Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Graph neural network approaches for drug-target interactions

Zehong Zhang[1,2], Lifan Chen[1,2], Feisheng Zhong[1,2], Dingyan Wang[1,2], Jiaxin Jiang[1], Sulin Zhang[1,2], Hualiang Jiang[1,2,3], Mingyue Zheng[1,2] and Xutong Li[1,2]

Abstract

Developing new drugs remains prohibitively expensive, time-consuming, and often involves safety issues. Accurate prediction of drug-target interactions (DTIs) can guide the drug discovery process and thus facilitate drug development. Non-Euclidian data such as drug-like molecule structures, key pocket residue structures, and protein interaction networks can be represented effectively using graphs. Therefore, the emerging graph neural network has been rapidly applied to predict DTIs, and proved effective in finding repositioning drugs and accelerating drug discovery. In this review, we provide a brief overview of deep neural networks used in DTI models. Then, we summarize the database required for DTI prediction, followed by a comprehensive introduction of applications of graph neural networks for DTI prediction. We also highlight current challenges and future directions to guide the further development of this field.

Addresses

[1] Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China

[2] University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing 100049, China

[3] School of Life Science and Technology, ShanghaiTech University, 393 Huaxiazhong Road, Shanghai 200031, China

Corresponding authors: Zheng, Mingyue (myzheng@simm.ac.cn); Li, Xutong (lixutong@simm.ac.cn)

Current Opinion in Structural Biology 2022, **73**:102327

This review comes from a themed issue on **Artificial Intelligence (AI) Methodologies in Structural Biology**

Edited by **Feixiong Cheng** and **Nurcan Tuncbag**

For complete overview of the section, please refer the article collection - Artificial Intelligence (AI) Methodologies in Structural Biology

Available online 21 January 2022

https://doi.org/10.1016/j.sbi.2021.102327

Introduction

Most drugs achieve therapeutic effects through in vivo interactions with specific target molecules such as enzymes, nuclear receptors, G-protein coupled receptors (GPCRs), and ion channels [1]. Therefore, the identification of drug-target interactions (DTIs) is an important area in the drug discovery pipeline, including lead generation and optimization, drug repositioning, polypharmacology, virtual screening and other related fields [2]. As traditional pharmacology assays for DTIs identification are costly and time-consuming [3], there is high demand for accurate computational determination of DTIs in order to effectively complement experimental wet-lab techniques by narrowing the search space for subsequent wet experiments and thus accelerating drug development [4].

Machine learning methods have long served as important tools in drug discovery [5]. Traditional machine learning methods such as Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), K-Nearest Neighbour (KNN), and artificial neural networks (ANNs) are widely used in quantitative structure activity relationship (QSAR), proteochemometric (PCM) approach, and molecular docking to model DTI [6].

DTI models based on traditional machine learning methods can take advantage of high-dimensional complex data, which is usually constructed using predefined chemical and protein descriptors and fingerprints. Deep learning approaches, including the deep neural network (DNN) and its variants, are capable of learning data representation directly without using predefined descriptors and thereby have shown promising potential in DTI predictions. DTI models build on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) can automatically extract features from texts (protein sequences and SMILES) and images (3D grids) [7,8]. However, all these features are Euclidean data. In fact, chemical molecules graph [9] and protein structures [10] can be processed into non-Euclidean form and used as inputs of graph neural networks (GNNs), as well as other data associated with DTIs such as protein–protein interaction networks [11] and knowledge graphs [12]. The GNN has recently become a widely used deep learning architecture because of its impressive performance and high interpretability [13]. Its applications on DTI prediction have achieved some

measure of success. This review provides a brief overview of neural networks used in DTI models, and then summarizes the advances of the investigations that have applied GNNs on DTI prediction. Finally, we discuss the challenge and developmental trend of the applicability of deep learning algorithms in the field of DTI prediction.

Brief introduction of neural networks

In the past few years, deep learning has achieved phenomenal success in solving many challenging tasks, such as data mining and classification, and thus aroused considerable interest across various disciplines [14—16]. Newly emerging neural network approaches have shown superior performance in recognizing, processing and extrapolating complex patterns in molecular data than traditional machine learning techniques [8,17].

Derived from the artificial neural network (ANN), CNN, RNN and GNN are three of the most widely used deep learning algorithms. Although they are generally comprised of an input, an output layer, and zero or more hidden layers, their application scope is different because of their different input data forms. CNNs and RNNs are efficient in extracting the features of Euclidean data such as images (2D grids) and texts (1D sequences). Euclidean data could be sampled on a grid and sensibly modelled as being plotted in n-dimensional linear space [18]. They could be considered special cases of graphs whose nodes are arranged regularly. This kind of data has translation invariance and local connectivity. Take the image as an example, by regarding pixels as nodes, every node has the same number of neighbor nodes. The position of nodes can be described in Cartesian coordinates, and the Euclidean distance between two nodes in the mapping space can represent the distance in the real world. Therefore, the same structural information in the image can be extracted in a same way by defining a globally shared convolution kernel (Figure 1 (a)). On the other hand, non-Euclidean data has irregular shapes and sizes, whose nodes do not have a spatial order and thus do not have properties of translation invariance, local connectivity or a common system of coordinates. Consequently, it is hard to define a localized convolutional kernel to extract non-Euclidean data such as structural information of target—ligand complex and PPI network (Figure 1 (b)). GNNs take these types of data as graphs, namely sets of objects (nodes) and their relationships (edges), to learn low-dimensional node embedding or graph representation. These embeddings are then employed to solve many graph analysis tasks, such as node classification, graph classification, and link prediction.

More specifically, the hidden layer of CNN [19] includes three major building blocks: convolution layer, pooling layer, and full connection layer. The function of the convolution layer is tantamount to extracting features from the input data (Figure 1 (c)). Features extracted by a convolution operation are often local, and thus multi-layer convolution is needed to extract global features. After feature extraction at the convolution layer, the output feature map will be transferred to the pooling layer for feature selection and information filtering.

Although general CNN can model 1D sequences, it is difficult for CNN to complete such tasks if the sequence length is variable. For example, in the translation task, the number of words in inputs and outputs is variable. As a sequence can be broken down into a number of unit tokens, RNNs treat each sequence token as an individual input/output by using their internal state (memory) [20]. Besides, as the unit tokens in the same sentence are not independent of each other, RNNs differ from CNNs because their hidden layers can receive the hidden state of the previous moment, which can be thought of as the memory of previous instances (Figure 1 (d)).

The graph neural network (GNN) has recently become a widely used deep learning architecture because of its impressive performance and high interpretability [13]. GNNs process non-Euclidean data as graphs, aiming to transform data into a low-dimensional and more discriminative feature space on the premise of maintaining some geometric characteristics of the current space through representation learning, namely graph embedding. DeepWalk [21] is the most representative method. It samples nodes in the graph by random walk, and then learns the vector representation of nodes by using a method similar to word2vec [22]. Four key ideas behind CNN inspired the proposal of GNN: local connection, shared weight, pooling, and the use of multiple layers [23]. Graphs are typical locally connected structures; shared weights reduce the computational cost; pooling layers can merge semantically similar features; multi-layer structure captures features of various sizes. According to different calculation methods, GNNs can be divided into spectral graph convolution and spatial graph convolution. The spectral domain graph convolutional networks (GCN) [24] transforms graph signal from the spatial domain to the spectral domain with Fourier basis and then defines graph convolution operation in the spectral domain. For the spatial domain, a representative example is GraphSAGE [14], which targets a learning aggregator whose job is to complete information aggregation of neighboring nodes. Furthermore, by combining some techniques, some variants of the GNN have demonstrated ground-breaking performances in many deep learning tasks. GAT [25], for instance, is produced by combining GNN and the self-attention mechanism.

