large independent datasets.

4.1.2 Few-shot Learning

Few-shot learning approaches aim to predict molecular properties with minimal labeled data [67]. HSL-RG [19] explored both global and local structural semantics for few-shot learning: global information was captured via molecular relation graphs built from graph kernels, while local information was learned through transformation-invariant representations. Similarly, MHNfs [68] employed a context module that retrieved and enriched molecule representations from a large pool of reference molecules, although the requirement for such large reference sets may limit its practical use.

GS-Meta [69] extended few-shot learning to simultaneously handle multiple properties or labels. PACIA [20] introduced hypernetworks to generate adaptive parameters for modulating the GNN encoder, reducing overfitting while maintaining flexibility. However, designing effective hypernetworks demands significant domain expertise and may constrain the method's generalizability across tasks. An alternative strategy to combat label scarcity involves grammar-based generation, as exemplified by Geo-DEG [70], which employed a hierarchical molecular grammar to create molecular graphs, using production paths as informative priors for structural similarity.

4.1.3 Incorporating Auxiliary Information

Beyond the standard 2D graph representation, researchers have explored integrating additional molecular information during learning. For example, PhysChem [22] showed that incorporating physical and chemical properties improved molecular representations, though its cooperation mechanism was less interpretable than traditional ensemble strategies. O-GNN [23] incorporated ring priors into the modeling, leveraging their importance in determining molecular properties. MoleOOD [71] introduced invariant substructure learning to better handle distribution shifts across environments. However, in datasets with large environmental variations, MoleOOD's advantage over simpler methods like ERM [72] may diminish.

Some recent methods have also integrated both one-dimensional sequential encodings (*e.g.*, SMILES strings) and 2D graphs to jointly leverage information from both views. DVMP [73], for instance, encoded molecular graphs via GNNs and SMILES sequences via Transformers, using a dual-view consistency loss to maintain semantic coherence. However, involving both Transformer and GNN branches significantly increases training costs compared to single-branch models.

4.2 Property Prediction based on 3D-graphs

Recent advancements in deep learning and molecular science, driven by the increasing availability of large-scale 3D molecular datasets, have significantly advanced property prediction based on 3D molecular graphs. SphereNet [45] introduced an innovative approach for 3D molecular representation learning, proposing a spherical message-passing scheme that explicitly incorporates 3D spatial information. While SphereNet demonstrated strong predictive performance, it lacked transparency and interpretability, which hinders the understanding of its decision-making processes. MolKGNN [74] addressed this limitation in the context of quantitative structure–activity relationship (QSAR) modeling. It enhanced 3D molecular representation learning by employing molecular graph convolution with learnable molecular kernels, effectively capturing chemical patterns. Importantly, MolKGNN incorporated molecular chirality, a critical aspect often neglected in previous models. However, the method emphasizes distinguishing specific molecular substructures, which may limit its generalizability across diverse chemical variations encountered in practical drug discovery scenarios.

Several studies have explored integrating both 2D and 3D information for property prediction. For instance, GraphMVP [75] developed a 2D graph encoder enriched by discriminative 3D geometric informa-

tion. It employed a self-supervised pre-training strategy that leveraged the correspondence and consistency between 2D topologies and 3D conformations. Similarly, 3D-Informax [30] proposed a transfer learning framework that pre-trained on molecules with both 2D and 3D data and then transferred the learned knowledge to molecules with only 2D structures. However, such approaches may risk overfitting, as evidenced by performance variability across different datasets.

UnifiedPML [76] further improved representation learning by jointly considering 2D and 3D information in its pre-training scheme. The framework employed three complementary tasks: reconstruction of masked atoms and coordinates, generation of 3D conformations conditioned on 2D graphs, and generation of 2D graphs conditioned on 3D conformations. GeomGCL [77] adopted a dual-channel message-passing neural network to effectively capture both topological and geometric features of molecular graphs.

MoleculeSDE [78] unified 2D and 3D molecular representations by treating them as separate modalities in a multi-modal pre-training framework. 3D-PGT [79] proposed a generative pre-training approach on 3D molecular graphs, which was subsequently fine-tuned on molecules lacking 3D structural data. It employed a multi-task learning strategy based on three geometric descriptors–bond lengths, bond angles, and dihedral angles–and used total molecular energy as an optimization target. While promising, the effectiveness of this framework remains to be validated on larger and more structurally complex molecules, as current evaluations have primarily focused on small molecules.

4.3 Interaction Prediction

Beyond property prediction, interaction prediction has been extensively explored, especially for drug-target or drug -disease interaction predictions. Many researchers (e.g., NeurTN [17]) have utilized drug-target interaction networks as input, where the nodes are drugs and targets and links are known drug-target relationship. These models typically infer new interactions based on the guilt-by-association principle and are fundamentally different from the methods discussed in this work, which rely primarily on drugs' structure information represented as molecular graphs. Therefore, interaction network based approaches are excluded from further discussion in this section.

In drug discovery, one of the most critical and extensively studied relationships is the interaction between drugs (or chemical compounds) and their protein targets. In the literature, various terms are used to describe these interactions, each emphasizing different aspects. These include drug-target interaction, protein-ligand interaction, drug-target binding affinity, protein-ligand binding affinity, molecular docking. Computationally, given the 2D or 3D structures of two molecules, the interaction can be studied at three levels: (1) binary interaction (*i.e.*, whether an interaction occurs), (2) binding affinity (a numerical value, typically reflecting binding free energy), and (3) docking or protein-ligand binding dynamics.

GNN-based approaches have been proposed to predict drug-target interactions based on their 2D structure. For example, CGIB [28] predicted interactions primarily using substructure information from paired graphs. MGraphDTA [27] predicted drug–target binding affinities based on 2D compound graphs and protein sequences. It utilized a deep GNN to capture both local and global molecular structures and a multi-scale convolutional neural network (CNN) to extract features from protein sequences. However, capturing long-range dependencies within complex molecular graphs remains a challenge for such models. Given the superior performance generally observed when utilizing 3D molecular geometries, the trend shows that more approaches incorporate 3D information for interaction prediction, especially for binding affinity prediction and docking.

For binding affinity prediction, the inputs are usually protein-ligand complexes, and the objective is to predict a binding score that reflects the strength of interaction, typically in terms of free energy. Recent developments have leveraged 3D graph representation learning to tackle this problem. For instance, Jones

et al. [24] proposed a fusion model that combined complementary molecular representations. Their method utilized a 3D CNN to capture local spatial features and a spatial GNN to encode global structural information, integrating both in a fused architecture.

The IGN framework [80] modeled protein-ligand complexes using three distinct molecular graphs, each incorporating both 3D structural and chemical properties. MP-GNN [81] introduced a multiphysical molecular graph representation, which systematically captured a wide range of molecular interactions across different atom types and physical scales. However, most existing biomolecular GNNs rely on covalent-bond-based graph constructions, which often fail to effectively characterize non-covalent interactions essential for modeling biomolecular complexes.

