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

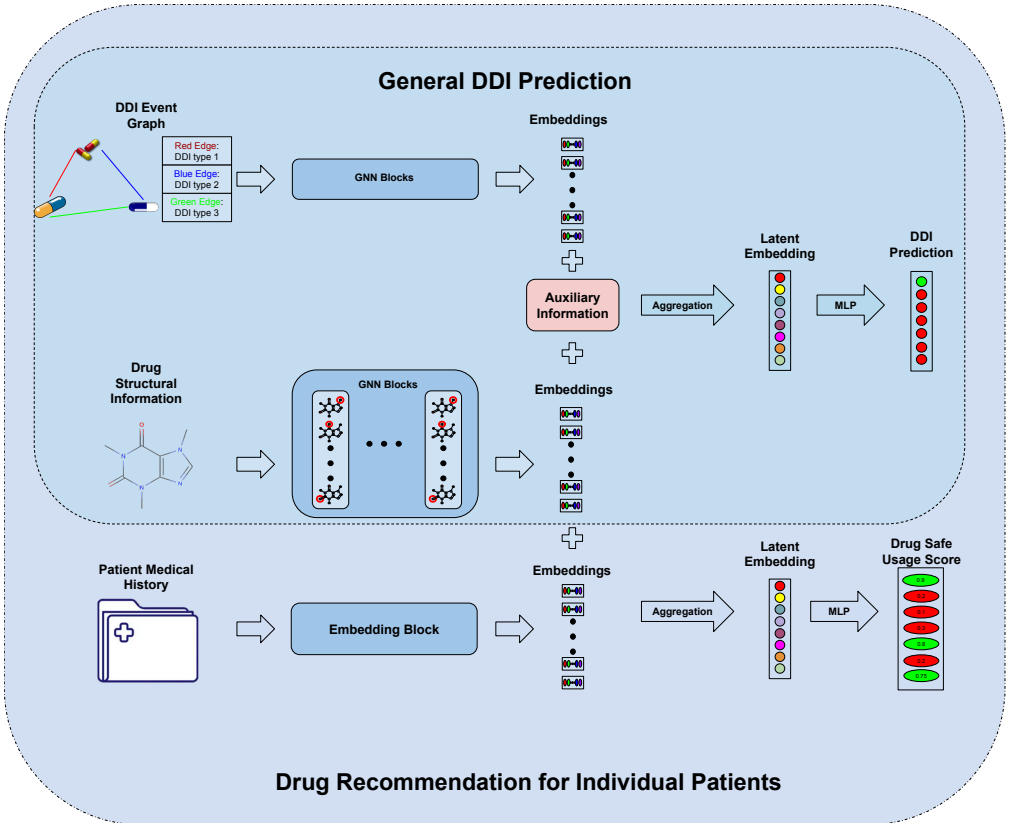

Figure 4: The general process of DDI prediction based on GNN models. Possible inputs for the general DDI prediction (the one inside the small rectangle) include DDI event graphs, and drug molecular structures, either individually or jointly. Additional auxiliary data can be incorporated into the models. 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