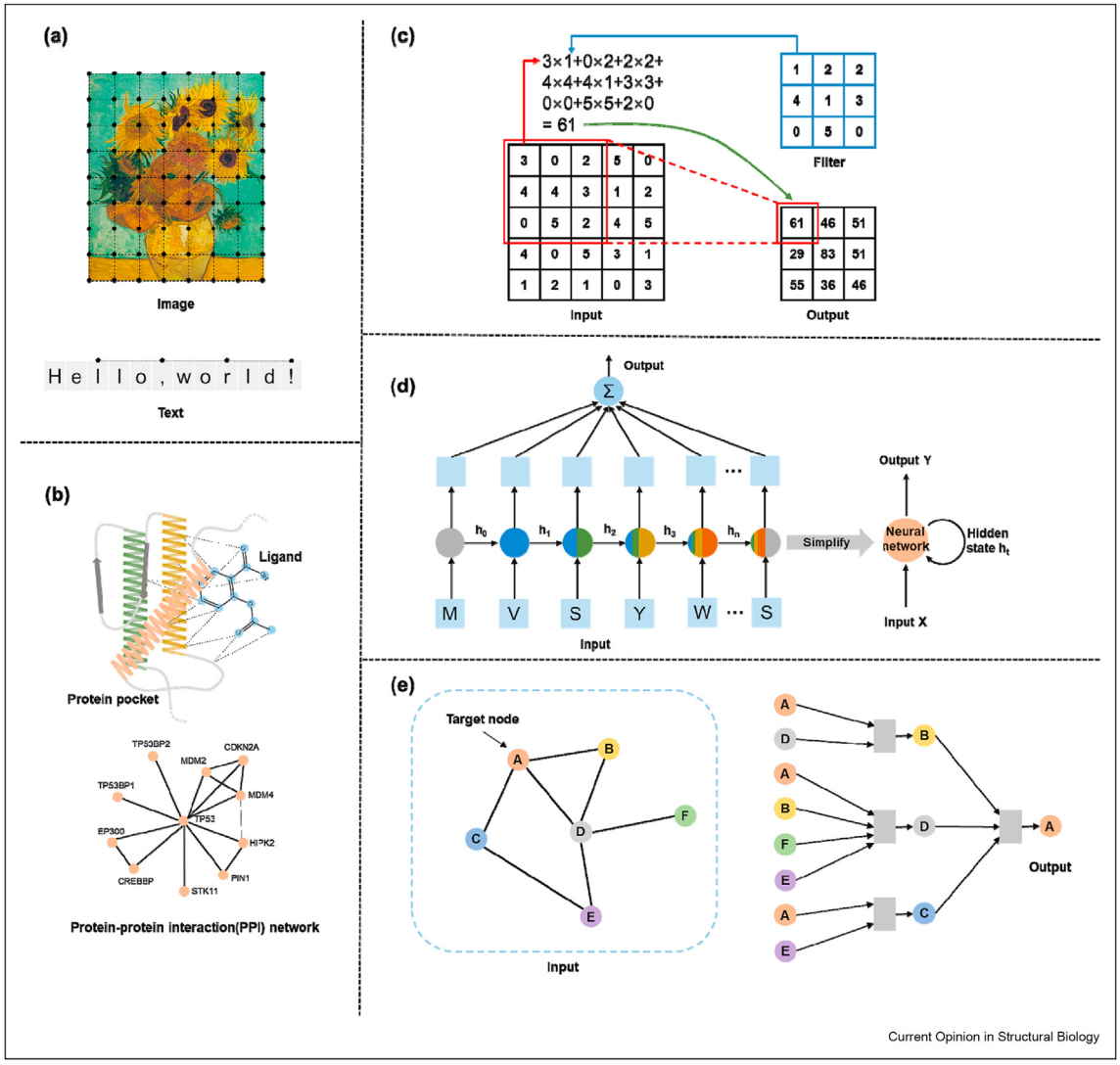
Conventional neural networks such as CNNs and RNNs stack the features of nodes in a specific order and traverse all possible sequences to characterize a graph completely,

which increases computational cost [13]. GNNs, however, are capable to circumvent this defect because the feature aggregation and propagation on each node does not depend on its order or the number of its neighborhood nodes (Figure 1 (e)), making it suitable for non-Euclidean graphs. For example, by regarding atoms as nodes and bonds as edges, chemical structural formulae are typical non-Euclidean graphs whose nodes are not naturally ordered. On the other hand, molecule structures can be transformed into unique text sequences, namely SMILES (Simplified Molecular Input Line Entry System) [26]. But the text sequence implicitly specifies the order of the atoms that may interfere with the model. On the basis of DeepDTA [17], GraphDTA [27] changed part of the CNN layer to the GNN layer without changing the structure of other parts of the model. On the same dataset, the performance of GraphDTA has improved to some extent. Hence, compared with CNN, GNN may be more suitable for extracting information from chemical structures. More details about the GNN algorithm could be gained in the review [13].

Databases

Getting enough high-quality data is a prerequisite for an accurate and efficient AI model. Driven by the technical advancement of high-throughput screening (HTS) [28] and parallel chemical synthesis [29], high-quality datasets of millions of molecules and their profile against a

Figure 1



Euclidean data, non-Euclidean data and comparison of neural networks. **(a)** Euclidean data, including image and text; **(b)** Non-Euclidean data, including chemical structure and interaction network; **(c)** The calculation method of CNNs' convolutional layer; **(d)** The structure of RNN; **(e)** Information aggregation of GNN. The information of the neighbor node is propagated to node A through two information aggregations.

multitude of biological targets are growing rapidly, making it possible to build in silico DTI models. In this section, we provide a summary of databases related to the DTI model in Table 1.

These databases can be divided into three categories: interaction databases, protein structure databases, and benchmark databases. Interaction databases are the broadest and include molecular property and activity (ChEMBL [35], PubChem [42], CTD [49]), clinical information (DrugBank [30], DrugCentral [40]), genomic information (KEGG [51], DisGeNET [44]), compound–protein interactions (Matador [34], PDBbind [37]), protein–protein interactions (HuRI [53], HPRD [46]) and drug–gene interaction (DGIdb [43]). We can use the information in the interaction databases to construct interaction networks. The structure data of the compound can also be found in the interaction databases, such as PubChem [42].

Protein structure databases mainly include protein sequences and 3D structure information. Benchmark database can be used to construct external test set s to evaluate the performance of the model. DUD-E [56] can establish the data of decoy compounds to verify the effect of the model. On the basis of PubChem bioactivity data, MUV [57] was established as an unbiased baseline dataset for virtual screening. The repoDB [58] database contains data on the success and failure of drug repositioning experiments. More detailed descriptions and comparisons of these databases can be found in Refs. [59,60].

Graph neural network in drug-target interactions

Structure-based predictions

With advances in theory and computational methods, molecular dynamics and quantum mechanics have been able to produce reliable results for structure-based prediction of ligand-protein binding affinity [61]. However, the huge computational cost limits their use in high-throughput screening. On the other hand, molecular docking methods including DOCK, AutoDock, GOLD, etc., have been used to predict binding affinity. Although these methods speed up computation through principled parameter fitting, their computational accuracy is not satisfactory. Therefore, deep learning methods have been favored by more and more researchers. However, using 3D grid representations of molecules makes 3D CNNs have a high computational cost, while the 1D representation method loses many important features. For example, SMILES (Simplified Molecular Input Line Entry System) [26] is a specification for describing the structure of chemical species, transforming molecule structures into unique 1D text sequences. Although SMILES implicitly contains the structural information of molecules, the correct structural information can only be extracted through specific methods. Using it as an input representation of neural network may lose some structure features. The emerging GNNs are introduced to structure-based DTI prediction workflow (Figure 2(a)). The comparison of docking, 3D CNN, GCN and GAT in predicting DTIs is shown in Table S3.

In structure-based GNN approaches, the input representation can be roughly divided into three categories: 1D sequences that represent proteins and graphs that represent molecules; protein pockets and molecules are represented by graphs separately; graphs for structures of protein pockets in complex with their molecular ligands (Figure 2(b)). The input representations of the model can be obtained by database query or RDKit conversion. After that, the representations of proteins and compounds are obtained through different neural network layers respectively, and then concatenated together. Finally, the output is obtained through training a neural network with the DTI prediction task (Figure 2(a)).