GraphscoreDTA [82] advanced this field by integrating a bitransport information mechanism and Vina distance optimization terms to better capture the mutual information between proteins and ligands. This method also highlighted critical atomic and residue-level features. In contrast to the above, NERE [25] proposed an unsupervised approach to binding energy prediction, framing it as a generative modeling task. Their method, based on Neural Euler's Rotation Equations (NERE), predicted molecular rotations by modeling the forces and torques between ligand and protein atoms. However, the current implementation of NERE for antibody modeling only considers backbone atoms and omits side-chain atoms, which are crucial for accurately estimating binding affinity.

Docking, a central process in drug discovery, has also seen innovation through GNN-based approaches. E3Bind [26] introduced an end-to-end model that directly generates ligand coordinates, thus eliminating the need for traditional sampling procedures and coordinate reconstructions. Similarly, FABind [83] combined pocket prediction and docking in an integrated model for fast and accurate binding pose prediction. A unique ligand-informed pocket prediction module was used to guide the docking process, with successive refinements optimizing the ligand-protein binding pose. The model further enhanced the docking process by incrementally integrating the predicted pockets to optimize protein-ligand binding. However, ablation studies indicated that different components contribute to the model's performance in varying degrees, suggesting potential inefficiencies in the overall architecture. More recently, NeuralMD [84] provided a fine-grained simulation of protein-ligand binding dynamics. The model included BindingNet, which adhered to group symmetry and captured multi-level interactions, and a neural ordinary differential equation (ODE) solver that modeled the physical trajectories of atoms based on Newtonian mechanics.

EquiPocket [85], distinct from the aforementioned methods, focused specifically on predicting ligand binding sites for given 3D protein structures. It introduced three novel modules: a local geometric modeling module to extract features from individual surface atoms, a global structural module to encode the chemical and spatial context of the entire protein, and a surface message-passing module to learn surface-level geometric patterns. In contrast to CNN-based methods, which suffer from inefficiencies due to voxelization of irregular protein surfaces, EquiPocket avoids computational redundancy and excessive memory usage through its surface-based geometric design.

5 Prediction of Drug-Drug Interactions

Predicting and understanding DDIs is a critical step in computational drug discovery, especially in the context of drug combination therapies [86, 87, 88], in which case multiple drugs are commonly used together in clinical practice to treat complex diseases such as cancer [89, 90, 91]. However, polypharmacy elevates the risk of adverse DDIs, potentially compromising therapeutic efficacy, posing serious health risks, and increasing healthcare costs [92, 93, 94]. Historically, many DDIs were discovered via clinical case reports or mined from electronic health records (EHRs) [95, 96]. Computational approaches, particularly those

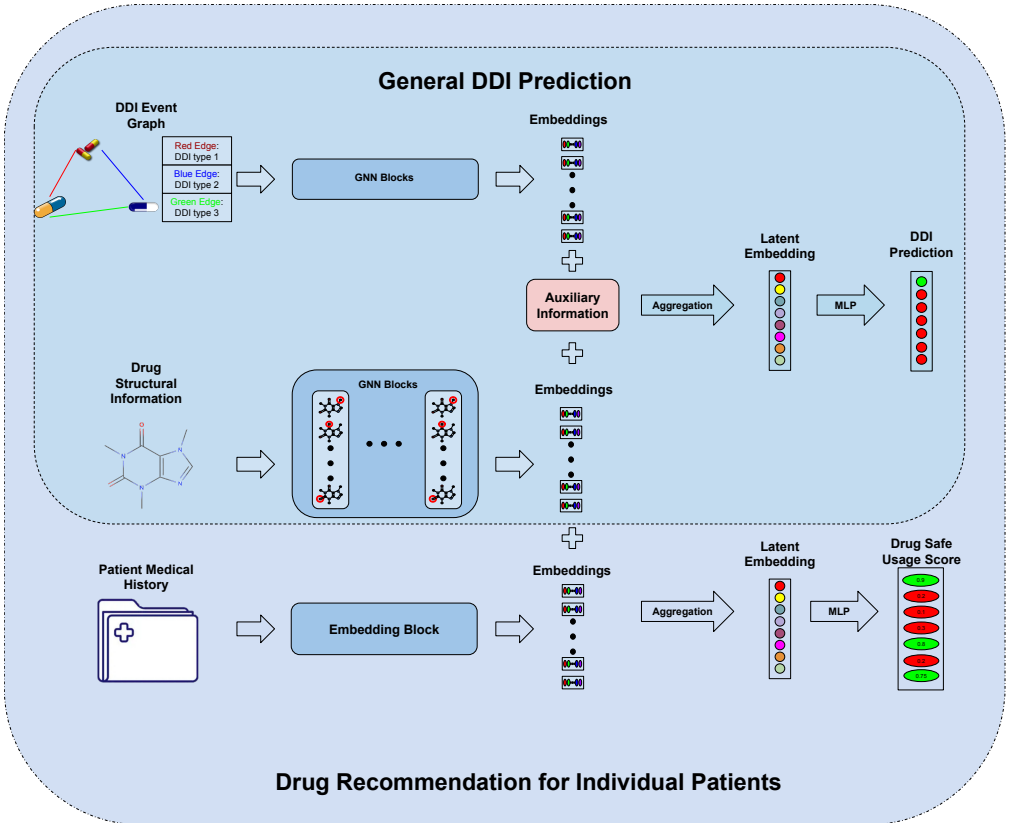

Figure 4: The general process of DDI prediction based on GNN models. Possible inputs for the general DDI prediction (the one inside the small rectangle) include DDI event graphs, and drug molecular structures, either individually or jointly. Additional auxiliary data can be incorporated into the models. GNN blocks map the inputs into the latent space, which will be utilized for DDI prediction. By including patient medical history, the model can be extended to perform patient-specific drug safety recommendations.

based on machine learning, now offer scalable and cost-effective alternatives to identify novel candidate interactions, either synergistic or adverse ones, beforehand.

Recent computational strategies for DDI prediction can be broadly categorized into two paradigms, illustrated in Fig. 4: (1) *general DDI prediction*, which identifies potential interactions across large drug populations; and (2) *personalized drug combination recommendation*, which tailors treatment regimens by further considering individual patient health profiles and personal DDI risk. In this subsection, we will first focus on general DDI prediction, and examine commonly used input data types, followed by the discussion of recent developments in problem formulations and model architectures. We conclude this subsection with discussions on personalized drug combination recommendation.