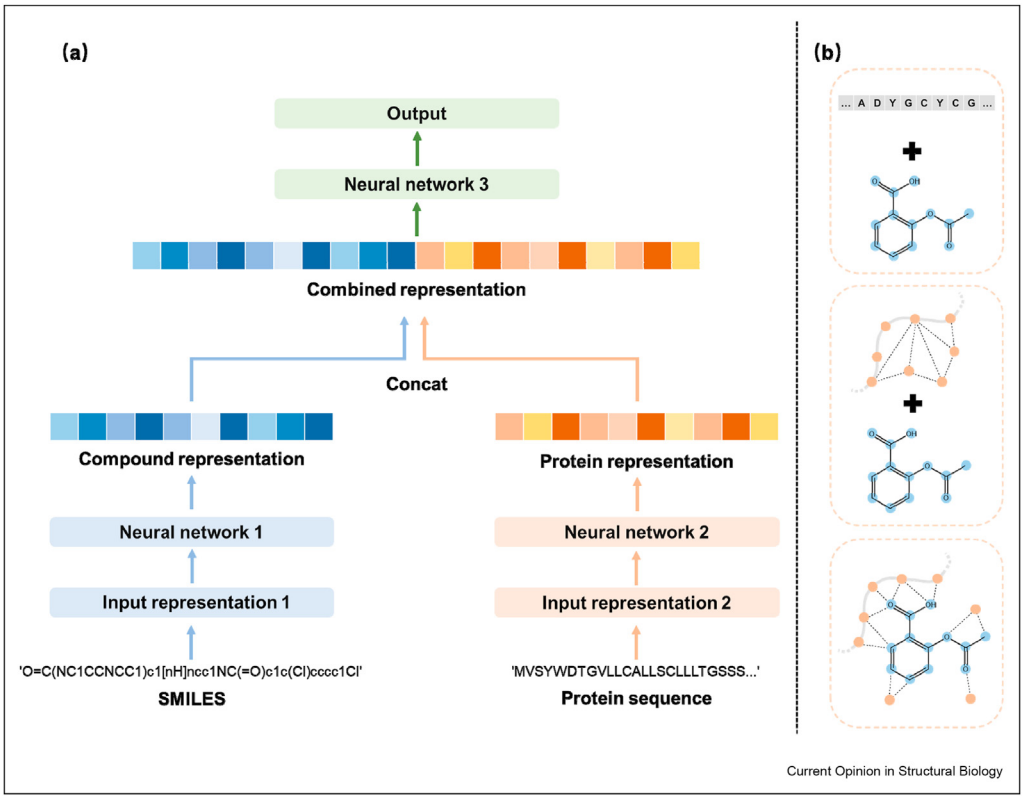
Some structure-based approaches treat the DTI prediction task as a regression task and output final results as continuous values, i.e., drug target affinity DTA prediction. In these tasks, model performance can be significantly improved by introducing GNNs compared to only using CNNs. DeepDTA uses two CNN building blocks to learn representations from SMILES of drugs and protein sequences, respectively, and in conjunction with DNN to predict drug-target affinity values. PADME [62] adds GNNs to DeepDTA and suggests using fixed rule descriptors to represent proteins, rather than learning the underlying feature vectors of proteins directly. PADME is the first method to use molecular graph convolution (MGC) for DTIs. It extracts information from molecular graph representation constructed by SMILES through GraphConv model, and merges the generated latent compound vectors and protein descriptors into Combined Input vector (CIV). Feedforward neural networks receive CIVas the input, and then output a real–value interaction strength as the prediction of DTIs. However, PADME is reported to have similar performance to DeepDTA. GraphDTA [27] uses RDKit [63] to construct molecular graphs and extract atomic features, and describes node features through DeepChem [64], such as the atomic symbol, the total number of hydrogen atoms, and the implicit value of atoms. Unlike PADME, GraphDTA investigates several GNN models to extract the features of molecular graphs. Finally, it is found that the performance of the model using graph-isomorphic network (GIN) [65] in both the Davis dataset [66] and Kiba dataset [67] is better than that of DeepDTA. DGraphDTA [68] constructs not only drug molecular maps, but also uses PconsC4 [69] to construct contact maps based on protein sequences. These two graphs are input into the two GNN building blocks to extract the representations, respectively, and then concatenated for affinity prediction. DGraphDTA not only greatly improves the accuracy of

Table 1

Databases used in papers in this survey.

Category	Database	Link	Main content	Model examples
Interaction databases	DrugBank [30]	https://go.drugbank.com/	Clinical level information and molecular level data about drugs	DeepCPI [31], GraphCPI [32], TriModel [33],
	Matador [34]	http://matador.embl.de/	Protein-chemical interactions	DeepCPI [31], GraphCPI [32]
	ChEMBL [35]	https://www.ebi.ac.uk/chembl/	chemical, bioactivity and genomic data of bioactive molecules	Wen Torng et al. [36]
	PDBbind [37]	http://www.pdbbind.org/(no longer available)	Binding affinities for the protein–ligand complexes	InteractionNet [38], Jaechang Lim et al. [39]
	DrugCentral [40]	https://drugcentral.org/	Active ingredients chemical entities, pharmaceutical products, drug mode of action, indications, pharmacologic action	Wang et al. [41]
	PubChem [42]	https://pubchem.ncbi.nlm.nih.gov/	Chemical structures, identifiers, chemical and physical properties, biological activities, patents, health, safety, toxicity data, and many others	Wang et al. [41]
	DGIdb [43]	https://www.dgidb.org/	Drug-gene interaction	Wang et al. [41]
	DisGeNET [44]	https://tdcommons.ai/	Human disease-associated genes and variants	Wang et al. [41], SkipGNN [45]
	HPRD [46]	https://www.hsls.pitt.edu/obrc/index.php?page=URL1055173331	Proteomic information pertaining to human proteins	Hafez Eslami Manoochehri et al. [47], GANDTI [48]
	CTD [49]	http://ctdbase.org/	Associations between chemicals, gene products, phenotypes, diseases, and environmental exposures.	Wang et al. [41], Hafez Eslami Manoochehri et al. [47]
	SIDER [50]	http://sideeffects.embl.de/	Drugs and side effects	Hafez Eslami Manoochehri et al. [47]
	KEGG [51]	https://www.genome.jp/kegg/pathway.html	Genomic, chemical and systemic functional information	TriModel [33]
	BIOSNAP [52]	http://snap.stanford.edu/biodata/	Associations between side effects, chemical, gene, drug, target, and disease	SkipGNN [45]
	HuRI [53]	http://www.interactome-atlas.org/	Protein–protein interaction	SkipGNN [45]
Protein structure databases	Uniprot [54]	https://www.uniprot.org/uploadlists/	Protein sequence and functional information	TriModel [33], GraphCPI [32]
	PDB [55]	https://www.rcsb.org/	3D shapes of proteins, nucleic acids, and complex assemblies	Wen Torng et al. [36]
Benchmark databases	DUD-E [56]	http://dude.docking.org/	Decoys	Wen Torng et al. [36], Jaechang Lim et al. [39]
	MUV [57]	https://www.tu-braunschweig.de/en/pharmchem/forschung/baumann/translate-to-english-muv	Maximum Unbiased Validation (MUV) of virtual screening methods	Wen Torng et al. [36]
	repoDB [58]	http://apps.chiragjpgroup.org/repoDB/	True positives (approved drugs), and true negatives (failed drugs) for drug repositioning	Wang et al. [41]

Figure 2



Structure-based predictions models. **(a)** The pipeline of structure-based model. Different models have similar skeletons; **(b)** Three common input types of structure-based models. In the first, proteins are input as sequence formats and small molecules are constructed as graphs. The second is that protein pockets and small molecules are constructed into different graphs; The third is that protein pockets and small molecules are converted into the same graphs.

DTA prediction, but also creatively uses protein sequences to establish graphs, providing a robust protein descriptor for drug design.

In addition to treating DTI prediction task as a regression task, some other studies modelled DTI prediction as a classification task. DeepCPI [31] can obtain the information of 3D structural interaction sites by representation learning, which is extracted from 2D molecular graphs and 1D protein sequence information. By embedding compounds using r-radius subgraphs, it overcomes the lack of learning parameters and ineffective representation learning due to the insufficient types of atoms and bonds in the molecule. Then, GNNs are used to obtain the low-dimensional representations of the molecular graphs. For protein features extraction, DeepCPI uses CNNs with a filter function. Finally, the attention mechanism is implemented to simulate the interaction and capture the interaction sites between a compound and a protein, rather than simply summing up their embeddings. It assigns a higher weight to a subsequence in a protein if it is important to the compound. GraphCPI [32] is another research that also learns the low-dimensional representations of protein sequences and molecular graphs from CNN and GNN building blocks, respectively. Compared with DeepCPI, it obtains the topological information of the compound from GNNs and uses Prot2vec to encode amino acid sequences into D-dimensional vectors to facilitate protein representation learning. In addition, GraphCPI allows the integration of any popular GNN model, making it more flexible. Since embedding single amino acid is usually meaningless, GraphCPI uses a fixed-length N-gram splitting method to partition the sequences to represent the local chemical context of protein sequences. Although its performance has only moderately improved compared to DeepCPI, GraphCPI is helpful for understanding the compound–protein interaction in combination with the local chemical context and topological structure.

All of these methods consider the entire protein sequence, while others focus on local pockets. Wen Torng and Russ B. Altman [36] developed a two-step graph convolution framework to predict DTIs using 2D complex structures. In their model, features of the protein pocket

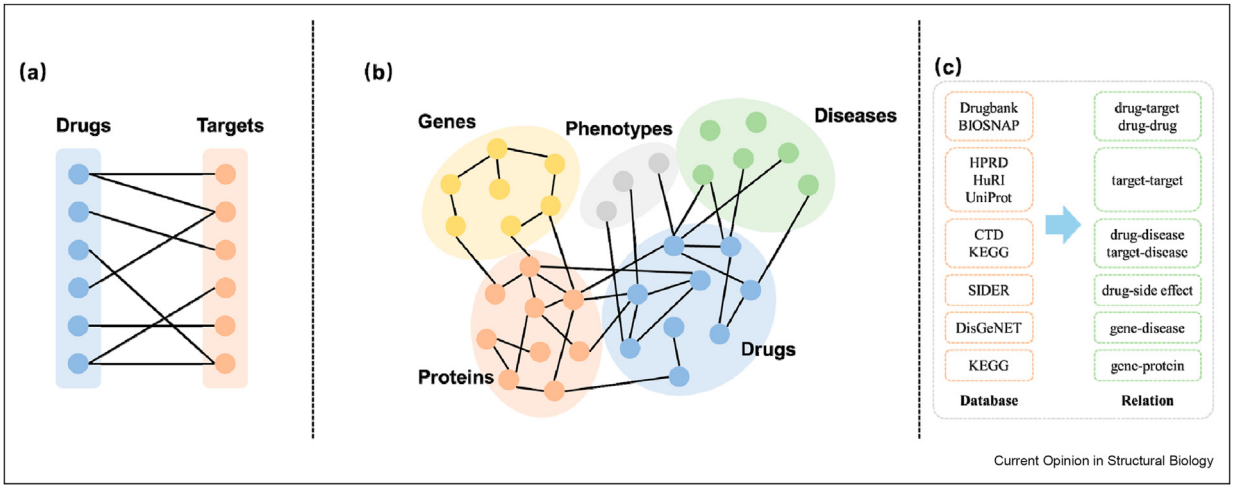
graphs and 2D ligand graphs were extracted by two GCNs, respectively, and then input into fully connected layers to predict DTIs. Remarkably, instead of using the actual chemical bonds as edges, they considered residues less than 7 Å away to be edge-attached, which adds more information about the relative positions of the actual 3D structures. As a result, their model greatly improves the accuracy of DTI prediction for DUD-E datasets. InteractionNet [38] divides the compound and pocket graphs into covalent and non-covalent. Covalent graphs are the actual chemical structures of the compound and protein pocket, while non-covalent maps regard all possible protein-ligand interactions as edges. Compared with the traditional strategy based on covalent bond composition, it provides a novel perspective. Through post-hoc layer-wise relevance propagation (LRP) analysis, InteractionNet successfully captures a series of important non-covalent interactions between proteins and ligands in specific complexes, including hydrogen bonds. Jaechang Lim et al. [39] used GNN to integrate 3D structural information of protein-ligand binding sites directly. They designed a distance—aware graph attention mechanism that enabled the model to distinguish the contribution of each interaction to binding affinity, taking into account both atomic distance and interaction.

Interaction network-based predictions

Interaction networks are ubiquitous in biological systems. In recent years, computational methods based on interaction networks have been applied to solve various biological problems [70]. Moreover, they have enabled us to discover biologically significant but previously unmapped DTIs. Most approaches are based on the assumption of "guilt-by association", that is, similar drugs may have similar targets, and vice versa [71]. With the development of public biomedical datasets on protein—protein interactions, adverse reactions, genome mapping, etc., the advantages of computational methods based on interaction networks have gradually emerged [72]. In addition, some methods that need protein 3D structures use sequences to predict uncharacterized protein crystal structures, which may introduce new biases. Therefore, DTI prediction based on interaction networks is often adopted in the face of emerging diseases, such as the in silico screening of anti-COVID-19 agents. In traditional network analysis, only direct interaction is considered, but the local role (such as neighbors, edge direction) and global position (such as global topology or structure) of the node are ignored [73]. GNNs sample and aggregate features from local neighbors, which can preserve local role and global position information of nodes in the graph [14]. This nature of GNNs benefits the integration of multimodal and complex relationships in biomedical networks.

According to different structures of the interaction network used in DTI prediction, its input can be categorized into bipartite and heterogeneous graphs (Figure 3). The bipartite graph consists of two disjoint and independent sets of nodes [74], such as drugs and targets sets. The edge is only connected between the two sets, i.e., the neighbor of a drug can only be a drug target, and vice versa. There is no edge such as drug—drug or target—target inside these two parts. Hafez Eslami Manoochehri et al.

Figure 3



The construction of interaction network and the comparison of interaction network-based models. **(a)** Bipartite graph; **(b)** Heterogeneous graph; **(c)** The source of the various edges in the interaction network. The left column is the database, and the right column is the kind of edges that can be built. Bipartite graphs require only drug-target interaction edges, while heterogeneous graphs require more kinds of edges.

[75] used an improved Weisfeiler-Lehman Neural Machine (WLNM) to make link predictions for bipartite graphs. Their approach is purely based on network topology information to exclude the "guilt-by association" assumptions. That is, they predicted links based on the structure of the local network, rather than the similarity within the set of drugs or targets. In the follow-up work, Hafez Eslami Manoochehri et al. [76] added drug—drug and protein—protein similarities to the drug-target bipartite graph to construct the semi-bipartite graph model, and made linkage prediction by considering geometric distances in drug target nodes and drug—drug and protein—protein similarities. In the Bifusion model established by Wang et al. [41], disease—drug association information was added. Compared with the information extracted and fused from the PPI network alone, the performance is improved to a certain extent, indicating that the addition of some relevant information in the bipartite graph has an auxiliary effect on the prediction of DTIs.

The heterogeneous graph integrates networks such as drug side effects, drug—disease association data, gene expression data, and protein function data, rather than just drug-target interaction data, which provides diverse information and multi-view perspectives for predicting novel DTIs. The research of Luo et al. [71] showed that the prediction performance of the model with multiple heterogeneous information was better and more robust than that of the simple bipartite graph. However, with the increase in information types, the models suitable for bipartite graphs are difficult to meet the needs of heterogeneous graphs. Hafez Eslami Manoochehri et al. [47] developed an encoder-decoder based GCN method for DTI prediction by constructing a heterogeneous graph consisting of drugs, proteins, diseases, and side effects. Compared with the most advanced heterogeneous graph-based DTI prediction methods at that time, such as NetLapRLS [77], HNM [78], and CMF [79], the performance of this model has been significantly improved, indicating that the principle of GCN method is suitable for this type of task. TriModel [33] proposed a knowledge-graph embedding approach that predicts DTIs in a multi-stage process. Facts in the knowledge-graph were modelled as (subject, predicate, object) (SPO) triples, and the subject entity (drug) is connected to the object entity (target protein) through predicate relationships (drug-target). Tianyi Zhao et al. [80] integrated the association between drug protein pairs (DPPs) into DTI modelling, instead of building a separate drug and protein network. They took DPPs as the nodes of the network and the association between DPPs as the edges of the network, and used GCN to predict DTIs, which also achieved satisfactory results. In addition to adopting different construction methods for heterogeneous graphs, some other studies introduce more advanced machine learning and GNN architectures. SkipGNN [45] receives information from two-hop neighbors as well as adjacent neighbors and predicts molecular interactions not only by aggregating information from direct neighbors, but also from second-order neighbors. Experiments showed that SkipGNN can not only learn biologically significant embedding, but also overcome the disadvantages of high noise and poor integrity in some interaction networks. GANDTI [48] introduced GAN to regularize the feature vectors of nodes into Gaussian distribution, and built a LightGBM classifier, exploiting unknown DTIs to offset the negative effects of class imbalance.