5.1 Types of Input Data

Similar to drug–target interaction prediction discussed in the previous section, many early approaches to DDI prediction primarily utilized drug–drug interaction graphs, where nodes represent drugs and edges encode their interactions. With advancements in the field, more recent methods have begun to incorporate drug molecular structures, represented as molecular graphs (as introduced previously), in which nodes denote atoms and edges correspond to chemical bonds. In both types of graphs, nodes and edges are typically enriched with additional features or attributes that capture relevant properties. For example, in drug–drug interaction graphs, node attributes may encode drug-specific properties, while edges can be labeled to indicate interaction types (e.g., synergistic or antagonistic effects). In molecular graphs, node features represent atomic properties, and edge features describe bond characteristics. GNN-based models are capable of processing both graph types, often alongside auxiliary information such as drug similarity matrices, to learn latent representations from different perspectives and at multiple levels—capturing both relational patterns in drug interaction networks and structural characteristics at molecular and sub-molecular scales. Increasingly, recent approaches aim to integrate both types of input within a unified learning framework, jointly capturing topological and structural information to enhance predictive performance.

5.2 Problem Formulations

The choice of input data and their representations not only inspires model design but also significantly influences problem formulations. Early models primarily focused on drug–drug interaction graphs combined with simple node-level drug features. Various GNN architectures were employed to learn low-dimensional representations of drugs from these graphs [97, 98]. Building on this idea, models such as GCNMK [36] further decomposed the DDI graph into two separate graphs, one representing interactions where a drug increases the activity of another, and the other where it decreases activity, and applied two GNNs to learn drug representations from these differentiated views. Subsequent studies expanded the input space by incorporating drug molecular graphs, thereby enabling the integration of both structural and relational perspectives to enhance model performance. For instance, MRCGNN [39] employed a GNN to process the relational DDI graph while enriching each drug's feature representation with molecular-level information extracted by a separate GNN operating on its molecular graph. This multimodal approach allowed the model to simultaneously capture both chemical and interaction-level knowledge.

The evolution of problem formulation also extends to the design of prediction tasks. Some models framed DDI prediction as a binary classification problem, aiming solely to determine the existence of an interaction [98]. Others formulated it as a multi-label classification task, predicting both the presence and the specific type of interaction from a predefined label set [99, 100, 101, 102]. Many researchers have further distinguished between adverse DDI prediction [103, 104, 105] and drug combination or synergy prediction [106, 105, 107, 108], providing a more nuanced understanding of interaction consequences.

Overall, advancements in problem formulation aim to enrich input representations with biologically meaningful information and to enable more fine-grained, application-specific predictions. Future directions are likely to emphasize input data that better reflect the underlying biological complexity of DDIs. For the prediction outcomes, knowledge of specific side effects from adverse interaction prediction and information of targeted diseases in drug combination modeling are welcome additions.

5.3 Advancements in Model Structure

Beyond problem formulation, significant research has focused on improving model architectures to more effectively aggregate information across different data modalities and drug representations. These approaches

typically employ distinct architectures (e.g., GNNs, CNNs) for different modalities or representations, and combine their learned features using various fusion strategies. Earlier models such as MDNN [37], GC-NMK [36], MRCGNN [39], and DeepDDS [107] fused features from each modality or representation through simple concatenation. While this strategy preserves feature information from all modalities, it neither accounts for the relative importance of each data source nor captures potential inter-modal relationships.

To address these limitations, more recent models have incorporated attention mechanisms to fuse latent features from different modalities, or from different drugs within the same modality, via cross-attention. For instance, SSF-DDI [109] utilized two drug representations: the 1D SMILES sequence and the 2D graph structure. Separate architectures (CNN for SMILES and GNN for molecular graphs) were used, and a cross-attention mechanism was employed to integrate the latent features generated by the two models. Similarly, SRR-DDI [110] constructed 2D molecular graph representations for drug pairs and applied cross-attention to fuse the learned latent features of the two drugs.

MD-Syn [111] proposed a multi-modal architecture with both one-dimensional and two-dimensional feature embedding modules, which allows incorporation of SMILES sequences, cell line information, drug molecular graphs, and protein–protein interaction (PPI) networks. Rather than using cross-attention, MD-Syn introduced a graph-trans pooling module within the 2D-feature embedding module, employing Transformer encoder layers with multi-head self-attention to process the concatenated latent representations from the PPI network and drug graphs.

Another direction in architectural advancement focuses on multi-level feature aggregation across GNN layers, particularly for molecular graphs. For example, DAS-DDI [112] introduced weighted layer-wise aggregation, where each GNN layer contributes differently to the final embedding. This enables molecular substructures of varying granularity to inform the final drug representation, thereby enhancing the expressiveness and robustness of the model in capturing complex inter-drug relationships.

5.4 Personalized Drug Combination Recommendation

A distinct line of research focuses on personalized DDI prediction by incorporating patient-specific medical histories. These models are less common due to the fact that data privacy concerns hinder the availability of clinical data, but they offer unique insights. For instance, SafeDrug [40] used GNNs and RNNs to align molecular features with patient treatment histories, producing compatibility scores for candidate drug combinations. MoleRec [41] leveraged attention mechanisms to integrate patient records and drug representations for safe prescription generation.

Despite their promise, challenges remain. GNNs often generate nearly identical embeddings for structurally similar molecules, regardless of therapeutic context. Carmen [42] addressed this with a context-aware GNN that incorporated medication context during atom-level message aggregation. This architecture produced distinct embeddings based on therapeutic relevance, offering a refined strategy for personalized drug combination recommendations. These models incorporating personal information represent a significant step toward safer, more effective treatment planning, highlighting the value of integrating biomedical knowledge with patient-specific data.

Finally, for easy reference, all the approaches discussed in this review for all the three tasks are organized in Table 1.

6 Benchmark Databases

In addition to newly developed methodologies, benchmark datasets play a vital role in advancing the field of computational drug discovery. High-quality data is essential for all the tasks ranging from molecular design

Table 1: GNN-based models discussed in this review and their characteristics. Each row includes the name of the approach, the main model architecture, the prediction task and the datasets used. The approaches are grouped into different bucket based on their tasks. The background with yellow color indicates that the approaches primarily utilized 2D structure and the blue color indicates that the approaches primarily utilized 3D structure. Methods using pre-training are labeled with ♯ and methods using few-shot learning are labeled with ⋆.

Name	architecture	Task	Datasets
ConfVAE [6]	MPNN	Unconstrained Generation w/ CVAE and uses 2D&3D	GEOM-QM9, GEOM-Drugs
VonMisesNet [7]	GCN	Unconstrained Generation w/ Von Mises distribution	NMRShiftDB, GDB-17
MoLeR [8]	GNN	Constrained Generation w/ motifs-based substructures	GuacaMol
MiCam [9]	GNN	Constrained Generation w/ connection-aware motif vocabulary	QM9, ZINC, GuacaMol
GEAM [10]	MPNN	Constrained Generation w/ soft-actor critic	ZINC250k
AR [11]	GNN	Ligand-Protein Based Generation w/ auxiliary network	CrossDocked
GraphBP [12]	GNN	Ligand-Protein Based Generation w/ spherical coordinates	CrossDocked
Pocket2Mol [13]	GNN	Ligand-Protein Based Generation w/ auxiliary atom positioning	CrossDocked
FLAG [14]	GNN	Ligand-Protein Based Generation w/ auxiliary motif attachment	CrossDocked
SQUID [15]	GNN	Ligand-Protein Based Generation w/ 3-D shape	MOSES
NeurTN [17]	GNN	Property Prediction w/ powerful nonlinear relationships	CTD,DrugBank,UniProt4
PhysChem [22]	MPNN	Property Prediction w/ physical&chemical information	QM7,QM8,QM9,Lipop,FreeSolv,ESOL,COVID19
O-GNN [23]	GNN	Property Prediction w/ ring substructures	BBBP,Tox21,ClinTox,HIV,BACE,SIDER,FS-Mol
MoleOOD [71]	SAGE	Property Prediction w/ invariant substructure across environments	BACE,BBBP,SIDER,HIV,DrugOOD
MGSSL [63]	GNN ♯	Property Prediction w/ motif-based self-supervised learning	MUV,ClinTox,SIDER,HIV,Tox21,BACE, ToxCast,BBBP
MoCL [18]	GIN ♯	Property Prediction w/ knowledge-aware contrastive learning	BACE,BBBP,ClinTox,Mutag,SIDER,Tox21,ToxCast
KCL [64]	MPNN ♯	Property Prediction w/ domain knowledge contrastive learning	BBBP,Tox21,ToxCast,SIDER,ClinTox,BACE, ESOL,FreeSolv
MCHNN [65]	GCN ♯	Property Prediction w/ multi-view contrastive learning	PubChem,MDAD,DrugVirus,HMDAD,Disbiome, gutMDisorder,Peryton
HiMol [66]	GNN ♯	Property Prediction w/ boundaries self-supervised learning	BACE,BBBP,Tox21,ToxCast,SIDER,ClinTox, ESOL,FreeSolv,Lipop,QM7,QM8,QM9
HSL-RG [19]	GNN ♯⋆	Property Prediction w/ few-shot learning&self-supervised learning	Tox21,SIDER,MUV,ToxCast
MHNfs [68]	GNN ⋆	Property Prediction w/ few-shot learning&context module	FS-Mol
GS-Meta [69]	GNN ⋆	Property Prediction w/ few-shot learning&simultaneous multiple labels	Tox21,SIDER,MUV,ToxCast,PCBA
PACIA [20]	GNN ♯⋆	Property Prediction w/ few-shot learning&adaptive parameters	Tox21,SIDER,MUV,ToxCast,FS-Mol
Geo-DEG [70]	MPNN	Property Prediction w/ hierarchical molecular grammar	CROW,Permeability,FreeSolv,Lipop,HOPV,PTC,ClinTox
DVMP [73]	GCN ♯	Property Prediction w/ pre-train for dual-view 1D&2D molecule	BBBP,Tox21,ClinTox,HIV,BACE,SIDER,ESOL
GraphMVP [75]	GNN ♯	Property Prediction w/ pre-train consistency between 2D&3D	BBBP,Tox21,ToxCast,SIDER,MUV,HIV,BACE
SphereNet [45]	MPNN	Property Prediction w/ spherical message passing	QM9
UnifiedPML [76]	GN Blocks ♯	Property Prediction w/ pre-train on multi-tasks for 2D&3D	BBBP,Tox21,ClinTox,HIV,BACE,SIDER
GeomGCL [77]	MPNN ♯	Property Prediction w/ dual-channel message passing for 2D&3D	ClinTox,SIDER,Tox21,ToxCast,ESOL,FreeSolv,Lipop
MolKGNN [74]	GNN	Property Prediction w/ molecular chirality	PubChem
3D-Informax [30]	MPNN ♯	Property Prediction w/ transfer learning for 2D&3D	QM9,GEOM-Drugs
MoleculeSDE [78]	GNN ♯	Property Prediction w/ multi-modal pre-train for 2D&3D	BBBP,Tox21,ToxCast,SIDER,ClinTox,MUV,HIV,BACE
3D-PGT [79]	GNN ♯	Property Prediction w/ multi-task generative pre-train on 3D	BBBP,Tox21,ToxCast,SIDER,ClinTox,MUV, HIV,BACE,ESOL,Lipop,Malaria,CEP,Davis,KIBA
MGraphDTA [27]	GNN	Molecular Interactions Prediction w/ super-deep GNN	Davis,KIBA,Metz,Human,C. elegans,ToxCast
CGIB [28]	MPNN	Molecular Interactions Prediction w/ substructure information	MNSol,FreeSolv,CompSol,Abraham,CombiSolv
SG-CNN [24]	GNN	Binding Affinity Prediction w/ complementary representations	PDBbind
IGN [80]	GNN	Binding Affinity Prediction w/ chemical information	PDBbind
MP-GNN [81]	GNN	Binding Affinity Prediction w/ multiphysical representations	PDBbind,SARS-CoV BA
GraphscoreDTA [82]	GNN	Binding Affinity Prediction w/ bitransport information	PDBbind
NERE [25]	MPNN ♯	Binding Affinity Prediction w/ Neural Euler's Rotation Equations	PDBbind
E3Bind [26]	GIN	Binding Affinity Prediction w/ docking	PDBbind
FABind [83]	GCN	Binding Affinity Prediction w/ pocket prediction and docking	PDBbind
NeuralMD [84]	MPNN	Protein-Ligand Binding Dynamics Simulations	MISATO
EquiPocket [85]	GNN	Ligand Binding Site Prediction w/ geometric and chemical	scPDB,PDBbind,COACH420,HOLO4K
MDNN [37]	GNN	DDI Prediction w/ knowledge graphs	DrugBank
DPDDI [98]	GCN	DDI prediction w/ extraction of the network structure features of drugs from DDI network	DrugBank,ZhangDDI
GCNMK [36]	GCN	DDI Prediction w/ dual-block GNN	DrugBank
MRCGNN [39]	GCN	DDI Prediction w/ incorporation of negative DDI event	Deng's dataset,Ryu's dataset
SRR-DDI [110]	MPNN	DDI Prediction w/ self-attention mechanism	DrugBank,Twosides
DAS-DDI [112]	GCN	DDI Prediction w/ dual-view framework	DrugBank,ChChMiner,ZhangDDI
SSF-DDI [109]	MPNN	DDI Prediction w/ on sequence and substructure features	DrugBank,Twosides
DeepDDS [107]	GAT, GCN	synergetic DDI Prediction w/ attention mechanism	O'Neil's dataset,Menden's dataset
MD-Syn [111]	GCN	synergistic DDI Prediction w/ chemicals and cancer cell line gene expression profiles	O'Neil's dataset, DrugCombDB
SafeDrug [40]	MPNN	Drug Combinations Recommendation w/ explicit leverages of drugs' molecule structures and model DDIs	MIMIC-III
MoleRec [41]	GIN	Drug Combinations Recommendation w/ molecular substructure-aware encoding method	MIMIC-III
Carmen [42]	GNN	Drug Combinations Recommendation w/ context-aware GNN	MIMIC-III,MIMIC-IV

Graph Isomorphism Network(GIN), GraphSAGE(SAGE), Graph Convolutional Network(GCN), Graph network block(GN blocks)

and property prediction to the characterization of drug–drug interactions and it serves as a foundation for objectively evaluating the effectiveness of various predictive models.