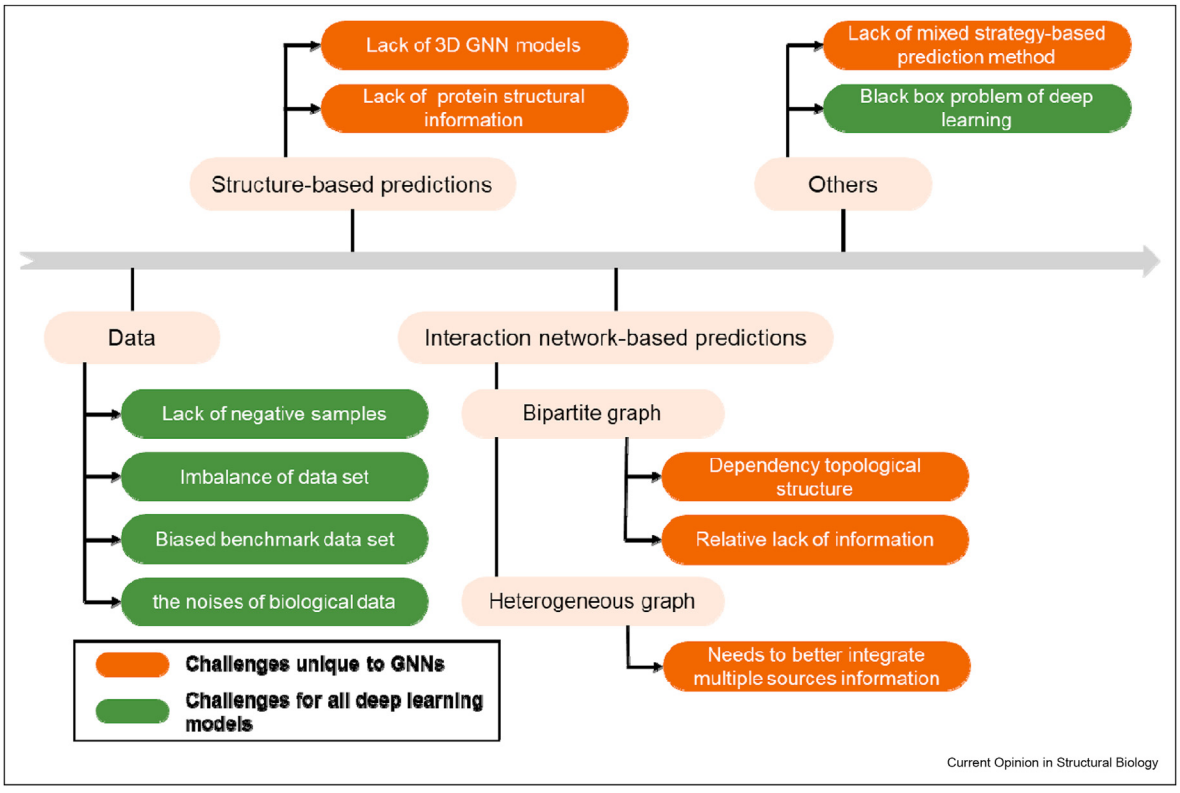
The emergence of SARS-CoV-2 triggered a global pandemic, causing an urgent need to develop effective treatments rapidly. Considering cost, safety, and development speed, drug repurposing is a good way to rapidly screen potential drugs. However, due to the lack of structural data of proteins associated with emerging infectious diseases, most current studies on DTI prediction based on GNNs focus on the interaction network, especially the heterogeneous network. Deisy Morselli Gysi et al. [81] integrated multiple approaches, including GCN, to obtain priority ranking of drug candidates. Drug repurposing knowledge map (DRKG) [82] is a heterogeneous graph composed of genes, compounds, diseases, biological processes, side effects, and symptoms [73,83,84]. built GNN models and used DRKG to rank drug candidates, among which [73] also introduced electronic health records (EHRs) to validate drug effectiveness from large-scale clinical data.

Discussion and outlook

Since GNN was proposed, it has been extensively developed and explored for more than a decade [85], and has been successfully applied in molecular bioinformatics and other fields in recent years. GNNs are applicable to non-Euclidean data that can be graphically represented, so they are of great significance in molecular structures and interaction networks. GNNs can capture the essential structural features of molecules [13,86], which may be one of the reasons why they are more accurate than other deep learning methods in structure-based DTI prediction. Furthermore, combined GCNs and the attention mechanism, SumGNN [87] generates a short reasoning path to provide clues for understanding drug interactions. However, there are still challenges to fully exploring the potential of GNN for DTI prediction (Figure 4).

At present, most GNN approaches for DTI prediction are end-to-end models. Compared with a pipeline of separate components involving feature calculation and feature selection, the end-to-end model requires a large amount of data to understand the complex relationship between the input and the target. Although many large databases have been available, they still cannot meet the needs well in some aspects. First, since most existing databases only contain positive samples, many supervised learning methods simply treat all unlabelled drug-target pairs as

Figure 4



Challenges to fully explore the potential of GNNs for DTI prediction.

negative samples, resulting in inaccurate predictions. Developing semi-supervised learning methods and adding negative samples to databases can solve this problem effectively. Second, the class imbalance between positive and negative data in the training set for the DTI prediction task is also a challenge to deep learning methods, including GNN. In the real scene, the vast majority of compounds are negative samples of a given protein. Thus, random oversampling was frequently used to increase the proportion of positive samples when training the DTI model, which may affect the generalization performance of the model in real-world prediction [88]. Third, the prediction performance of most models for experimental data is significantly lower than that for the DUD-E test set [39]. Therefore, the benchmark dataset may have a biased pattern for classifying active and inactive molecules, which is easily captured by neural networks. There is high demand for building high-quality unbiased benchmark datasets that consist of active and inactive molecules obtained from experiments. Finally, there are a variety of experimental assays and criteria for the determination of bioactivity data. Besides, some active molecules miss quantitative activity data to make quantitative comparisons. As the heterogeneous biological data obtained by different experimental procedures and instruments are noisy and fuzzy [89], it is highly demanding to provide guidelines to find, access, interoperate and reuse DTI data.

The function of macromolecules, especially proteins, is greatly affected by their 3D folding structures. Obviously, the prediction method based on 3D structures provides more direct characterizations of physical interactions between a drug and its protein target. However, the current GNNs mostly operate on flat 2D graphs, ignoring the structural information in 3D space. Recently, some researchers have used 3D protein pockets as input [39], confirming the validity of the GNNs method for 3D molecular structure problems. Such an approach of using local pockets as inputs points the way to structure-based DTI prediction in the future. It captures the 3D structure features of the local reaction pocket rather than the entire protein to reduce the calculation cost. In addition, the structure-based DTI prediction method also has a problem that cannot be ignored, that is, there are still many proteins without structural information. Using homology modelling and some protein structure prediction softwares to obtain the 3D structures of proteins may introduce more bias [68]. Recently, AlphaFold2 [90] has made a breakthrough in protein structure prediction, and structural

biologists are constantly learning more about the structure of proteins. It is believed that the 3D structure of proteins will be easier to obtain in the future, and structure-based DTI prediction methods will also be more accurate than before.

The interaction networks can be divided into bipartite and heterogeneous graphs. The advantage of bipartite graphs is that modelling is simple. As the neighbor nodes of drugs can only be targets, there is no need to integrate multiple-interaction information. However, its disadvantages cannot be ignored. Since many drug targets have not been identified completely, the performance of topological structure-based bipartite graphs is limited because of the missing link. Some studies added other relevant information into the bipartite graph to form a semi-bipartite graph to fill the vacant value to a certain extent. Heterogeneous graphs integrate more information than semi-bipartite graphs. The heterogeneous information is used for DTI prediction, which takes into account the role of entities in different networks, and is more in line with biological and chemical significance. However, there are also some problems that need to be solved. On the one hand, heterogeneous graphs put forward higher requirements for modelling theory. With the continuous improvement and perfection of GNNs theory, researchers are increasingly inclined to use heterogeneous graphs to make DTI predictions. Some studies [33,45] have introduced skip connection and knowledge graph to predict DTIs and achieved good results. On the other hand, the network structure needs to be optimized to integrate information from multiple sources, such as electronic medical records. Knowledge graphs can use such information more effectively. Traditional knowledge graphs need to be populated by experts via manual curation, requiring considerable time and effort [91,92]. Current studies aim to build knowledge graphs based on such data through automated processes [93], but there is still room for improvement. Figuring out how to use interaction network information more effectively is both a challenge and an opportunity. Because the structural and pathological mechanisms of related proteins are not fully understood, many drug repurposing studies for COVID-19 are based on interaction network [73,81,83], demonstrating the necessity of developing prediction methods based on interaction networks.

At present, there is still a lack of DTI prediction methods based on mixed strategies. Perhaps combining structure with interaction network can take into account both chemical action and biological network location, and more comprehensive information can contribute to better prediction. In addition, "black box" deep neural networks are often criticized for their lack of interpretability, which is necessary in the field of biomedicine. Interpretable models not only enable prediction of DTIs, but also help us understand the underlying mechanisms better and facilitate the discovery of new active compounds and new targets.

Conclusions

Over the past few years, GNNs have become a powerful and useful tool for DTI prediction tasks. This progress is attributed to the applicability of GNNs to non-Euclidean data and the explosive growth of GNNs theoretical research in recent years. In this review, we provide a comprehensive review of the recent applications of GNNs in the field of DTI prediction. According to the difference of input, we categorize the DTI prediction tasks into structure-based and interaction network-based and introduce several representative research of them. In addition, we summarize the databases involved in the relevant papers and classify them according to different content. Finally, we propose the current problems that need to be solved in GNN approaches for DTI prediction task, and make a preliminary discussion on the future research direction of some problems. It is foreseeable that GNNs will be used widely in the field of drug research in the future, which could significantly shorten the cycle of drug research and development.

Conflict of interest statement

Nothing declared.

Acknowledgments

This work was supported by the Lingang Laboratory (LG202102-01-02), the National Natural Science Foundation of China (81903639), Shanghai Municipal Science and Technology Major Project, and Shanghai Sailing Program (19YF1457800).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.sbi.2021.102327.

References

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Landry Y, Gies JP: **Drugs and their molecular targets: an updated overview**. *Fund Clin Pharmacol* 2008, **22**:1–18, https://doi.org/10.1111/j.1472-8206.2007.00548.x.

2. Masoudi-Nejad A, Mousavian Z, Bozorgmehr JH: **Drug-target and disease networks: polypharmacology in the post-genomic era**. *In Silico Pharmacol* 2013, **1**:17, https://doi.org/10.1186/2193-9616-1-17.

3. Whitebread S, Hamon J, Bojanic D, Urban L: **Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development**. *Drug Discov Today* 2005, **10**:1421–1433, https://doi.org/10.1016/S1359-6446(05)03632-9.

4. Sliwoski G, Kothiwale S, Meiler J, Lowe EW: **Computational methods in drug discovery**. *Pharmacol Rev* 2014, **66**:334–395, https://doi.org/10.1124/pr.112.007336.

5. • D'souza S, Prema K, Balaji S: **Machine learning models for drug–target interactions: current knowledge and future directions**. *Drug Discov Today* 2020, **25**:748–756, https://doi.org/10.1016/j.drudis.2020.03.003.

This work provides an extensive review of machine learning and deep learning models for virtual screening and DTI prediction.

6. BI I, David W, TI V: **A renaissance of neural networks in drug discovery**. *Expet Opin Drug Discov* 2016, **11**:785–795, https://doi.org/10.1080/17460441.2016.1201262.

7. Velazquez M, Anantharaman R, Velazquez S, Lee Y: **RNN-based Alzheimer's disease prediction from prodromal stage using diffusion tensor imaging**. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; 2019: 1665–1672, https://doi.org/10.1109/BIBM47256.2019.8983391.

8. Lee I, Keum J, Nam H: **DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences**. *PLoS Comput Biol* 2019, **15**, e1007129, https://doi.org/10.1371/journal.pcbi.1007129.

9. Chi C, Weike Y, Yunxing Z, Chen Z, Ping OS: **Graph networks as a universal machine learning framework for molecules and crystals**. *Chem Mater* 2019, **31**:3564–3572, https://doi.org/10.1021/acs.chemmater.9b01294.

10. Gil A, Arye S, Einat S, Maxim S, Dvir N, Ilya V, Shmuel P: **Network analysis of protein structures identifies functional residues**. *J Mol Biol* 2004, **344**:1135–1146, https://doi.org/10.1016/j.jmb.2004.10.055.

11. Fout A, Byrd J, Shariat B, Ben-Hur A: **Protein interface prediction using graph convolutional networks**. *Adv Neural Inf Process Syst* 2017:6530–6539.

12. Hamaguchi T, Oiwa H, Shimbo M, Matsumoto Y: **Knowledge transfer for out-of-knowledge-base entities: a graph neural network approach**. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*; 2017: 1802–1808, https://doi.org/10.24963/ijcai.2017/250.

13. •• Zhou J, Cui G, Zhang Z, Yang C, Liu Z, Wang L, Li C, Sun M: **Graph neural networks: a review of methods and applications**. *AI Open* 2020, **1**:57–81, https://doi.org/10.1016/j.aiopen.2021.01.001.
In this paper, the existing GNN models are reviewed in detail, and their applications are systematically classified.

14. Hamilton WL, Ying R, Leskovec J: **Inductive representation learning on large graphs**. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017.

15. Battaglia PW, Pascanu R, Lai M: **Interaction networks for learning about objects, relations and physics**. *arXiv preprint: .00222* 2016. https://arxiv.org/abs/1612.00222; 2016.

16. Rhee S, Seo S, Kim S: **Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification**. *arXiv preprint :.05859* 2017. https://arxiv.org/abs/1711.05859; 2017.

17. Öztürk H, Özgür A, Ozkirimli E: **DeepDTA: deep drug–target binding affinity prediction**. *Bioinformatics* 2018, **34**:i821–i829, https://doi.org/10.1093/bioinformatics/bty593.

18. Bronstein MM, Bruna J, Lecun Y, Szlam A, Vandergheynst P: **Geometric deep learning: going beyond Euclidean data**. In *IEEE Signal Processing Magazine*; 2017:18–42, https://doi.org/10.1109/MSP.2017.2693418.

19. Lecun Y, Bottou L, Bengio Y, Haffner P: **Gradient-based learning applied to document recognition**. In *Proceedings of the IEEE*, vol. 86; 1998:2278–2324, https://doi.org/10.1109/5.726791.

20. Tsoi AC, Back A: **Discrete time recurrent neural network architectures: a unifying review**. *Neurocomputing* 1997, **15**: 183–223, https://doi.org/10.1016/S0925-2312(97)00161-6.

21. Perozzi B, Al-Rfou R, Skiena S: **Deepwalk: online learning of social representations**. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*; 2014:701–710.

22. Goldberg Y: **Levy OJaPA: word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method**. *arXiv preprint* 2014. https://arxiv.org/abs/1402.3722; 2014.

23. •• Lecun Y, Bengio Y, Hinton G: **Deep learning**. *Nature* 2015, **521**: 436–444, https://doi.org/10.1038/nature14539.
This article introduce neural networks, and points out four key ideas for CNN's success.