We assembled a comprehensive list of datasets referenced across the reviewed studies and organized

them by their data characteristics, as summarized in Table 2. Four primary categories capture the scope of these resources: Comprehensive Databases, Clinical Databases, Structural Information Databases, and Molecular Interaction Databases. Given the breadth and interrelated nature of the latter, we subdivided Molecular Interaction Databases into Protein–Ligand Binding and Drug–Drug Interaction collections, each distinguished by color coding in table. While not exhaustive, this selection emphasizes the most influential datasets driving progress in computational drug discovery.

Table 2: Commonly used benchmark databases and their brief descriptions. Consistent with discussion in the paper, we separate the datasets into four categories: Comprehensive Databases, Clinical Databases, Structural Information Databases, and Molecular Interaction Databases. The Molecular Interaction category is further divided into Protein–Ligand Binding and Drug–Drug Interaction.

Task	Dataset	Description
Comprehensive Databases	DrugBank [113]	Extensive repository of approved and investigational drugs linking chemical structures with pharmacological profiles and target interactions.
	PubChem [114]	Vast compound library annotated with high-throughput screening bioactivities and comprehensive chemical properties.
	MoleculeNet [21]	Aggregated benchmark collection covering diverse molecular properties and activities for algorithm evaluation.
Clinical Databases	MIMIC-III [115]	Detailed, de-identified ICU patient records including vitals, labs, and clinical interventions over time.
	MIMIC-IV [116]	A public EHR dataset with deidentified clinical data for 180,733 hospital and 50,920 ICU patients, covering patient tracking, billing, medications, and measurements.
	UK Biobank [117]	Population-scale cohort with deep phenotypic, genotypic, and long-term health outcome data.
Structural Information Databases	ZINC [118]	Vendor-curated set of purchasable compounds each with experimentally determined 3D conformers.
	GEOM [119]	High-precision quantum-mechanically optimized 3D molecular geometries for conformational analysis.
	MISATO [120]	Multigrained collection of protein–ligand complexes annotated with binding-site details.
	CrossDocked [121]	Large-scale docking dataset providing multiple poses and affinity estimates for protein–ligand pairs.
Molecular Interaction Databases Protein–Ligand Binding	ChEMBL [122]	Expert-curated database of small molecules linked to experimentally measured target binding affinities.
	Metz Dataset [123]	Collection of kinase inhibitor experiments reporting inhibition constants (K_i) across targets.
	KIBA Dataset [124]	Unified resource converting heterogeneous kinase-inhibitor bioactivities into standardized KIBA scores.
	Davis Dataset [125]	Comprehensive mapping of kinase–inhibitor dissociation constants (K_d) over multiple enzymes.
	PDBbind Dataset [126]	Annotated set of biomolecular complexes with experimentally determined binding affinities and structures.
Molecular Interaction Databases Drug–Drug Interaction	TwoSIDES [127]	Pharmacovigilance resource of adverse drug–drug event pairs mined from FAERS reporting data.
	Deng's Dataset [128]	Multimodal catalog of 570 approved drugs' interactions stratified by 65 mechanistic event types.
	ChChMiner [129]	A BioSNAP sub-dataset of 1,514 FDA-approved drugs and 48,514 DDI.
	DrugCombDB [130]	Dataset that contains 448,555 combinations of 2,887 drugs across 124 cancer cell lines, labeled as synergistic or antagonistic using multiple scoring models.
	O'Neil's dataset [131]	A dataset that contains 583 drug combinations across 39 cancer cell lines, identifying 287 synergistic and 178 antagonistic pairs among 38 drugs
	AstraZeneca's dataset [31]	A dataset that features 910 combinations of 118 drugs across 85 cell lines, with 797 pairs showing high synergy

6.1 Comprehensive Databases

Comprehensive databases are those that contain extensive molecular and biochemical information on drugs and chemical compounds, including but not limited to compound identifiers, structural representations (e.g., SMILES, 2D and/or 3D graphs), indications, and target information. Such data supports a wide range of applications. The compound structure information usually serves as input for computational models and their labels and properties serve as training data. In this subsection, we include three representative comprehensive databases in drug discovery: DrugBank [113], PubChem [114], and MoleculeNet [21].

DrugBank [113] is a comprehensive, freely accessible, online database containing reliable information on drugs and drug target and is a vital resource for computational drug discovery and pharmaceutical research. The latest release features more than seventeen thousand drug entries, including FDA-approved small molecule drugs, FDA-approved biotech (protein/peptide) drugs, nutraceuticals, and experimental drugs. For each drug entry, DrugBank contains chemical, pharmacological, and pharmaceutical properties of the drug as well as links to external databases. In addition, DrugBank also provide sequence, structure, and pathway information of around six thousand unique proteins, which are drug targets/enzymes/transporters/carriers associated with these drugs. The information about drug structures, indications, drug-target interactions, and pathways can support a wide range of tasks such as drug property prediction, drug activity analysis, drug repurposing, and drug-target interaction prediction.

PubChem [114] is another comprehensive database of chemical molecules and their activities against biological assays, which is maintained by the National Center for Biotechnology Information (NCBI). It serves as a comprehensive resource for information on the chemical structures, properties, biological activities, and

toxicity of small molecules, and is widely used in cheminformatics, bioinformatics, and computational drug discovery. PubChem is organized into three main interlinked databases: PubChem Compound, PubChem Substance, and PubChem BioAssay (PCBA). PubChem Compound database contains information of more than 100 million pure and characterized chemical compounds. The Substance section collects information of substances, including mixtures and uncharacterized substances, submitted by various data contributors. The BioAssay section contains bioactivity results from approximately 1.67 million biological assay experiments. PubChem Compound IDs are widely used across chemical databases for consistent referencing.

MoleculeNet [21] is a benchmarking platform designed to facilitate the development and evaluation of machine learning models for molecular property prediction. The authors curated a wide variety of datasets from other primary sources, covering different molecular properties and tasks. Although it is not as complex as the two database mentioned earlier, it had been utilized frequently in evaluating newly proposed machine learning approaches because the datasets were constructed for specific tasks and were organized in a very simple format for download. Briefly, the datasets cover four different types of properties, including Quantum Mechanics (including datasets QM7, QM8, QM9), Physical Chemistry (datasets ESOL, FreeSolv, Lipophilicity), Biophysics (datasets PCBA, MUV, HIV, BACE), and Physiology (datasets BBBP, Tox21, SIDER, ClinTox). The prediction tasks can be either classification or regression. As a reference, we provide a very brief summary for each of the datasets.

QM7, QM8, QM9 provide quantum mechanical properties and 3D molecular geometries that can be used as training data for quantum property prediction. QM7 includes 7,165 molecules computed atomization energies and Coulomb matrices. QM8 includes 21,786 molecules with calculated electronic spectra. QM9 expands to over 133,000 stable organic compounds with detailed quantum mechanical properties including energies, geometries, and vibrational frequencies.

ESOL is small dataset of 1,128 molecules in SMILES format offering water solubility data, useful for evaluating solubility predictions. **FreeSolv** contains 642 small molecules with both experimental and computed hydration free energies. **Lipophilicity** reports log D values of the octanol–water distribution coefficients for over 4,200 drug molecules, reflecting membrane permeability and solubility.

PCBA contains activity profiles for over 400,000 molecules against specific enzymes, receptors, and pathways, derived from PubChem BioAssay database. **MUV** is a filtered subset of PubChem BioAssay, designed to validate virtual screening techniques and includes 17 benchmark tasks. **HIV** contains more than 41 thousand molecules labeled for their ability to inhibit HIV replication based on biological assay data. **BACE** includes 1,513 inhibitors of human β-secretase 1 (BACE-1), with both binary activity labels and IC50 values.

BBBP includes more than two thousand molecules labeled based on whether they can cross the blood-brain barrier. **Tox21** contains toxicity data for close to eight thousand compounds across 12 targets, used in toxicology modeling. **ToxCast** extends Tox21, with bioactivity measurements on 617 biological targets for 8,576 compounds. **SIDER** focuses on close to fifteen hundred marketed drugs and their recorded adverse drug reactions (more than five thousand side effects). **ClinTox** contains approved drugs and compounds that failed clinical trials due to toxicity concerns.

Each dataset is accompanied by task definitions (e.g., classification or regression), standard metrics (e.g., ROC-AUC, RMSE), and data preprocessing techniques (e.g., scaffold splits, random splits) to promote consistent model evaluation.

6.2 Clinical Databases

Clinical databases are those databases that contain clinical and health information from patients that can be used for disease prediction, treatment outcome modeling, as well as drug recommendation and preci-

sion medicine. We therefore list two popular databases here: Medical Information Mart for Intensive Care database (MIMIC-III [115] and MIMIC-IV [116]), and the UK Biobank database [117].

MIMIC-III [115] and **MIMIC-IV** [116] are freely accessible, large-scale clinical databases developed by the MIT Lab for Computational Physiology. While MIMIC-IV is an updated and improved version of MIMIC-III, and there are overlapped samples in the two database, MIMIC-IV does not encompass all the data present in MIMIC-III. We briefly discuss both databases. MIMIC-III contains de-identified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center (BIDMC) between 2001 and 2012. The database includes a wide range of data types across 26 tables, such as demographics, vital signs, laboratory test results, medications, diagnostic codes (ICD-9), procedures, imaging reports, and clinical notes. Its structured and time-stamped data makes it especially valuable for developing and validating models for disease risk and patient trajectory prediction. In addition, the dataset has also been frequently employed in studies on drug recommendations and drug combination recommendations after extensive preprocessing. On the other hand MIMIC-IV includes detailed, de-identified clinical data for 180,733 patients for hospital admissions and 50,920 patients for ICU admissions from BIDMC between 2008 and 2019. Information available includes patient measurements, orders, diagnoses, procedures, treatments, and de-identified free-text clinical notes. Both datasets support a wide array of research studies and help to reduce barriers to conducting clinical research using patient level data.

UK Biobank [117] is another large-scale biomedical database and research resource containing in-depth genetic, lifestyle, and health information from approximately 500,000 volunteer participants aged 40–69 at the time of recruitment (2006–2010) across the United Kingdom. It is managed by a charitable organization and made available to approved researchers for health-related research. The dataset includes a broad array of data types, such as genotyping and whole-genome sequencing, biochemical assays, physical measures, imaging data (e.g., MRI, CT scans), hospital and primary care records, and detailed lifestyle and demographic questionnaires. UK Biobank is particularly valuable for population-level studies on complex diseases. The integration of genetic with phenotypic and clinical data makes it one of the most important resources for identifying drug targets, predicting drug effects, and accelerating drug development.

6.3 Structural Information Databases

Structural information datasets focus on providing 3D structures and/or conformers of isolated ligand and/or protein–ligand complexes. These resources combine experimental structures with computationally refined conformations to support a range of applications – from physics-based simulations such as free energy calculations to data-driven machine learning models that predict binding affinity or molecular properties. In this subsection, we highlight four widely used structural datasets: **ZINC** [118], **GEOM** [119], **MISATO** [84], and **CrossDocked** [121].

ZINC [118] is a meticulously curated repository containing over 230 million commercially accessible compounds. It includes 3D structures, physicochemical properties, and vendor metadata. Unlike theoretical libraries, ZINC focuses on experimentally testable molecules, facilitating streamlined drug discovery workflows. The database provides SMILES strings, 3D structures, and drug-like property classifications, organized into tranches for targeted virtual screening. As a benchmark for docking and virtual screening studies, ZINC accelerates the transition from computational predictions to experimental validation, serving as a critical tool for hit identification and lead optimization.

GEOM [119] contains quantum mechanics-optimized 3D geometries for approximately 30 million conformers representing 450,000 drug-like molecules. Each structure is refined using density functional theory (DFT) to capture realistic conformational landscapes, emphasizing ensembles of energetically feasible states rather than static geometries. This ensemble approach advances protein-ligand interaction modeling,

conformer generation algorithms, and force field validation. Its high accuracy has substantially advanced 3D-aware machine learning models for molecular property prediction and molecule generation.

MISATO [84] is a machine learning-oriented dataset comprising roughly 20,000 experimentally resolved protein–ligand complexes. Each complex undergoes structural refinement and quantum-mechanical optimization to address stereochemical, geometric, and protonation inconsistencies. Around 17,000 complexes are further subjected to explicit-solvent molecular dynamics (MD) simulations. This dynamic data captures conformational flexibility and binding pocket dynamics. In addition, MISATO includes quantum-derived electronic descriptors, partial charges, and preprocessing utilities tailored for machine learning pipelines. By integrating static structures with time-resolved dynamics, MISATO enables modeling of transient binding states and induced-fit effects, overcoming the limitations of single-conformation datasets.

CrossDocked [121] curates 18,450 non-redundant protein–ligand complexes derived from the Protein Data Bank (PDB) [132], using a systematic cross-docking approach. Ligands are docked into non-cognate, structurally similar binding pockets to produce a diverse set of over 22.5 million binding poses. The dataset features cluster-based predefined splits for evaluating model generalizability to unseen targets and provides dual metrics for assessing both pose accuracy and binding affinity predictions. By modulating the structural similarity between docking receptors and their native counterparts, CrossDocked enables rigorous evaluation of docking algorithms under realistic scenarios where exact receptor structures may be unknown. Designed as a comprehensive benchmark, CrossDocked supports the standardized training and evaluation of 3D CNNs and other ML models for non-native protein–ligand interaction modeling, with broad implications for virtual screening.