24. Bruna J, Zaremba W, Szlam A, Lecun Y: **Spectral networks and locally connected networks on graphs**. *arXiv preprint* 2013 https://arxiv.org/abs/1312.6203.

25. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y: **Graph attention networks**. *arXiv preprint* 2017:10903. https://arxiv.org/abs/1710.10903.

26. Weininger D: **SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules**. *J Chem Inf Comput Sci* 1988, **28**:31–36, https://doi.org/10.1021/ci00057a005.

27. • Nguyen T, Le H, Venkatesh S: **GraphDTA: prediction of drug–target binding affinity using graph convolutional networks**. *bioRxiv* 2019:684662.
On the basis of DeepDTA, this work changed part of CNN layer to GNN layer. Without changing the structure of other parts of the model, for the same data set, the model performance has been improved to some extent.

28. Mayr LM, Bojanic D: **Novel trends in high-throughput screening**. *Curr Opin Pharmacol* 2009, **9**:580–588, https://doi.org/10.1016/j.coph.2009.08.004.

29. Borman S: **Reducing time to drug discovery**. *Chem Eng News* 1999, **77**:33–34. 36,38,40,42-43,46,48, https://doi.org/10.1021/cen-v077n010.p033.

30. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets**. *Nucleic Acids Res* 2008, **36**:D901–D906, https://doi.org/10.1093/nar/gkm958.

31. Tsubaki M, Tomii K, Sese J: **Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences**. *Bioinformatics* 2019, **35**:309–318, https://doi.org/10.1093/bioinformatics/bty535.

32. Quan Z, Guo Y, Lin X, Wang Z-J, Zeng X: **Graphcpi: graph neural representation learning for compound-protein interaction**. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE; 2019:717–722, https://doi.org/10.1109/BIBM47256.2019.8983267.

33. •• Mohamed SK, Nováček V, Nounu A: **Discovering protein drug targets using knowledge graph embeddings**. *Bioinformatics* 2020, **36**:603–610, https://doi.org/10.1093/bioinformatics/btz600.
By formulating the DTI prediction as link prediction problem in knowledge graphs, this paper proposes a specific knowledge graph embedding model that is superior to all previous methods in terms of both area under ROC and precision-recalll curves.

34. Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiess A, Jensen LJ: **SuperTarget and Matador: resources for exploring drug-target relationships**. *Nucleic Acids Res* 2007, **36**:D919–D922, https://doi.org/10.1093/nar/gkm862.

35. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, Mcglinchey S, Michalovich D, Al-Lazikani B: **ChEMBL: a large-scale bioactivity database for drug discovery**. *Nucleic Acids Res* 2012, **40**:D1100–D1107, https://doi.org/10.1093/nar/gkr777.

36. Torng W, Altman RB: **Graph convolutional neural networks for predicting drug-target interactions**. *J Chem Inf Model* 2019, **59**: 4131–4149, https://doi.org/10.1021/acs.jcim.9b00628.

37. Wang R, Fang X, Lu Y, Yang C-Y, Wang S: **The PDBbind database: methodologies and updates**. *J Med Chem* 2005, **48**: 4111–4119, https://doi.org/10.1021/jm048957q.

38. Hyeoncheol C, Kyun LE, CI S: **InteractionNet: modeling and explaining of noncovalent protein-ligand interactions with noncovalent graph neural network and layer-wise relevance propagation**. *arXiv preprint :.13438* 2020. https://arxiv.org/abs/2005.13438.

39. •• Lim J, Ryu S, Park K, Choe YJ, Ham J, Kim WY: **Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation**. *J Chem Inf Model* 2019, **59**:3981–3988, https://doi.org/10.1021/acs.jcim.9b00387.
With introducing a distance-aware graph attention algorithm, the graphic features of intermolecular interactions can be extracted directly from the 3D structure of protein ligand binding pose.

40. Sorin A, BC G, Jayme H, Giovanni B, WT B, Dac-Trung N, Ramona C, Liliana H, Alina B, YJ J: **DrugCentral 2021 supports drug discovery and repositioning**. *Nucleic Acids Res* 2021, **49**: D1160–D1169, https://doi.org/10.1093/nar/gkaa997.

41. Wang Z, Zhou M, Arnold C: **Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing**. *Bioinformatics* 2020, **36**:i525–i533, https://doi.org/10.1093/bioinformatics/btaa437.

42. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B: **PubChem 2019 update: improved access to chemical data**. *Nucleic Acids Res* 2019, **47**:D1102–D1109, https://doi.org/10.1093/nar/gky1033.

43. CK C, WA H, Yang-Yang F, Susanna K, CA C, Gregory S, Alex W, SN C, GO L, Malachi G: **DGIdb 3.0: a redesign and expansion of the drug–gene interaction database**. *Nucleic Acids Res* 2018, **46**:D1068–D1073, https://doi.org/10.1093/nar/gkx1143.

44. Piñero J, À Bravo, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI: **DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants**. *Nucleic Acids Res* 2016:D833–D839, https://doi.org/10.1093/nar/gkw943.

45. • Huang K, Xiao C, Glass L, Zitnik M, Sun J: **SkipGNN: predicting molecular interactions with skip-graph networks**. *Sci Rep* 2020, **10**:1–16, https://doi.org/10.1038/s41598-020-77766-9.
In biological networks, similarity between nodes that do not interact directly is very useful. Unlike existing GNN model, SkipGNN uses this skipping similarity to aggregate second-order interaction information to predict interactions between molecules.

46. Keshava Prasad TT, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A: **Human protein reference database—2009 update**. *Nucleic Acids Res* 2009, **37**:D767–D772, https://doi.org/10.1093/nar/gkn892.

47. Manoochehri HE, Pillai A, Nourani M: **Graph convolutional networks for predicting drug-protein interactions**. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE; 2019:1223–1225, https://doi.org/10.1109/BIBM47256.2019.8983018.

48. Sun C, Xuan P, Zhang T, Ye Y: **Graph convolutional autoencoder and generative adversarial network-based method for predicting drug-target interactions**. *IEEE ACM Trans Comput Biol Bioinf* 2020, https://doi.org/10.1109/TCBB.2020.2999084.

49. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Mcmorran R, Wiegers J, Wiegers TC, Mattingly CJ: **The comparative toxicogenomics database: update 2019**. *Nucleic Acids Res* 2019, **47**: D948–D954, https://doi.org/10.1093/nar/gky868.

50. Kuhn M, Letunic I, Jensen LJ, Bork P: **The SIDER database of drugs and side effects**. *Nucleic Acids Res* 2016, **44**: D1075–D1079, https://doi.org/10.1093/nar/gkv1075.

51. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T: **KEGG for linking genomes to life and the environment**. *Nucleic Acids Res* 2007, **36**:D480–D484, https://doi.org/10.1093/nar/gkm882.

52. Zitnik M, SosiC R, Maheshwari S, Leskovec J: ***BioSNAP Datasets: stanford biomedical network dataset collection***. http://snap.stanford.edu/biodata/.

53. Vidal M: **How much of the human protein interactome remains to be mapped?** *Sci Signal* 2016, **9**:eg7, https://doi.org/10.1126/scisignal.aaf6030.

54. Consortium U: **UniProt: a worldwide hub of protein knowledge**. *Nucleic Acids Res* 2019, **47**:D506–D515, https://doi.org/10.1093/nar/gky1049.

55. BH M, John W, Zukang F, Gary G, BT N, Helge W, SI N, BP E: **The protein data bank**. *Nucleic Acids Res* 2000, **28**:235–242, https://doi.org/10.1093/nar/28.1.235.

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

57. Rohrer SG, Baumann K: **Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data**. *J Chem Inf Model* 2009, **49**:169–184, https://doi.org/10.1021/ci8002649.

58. BA S, PC J: **A standard database for drug repositioning**. *Sci Data* 2017, **4**:1–7, https://doi.org/10.1038/sdata.2017.29.

59. Sun M, Zhao S, Gilvary C, Elemento O, Zhou J, Wang F: **Graph convolutional networks for computational drug development and discovery**. *Briefings Bioinf* 2020, **21**:919–935, https://doi.org/10.1092/bib/bbz042.

60. • Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J, Zhang Y: **Drug–target interaction prediction: databases, web servers and computational models**. *Briefings Bioinf* 2016, **17**:696–712, https://doi.org/10.1092/bib/bbv066.
In this review, authors detail the databases commonly used in DTI prediction. In addition, they mainly introduced some state-of-the-art computational models for DTI prediction, including network-based method, machine learning-based method and so on.

61. Wang L, Deng Y, Wu Y, Kim B, Lebard DN, Wandschneider D, Beachy M, Friesner RA, Abel R: **Accurate modeling of scaffold hopping transformations in drug discovery**. *J Chem Theor Comput* 2017, **13**:42–54, https://doi.org/10.1021/acs.jctc.6b00991.

62. Feng Q, Dueva E, Cherkasov A, Ester M, Padme: **A deep learning-based framework for drug-target interaction prediction**. *arXiv preprint :.09741* 2018. https://arxiv.org/abs/1807.09741; 2018.

63. Landrum G: *RDKit: a software suite for cheminformatics, computational chemistry, and predictive modeling*. Academic Press; 2013.

64. Ramsundar B: ***deepchem.io***. https://github.com/deepchem/deepchem.

65. Xu K, Hu W, Leskovec J, Jegelka S: **How powerful are graph neural networks?** *arXiv preprint :.00826* 2018 https://arxiv.org/abs/1810.00826.

66. Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, Hocker M, Treiber DK, Zarrinkar PP: **Comprehensive analysis of kinase inhibitor selectivity**. *Nat Biotechnol* 2011, **29**: 1046–1051, https://doi.org/10.1038/nbt.1990.

67. Tang J, Szwajda A, Shakyawar S, Tao X, Aittokallio T: **Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis**. *J Chem Inf Model* 2014, **54**:735–743, https://doi.org/10.1021/ci400709d.

68. • Jiang M, Li Z, Zhang S, Wang S, Wang X, Yuan Q, Wei Z: **Drug–target affinity prediction using graph neural network and contact maps**. *RSC Adv* 2020, **10**:20701–20712, https://doi.org/10.1039/D0RA02297G.
In order to improve the prediction accuracy of DTA, the drug molecule map and protein map were established respectively, and the graph neural network was introduced to obtain their representation, which greatly improved the model performance.

69. Michel M, Menéndez Hurtado D, Elofsson A: **PconsC4: fast, accurate and hassle-free contact predictions**. *Bioinformatics* 2019, **35**:2677–2679, https://doi.org/10.1093/bioinformatics/bty1036.

70. Lenore C, Trey I, RB J, Roded S: **Network propagation: a universal amplifier of genetic associations**. *Nat Rev Genet* 2017, **18**:551, https://doi.org/10.1038/nrg.2017.38.

71. Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, Peng J, Chen L, Zeng J: **A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information**. *Nat Commun* 2017, **8**:1–13, https://doi.org/10.1038/s41467-017-00680-8.

72. Vilar S, Hripcsak G: **The role of drug profiles as similarity metrics: applications to repurposing, adverse effects**

detection and drug–drug interactions. *Briefings Bioinf* 2016, **18**:670–681, https://doi.org/10.1092/bib/bbw048.

73. Hsieh K, Wang Y, Chen L, Zhao Z, Savitz S, Jiang X, Tang J, Kim Y: **Drug repurposing for COVID-19 using graph neural network with genetic, mechanistic, and epidemiological validation**. *arXiv preprint :.10931* 2020. https://arxiv.org/abs/2009.10931; 2020.

74. Pavlopoulos GA, Kontou PI, Pavlopoulou A, Bouyioukos C, Markou E, Bagos PG: **Bipartite graphs in systems biology and medicine: a survey of methods and applications**. *GigaScience* 2018, **7**:giy014, https://doi.org/10.1093/gigascience/giy014.

75. Manoochehri HE, Kadiyala SS, Nourani M: **Predicting drug-target interactions using weisfeiler-lehman neural network**. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE; 2019:1–4, https://doi.org/10.1109/BHI.2019.8834572.

76. Manoochehri HE, Nourani M: **Drug-target interaction prediction using semi-bipartite graph model and deep learning**. *BMC Bioinf* 2020, **21**:1–16, https://doi.org/10.1186/s12859-020-3518-6.

77. Xia Z, Wu L-Y, Zhou X, Wong ST: **Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces**. BioMed Central *BMC Syst Biol* 2010:1–16, https://doi.org/10.1186/1752-0509-4-S2-S6.

78. Monica C, Michael K, Anne-Claude G, Juhl JL, Peer B: **Drug target identification using side-effect similarity**. *Science* 2008, **321**:263–266, https://doi.org/10.1126/science.1158140.

79. Zheng X, Ding H, Mamitsuka H, Zhu S: **Collaborative matrix factorization with multiple similarities for predicting drug-target interactions**. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*; 2013:1025–1033, https://doi.org/10.1145/2487575.2487670.

80. Zhao T, Hu Y, Valsdottir LR, Zang T, Peng J: **Identifying drug–target interactions based on graph convolutional network and deep neural network**. *Briefings Bioinf* 2020, https://doi.org/10.1092/bib/bbaa044.

81. Gysi DM, Do Valle Í, Zitnik M, Ameli A, Gan X, Varol O, Sanchez H, Baron RM, Ghiassian D, Loscalzo J: **Network medicine framework for identifying drug repurposing opportunities for COVID-19**. *Proc Natl Acad Sci Unit States Am* 2021, **118**, e2025581118, https://doi.org/10.1073/pnas.2025581118.

82. Song X, Ioannidis VN, Li M, Zheng D: ***Drug repurposing knowledge graph (DRKG)***. https://github.com/gnn4dr/DRKG.

83. Ioannidis VN, Zheng D, Karypis G: **Few-shot link prediction via graph neural networks for Covid-19 drug-repurposing**. *arXiv preprint :.10261* 2020. https://arxiv.org/abs/2007.10261; 2020.

84. Siddhant D, Prabhakar CS: **Dr-COVID: graph neural networks for SARS-CoV-2 drug repurposing**. *arXiv preprint :.02151* 2020. https://arxiv.org/abs/2012.02151; 2020.

85. Marco G, Gabriele M, Franco S: **A new model for learning in graph domains**. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* IEEE; 2005:729–734, https://doi.org/10.1109/IJCNN.2005.1555942.

86. Nyamabo AK, Yu H, Shi J-Y: **SSI–DDI: substructure–substructure interactions for drug–drug interaction prediction**. *Briefings Bioinf* 2021:1–10, https://doi.org/10.1092/bib/bbab133. 00.

87. Yu Y, Huang K, Zhang C, Glass LM, Sun J, Xiao C: **Sumgnn: multi-typed drug interaction prediction via efficient knowledge graph summarization**. *Bioinformatics* 2021, btab207, https://doi.org/10.1093/bioinformatics/btab207.

88. Kotsiantis S, Kanellopoulos D, Pintelas P: **Handling imbalanced datasets: a review**. *GESTS Int Transact Comp Sci Engin* 2006, **30**:25–36.

89. Zhou W, Koudijs KK, Böhringer S: **Influence of batch effect correction methods on drug induced differential gene expression profiles**. *BMC Bioinf* 2019, **20**:1–14, https://doi.org/10.1186/s12859-019-3028-6.

90. •• Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021:1–11, https://doi.org/10.1038/s41586-021-03819-2.
This is the most accurate protein structure prediction method to date.

91. Shwe MA, Middleton B, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP, Cooper GF: **Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base**. *Methods Inf Med* 1991, **30**:241–255, https://doi.org/10.1055/s-0038-1634846.

92. Nicholson DN, Greene CS: **Constructing knowledge graphs and their biomedical applications**. *Comput Struct Biotechnol J* 2020, **18**:1414–1428, https://doi.org/10.1016/j.csbj.2020.05.017.

93. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D: **Learning a health knowledge graph from electronic medical records**. *Sci Rep* 2017, **7**:1–11, https://doi.org/10.1038/s41598-017-05778-z.