6.4 Molecular Interaction Databases

Molecular interactions databases record relationships among different molecules in various formats and are fundamental in the study of molecular biology, computational chemistry, and drug discovery. Researchers use these datasets to elucidate biochemical pathways, predict binding affinities, evaluate selectivity, and simulate off-target effects. We include datasets that capture a range of molecular interactions, grouped into two categories: protein–ligand binding and drug–drug interactions.

6.4.1 Protein-Ligand Binding Databases

Protein–ligand binding describes the specific interactions between proteins (often therapeutic targets or receptors) and small-molecule ligands (including drugs). These interactions drive most biochemical modulation and are critical for drug discovery, off-target prediction, and mechanistic studies. Below, we summarize several widely used public datasets.

ChEMBL [122] is a manually curated database focusing on bioactive molecules with drug-like properties. To assess the binding affinity of small-molecule ligands to their targets, ChEMBL primarily uses experimental bioactivity data extracted from scientific literature. To facilitate comparison and analysis, data from different sources undergoes a standardization process so that measurement type, value, and units are comparable. In addition to binding affinity measurements, ChEMBL also contains rich information about compounds, targets, experimental assays, and original sources. The current release of ChEMBL (release 35) includes approximately 2.5 million distinct compounds and 21.1 million bioactivity measurements derived from 1.7 million biological assays across 16,000 biomolecular targets. ChEMBL supports a wide range of applications, including structure–activity relationship analysis, off-target prediction, and drug repurposing. Its datasets are available via web interfaces, APIs, and bulk downloads under open Creative Commons licenses.

PDBbind Dataset [126] was created to collect experimentally measured binding data from literature for the biomolecular complexes with high-resolution 3D structures in the Protein Data Bank (PDB). It provides an essential linkage between the energetic and structural information of those complexes, which is helpful for various computational and statistical studies on docking validation, scoring-function development, affinity prediction, molecular recognition, and drug discovery. The most recent version (2024) was released on a commercial platform called PDBbind+ with a free demo version. It currently contains experimental binding affinity data for 27,385 protein-ligand complex, 4,594 protein-protein complex, 1.440 protein-nucleic acid complex and 234 nucleic acid-ligand complex.

Metz Dataset [123] focused exclusively on kinase inhibition activities. It contains over 150,000 kinase inhibitory measurements, comprising more than 3,800 compounds tested against 172 different protein kinases. Based on these measurements, the authors constructed a comprehensive kinome interaction network, enabling systematic analysis of kinase–inhibitor interactions. This dataset is applicable not only to binding affinity prediction but also to the design of multi-kinase inhibitors.

Davis Dataset [125] was developed by Davis et al. to provide a broad target panel covering over 80% of the human catalytic kinome. It includes selectivity profiles of 72 kinase inhibitors tested against 442 human kinases. With more than 30,000 high-precision measurements obtained from a standardized binding assay, the dataset supports various tasks such as binding affinity prediction, selectivity profiling, off-target prediction, and regression model validation. The uniform experimental design and extensive kinase coverage make the dataset a preferred benchmark for visually screening compound libraries and identifying novel inhibitors.

Via a systematic evaluation of target selectivity profiles across three different biochemical assays of kinase inhibitors, Tang et al. [124] introduced a model-based approach and a unified affinity metric known as the "KIBA score", to integrate complementary information captured by different bioactivity types. The resulting **KIBA Dataset** comprises a drug–target bioactivity matrix involving 52,498 chemical compounds and 467 kinase targets, with a total of 246,088 KIBA scores. This statistically harmonized dataset, designed to minimize experimental variability, has become a widely used benchmark for training machine learning models in drug–target affinity prediction.

6.4.2 Drug–Drug Interaction Databases

While comprehensive resources such as DrugBank include DDI information, they often require additional processing to extract pure interaction data. Furthermore, they may not include all the measurements from high-throughput screening assays, for example, dose responses for different dose combinations of drug pairs. Below, we highlight some widely used datasets for both adverse and synergistic DDI prediction tasks. For adverse or general DDI prediction, commonly used datasets include TWOSIDES [127], Deng's Dataset [128], and ChChMiner [129].

TwoSIDES [127] is a database of polypharmacy side effects for pairs of drugs. It was constructed by mining the U.S. Food and Drug Administration (FDA) adverse event reporting systems (FAERS) [133]. The dataset consists of 868,221 statistically significant association between 59,220 drug pairs and 1,301 adverse events. Only associations that cannot be clearly attributed to either drug alone were included. It improves the detection and prediction of adverse effects of drug interactions.

Deng's Dataset [128] consists of 74,528 distinct drug–drug interactions among 572 approved drugs extracted from DrugBank entries by appling NLP algorithms. Each interaction is recorded as a four-element tuple: (*drug A*, *drug B*, *mechanism*, *action*), where the 'mechanism' means the effect of drugs in terms of the metabolism, the serum concentration, the therapeutic efficacy and so on. The 'action' represents increase or decrease. The categorization of interactions into different types of mechanism of actions is

useful for understanding the actual logic hidden behind the combined drug usage or adverse reactions and for evaluating algorithms that aim to recover them.

ChChMiner is a sub-dataset from the Stanford Biomedical Network Dataset Collection (BioSNAP) [129], which itself is a comprehensive database providing information about relationships between a variety of biological, chemical, or clinical entities including drugs. ChChMiner is a network of interactions among 1,514 FDA-approved drugs, extracted from drug labels, scientific publications, and DrugBank, with 48,514 drug–drug interactions.

For synergistic DDI prediction, we include an aggregated dataset from various sources (DrugCombDB [130]) and two high-throughput screening data of drug pairs on cancer cell lines (O'Neil's Dataset [131] and AstraZeneca's Dataset [31]). **DrugCombDB** [130] aggregated drug combination data from various data sources: high-throughput screening assays of drug combinations, manual curations from the literature, FDA-approved therapies and failed clinical trials, as well as some earlier drug combination databases such as DCDB [134]. The database comprises 448,555 drug combinations involving 2,887 unique drugs and 124 human cancer cell lines. In addition, DrugCombDB has more than 6,000,000 quantitative dose responses from which multiple synergy scores to determine the overall synergistic or antagonistic effects of drug combinations were calculated based on different models. As a comprehensive database with a large number of drug combinations, DrugCombDB would greatly facilitate and promote the discovery of novel synergistic drugs for the therapy of complex diseases.

By using a high-throughput platform for unbiased identification of synergistic and additive drug combinations, O'Neil et al. [131] created a dataset comprising 38 experimental or approved drugs tested across 39 diverse cancer cell lines, using a 4-by-4 dosing regimen for a total of 583 drug–drug combinations. More recently, a similar but larger dataset was created by AstraZeneca and shared with the research community through a DREAM Challenge [31], which consisted of 11,576 experiments from 910 combinations of 118 drugs across 85 molecularly characterized cancer cell lines. Both datasets can serve as benchmark datasets and can accelerate the development of computational approaches for drug combination synergy prediction.

6.5 Challenges in Data

Deep learning models require large volumes of high-quality data to effectively capture the complex relationships between molecular entities and to predict properties of chemical compounds. However, several limitations still hinder progress. First of all, substantial heterogeneity exists in different databases. Data often comes in different formats with different identifiers, making integration and standardization difficult. Data derived from different labs or measurement techniques may not be comparable, limiting transferability. In the extreme cases, different sources may report contradictory effects or properties for the same compound.

Secondly, considerable biases exist in different data sources. Almost all the databases have the coverage bias: most datasets focus on well-studied proteins, pathways, diseases, and only a small portion of theoretically limitless set of all possible chemical compounds, leaving gaps in knowledge for novel targets, rare diseases, and novel drug candidates. Underrepresented molecular classes such as biotech medicines/biologics are often excluded due to lack of structured data. In addition, many datasets consist of label biases, including imbalance labels, and noisy and incomplete labels. Many datasets consist of only positive (or negative) samples because of the nature of the datasets. For example, drug-drug interactions were mostly reported when side effects were observed, which only provided positive samples. On the contrary, failed clinical trials only contained negative results. Label imbalance makes the prediction challenging and motivates the development of pre-training and few-shot learning algorithms. Some benchmark datasets, especially those with smaller sizes, have dataset specific biases. Models trained on those datasets may suffer from overfitting and not generalize well.

Finally, there are limitations on the scopes of data. Many benchmark datasets are static and not routinely updated with new discoveries. Furthermore, due to privacy concern and regulatory and ethical constraints, individual level, patient-centric data is not accessible to the broad research community. In particular, few datasets combine molecular data with real-world patient electronic health record (EHR) data, hindering the progress in personalized medicine.

7 Discussion and Conclusion

In recent years, deep learning approaches – particularly GNNs – have garnered increasing attention in the field of computational drug discovery. In this survey, we systematically review studies published since 2021, highlighting the significant role GNNs have played across three core application areas: molecule generation, molecular property prediction, and drug–drug interaction prediction. We examine how GNNs effectively model chemical structures and capture complex molecular patterns, discussing the strengths and limitations of current approaches, along with the major challenges faced by the research community. The papers included in this review represent state-of-the-art advancements in the field and demonstrate that molecular graph representation learning has become a dominant paradigm across these applications.

Specifically, we observe that GNNs have been instrumental in advancing molecule generation by enabling the design of novel compounds with desired properties through both unconstrained and constrained generation strategies. In the area of molecular property prediction, the shift towards utilizing 3D molecular graphs has led to more accurate and robust outcomes, particularly when combined with message-passing mechanisms and contrastive learning techniques. For DDI prediction, GNNs have opened promising avenues in personalized medicine by identifying safe and effective drug combinations tailored to individual patient profiles. This is especially impactful for drug repurposing, where GNNs can significantly accelerate the development of combinatorial therapies.

Across the reviewed studies, several common trends emerge. First, pre-training and self-supervised learning have become widespread, substantially enhancing the performance of GNN-based models. These techniques are particularly effective in mitigating the issue of limited labeled data and contribute to the development of more generalizable models. Second, the incorporation of domain-specific knowledge into GNN architectures has led to noticeable improvements in model performance, suggesting a movement toward more specialized and biologically informed models. Third, many recent works adopt multi-modal approaches that integrate diverse input formats, including 2D and 3D molecular graphs, as well as SMILES strings, sometimes combining GNNs with other deep learning architectures. This fusion of complementary molecular representations enables a more comprehensive understanding of the data and has the potential to significantly enhance predictive performance.

Despite these advancements, several challenges remain, which also point to promising future directions. First, as previously discussed, data scarcity and the limited availability of high-quality, diverse datasets continue to constrain the full potential of GNNs in drug discovery. Second, the interpretability of GNN models remains a critical hurdle that must be addressed to foster trust and facilitate their adoption in real-world applications. As GNN architectures grow increasingly complex, ensuring interpretability is paramount. As emphasized by Henderson et al. [135], future research should prioritize models that not only provide accurate predictions but also offer clear explanations for their decisions. Bridging this gap between computational predictions and human understanding will be essential for generating actionable scientific insights. Lastly, multi-omics integration approaches combine data from various biological levels (genomics, transcriptomics, proteomics, metabolomics, epigenomics, etc.) to create a more comprehensive understanding of disease mechanisms. Integration of GNN-based models with multi-omics data holds significant promise and could

further revolutionize the landscape of drug discovery by identifying better drug targets, developing personalized therapies, and eventually improving treatment outcomes.

8 Competing Interests

No competing interest is declared.

9 Author Contributions Statement

Zhengyu Fang: Conceptualization, Writing – Original Draft. Xiaoge Zhang: Writing – Original Draft, Writing – Review & Editing. Anyin Zhao: Writing – Original Draft, Writing – Review & Editing. Xiao Li: Supervision. Huiyuan Chen: Conceptualization, Investigation. Jing Li: Supervision, Conceptualization, Writing – Review & Editing. All authors reviewed and approved the final manuscript.

10 Acknowledgments

This work is supported in part by NSF CCF-2200255, NSF CCF-2006780, NSF IIS-2027667, NIH U01AG073323, NIH R01HG009658, NIH 1R01HL159170 and NIH 1R01NR02010501.

References

[1] Mullard, A. New drugs cost us 2.6 billion to develop. *Nature reviews drug discovery* (2014).

[2] Gawehn, E., Hiss, J. A. & Schneider, G. Deep learning in drug discovery. *Molecular informatics* **35**, 3–14 (2016).

[3] Chen, H. & Li, J. Modeling relational drug-target-disease interactions via tensor factorization with multiple web sources. In *The World Wide Web Conference*, 218–227 (2019).

[4] Chen, H. & Li, J. DrugCom: Synergistic discovery of drug combinations using tensor decomposition. In *2018 IEEE International Conference on Data Mining (ICDM)*, 899–904 (IEEE, 2018).

[5] Wang, W., Yang, S., Zhang, X. & Li, J. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* **30**, 2923–2930 (2014).

[6] Xu, M. *et al.* An end-to-end framework for molecular conformation generation via bilevel programming. In *International conference on machine learning*, 11537–11547 (PMLR, 2021).

[7] Swanson, K., Williams, J. L. & Jonas, E. M. Von mises mixture distributions for molecular conformation generation. In *International Conference on Machine Learning*, 33319–33342 (PMLR, 2023).

[8] Maziarz, K. *et al.* Learning to extend molecular scaffolds with structural motifs. In *International Conference on Machine Learning* (2021).

[9] Geng, Z. *et al.* De Novo molecular generation via connection-aware motif mining. In *International Conference on Learning Representations (ICLR)* (2023).