Computers in Biology and Medicine 163 (2023) 107136

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A survey of drug-target interaction and affinity prediction methods via graph neural networks

Yue Zhang [a,*], Yuqing Hu [a], Na Han [a], Aqing Yang [a], Xiaoyong Liu [a], Hongmin Cai [b]

[a] School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, 510665, China
[b] School of Computer Science and Engineering, South China University of Technology, Guangzhou, 510006, China

ARTICLE INFO

Index Terms:
Drug–target interaction
Drug–target affinity
Graph neural networks
Deep learning

ABSTRACT

The tasks of drug-target interaction (DTI) and drug-target affinity (DTA) prediction play important roles in the field of drug discovery. However, biological experiment-based methods are time-consuming and expensive. Recently, computational-based approaches have accelerated the process of drug-target relationship prediction. Drug and target features are represented in structure-based, sequence-based, and graph-based ways. Although some achievements have been made regarding structure-based representations and sequence-based representations, the acquired feature information is not sufficiently rich. Molecular graph-based representations are some of the more popular approaches, and they have also generated a great deal of interest. In this article, we provide an overview of the DTI prediction and DTA prediction tasks based on graph neural networks (GNNs). We briefly discuss the molecular graphs of drugs, the primary sequences of target proteins, and the graph reSLBpresentations of target proteins. Meanwhile, we conducted experiments on various fundamental datasets to substantiate the plausibility of DTI and DTA utilizing graph neural networks.

1. Introduction

The approval or disapproval of a new drug by the U.S. Food and Drug Administration (FDA) takes 10–17 years and costs approximately $2.6 billion [1–4]. Hence, the development of drugs is very challenging, with significant losses in case of failures [5]. As regulatory oversight of marketed drugs becomes increasingly stringent, the number of new drugs entering the market continues to decline. Therefore, an efficient screening technology has emerged—drug repurposing and repositioning. Drug repurposing and repositioning can identify new uses for approved drugs, speed up the drug development process, and reduce the financial and time costs of development [6]. Thus, how marketed drugs interact with novel targets and how drug-target interactions (DTIs) can be predicted with high precision are topics that have attracted widespread attention.

Although methods based on biological trials (e.g., high-throughput screening experiments) can be effective at predicting DTIs, they are too cumbersome and demand substantial amounts of money and time [7]. At the same time, the presence of millions of drug-like compounds [8] and hundreds of potential target make this technology unreachable. Bioinformatics databases have accumulated large amounts of biological experimental data in recent years. The emergence of computational methods is a result of this volume of data.

Ligands are substances that bind to proteins. Conventional computational methods are broadly divided into two categories: ligand-based methods and structure-based methods. Ligand-based methods [9] assume that ligands with similar chemical characteristics also possess similar biological activities and can bind to similar target proteins [10–13]. This method depends on the information of the ligand of interest rather than the structure and data of the target protein. However, the accuracy of the predicted target decreases when the number of ligands is too insufficient. One of the most typical examples of a structure-based approach is docking simulation [14]. The three-dimensional structures of drug molecules and target proteins are used to predict the binding strengths of drugs and target proteins via molecular docking and molecular dynamics simulations [15–18]. This enables accurate prediction, but it is very time-consuming in practice.

Sequence-based methods are hot research topics in computational biology. With the development of deep learning, many earlier studies employed neural network frameworks to extract the features of protein sequences and drug molecules. Compared with traditional machine learning methods, deep learning can better learn the potential

* Corresponding author. School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China.
E-mail address: zhangyue@gpnu.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.compbiomed.2023.107136
Received 11 March 2023; Received in revised form 29 May 2023; Accepted 4 June 2023
Available online 7 June 2023

characteristics of target proteins and drugs. Sequence-based methods divide the relationships between drugs and targets into two categories: DTIs and drug-target affinities (DTAs). DTI prediction is a binary classification task that determines whether a drug and a target interact, and it is also a simplification of DTA prediction. For instance, transformer technology is used by TransformerCPI [19] to mine drug and protein sequence data for DTI prediction. NeoDTI [20] leverages data from heterogeneous networks to preserve the topologies of drugs and proteins for DTI prediction. [21] employed convolutional neural network (CNN) technology to obtain the local features of drug and protein sequences for DTI prediction. DTA prediction is a regression task, unlike DTI prediction, and its prediction process yields continuous values. DeepDTA [22] adopts two CNNs to learn the features of drug molecules and protein sequences for DTA prediction. Two CNNs are added to separately learn the features of protein domains and motifs (PDMs) and ligand maximum common substructures (LMCSs). WideDTA [23] is an improved version of DeepDTA. To capture the latent features of drug and protein sequences for affinity prediction, Co-VAE [24] employs variational autoencoding technology. While sequence-based methods have achieved some success, they ignore the topologies of drugs and protein molecules.

The use of CNN technology alone is not sufficient for capturing the sequence information of drugs and proteins. It is important to consider the relationships between each atom and automatically assign different weights and attention to them. Attention mechanism has emerged as a promising approach to address this challenge. In recent years, numerous DTI prediction methods based on attention mechanism have been proposed and achieved remarkable results. For instance, AttentionSiteDTI [25] constructs a graph using structural information between drugs and targets, combines NLP techniques to process relational information from textual data, and applies attention mechanism to weight the importance of different nodes and edges in the graph. DrugBAN [26] introduces a model based on bilinear attention networks to capture the interactions between features and weight the importance of different features using attention mechanism. Additionally, the model incorporates domain adaptation techniques to enhance prediction performance by adapting to different domain data. HyperAttentionDTI [27] utilizes CNN and attention mechanism to learn features of drugs and targets, and assigns attention vectors to each atom and amino acid for weighted representation. FusionDTA [28] employs an attention-based feature fusion module to combine the feature representations of drugs and targets, generating more informative feature representations. Furthermore, it leverages knowledge distillation to combine the predictions from multiple models, further improving the accuracy and robustness of the predictions.

Matrix factorization-based methods are crucial skills in computational DTI prediction models. Matrix factorization and matrix completion techniques have been widely applied in drug repositioning and target prediction tasks. They map the drug-target association matrix into a low-dimensional representation, thereby extracting low-dimensional features. The specific approach involves decomposing the interaction matrix into two low-dimensional feature matrices for drugs and targets through matrix factorization in DTI prediction tasks. This method has attracted considerable attention. For instance, DTINet [29] constructs a heterogeneous network using various drug-related information to learn low-dimensional vector representations of features, enabling the prediction of DTI. NRLMF [30] introduces regularization to incorporate neighborhood information, encouraging similar drugs and targets to remain close in the feature space. It models the probabilities of DTI through matrix factorization to achieve accurate predictions. NRLMFβ [31] addresses the issue of inaccurate predictions when DTI have limited sample information, building upon NRLMF. KBMF2K [32] converts drug and target information into kernel matrices and employs Bayesian matrix factorization to decompose the kernel matrix into two low-dimensional feature matrices, obtaining latent representations of drugs and targets. SPLCMF [33] utilizes a self-paced learning strategy to progressively select training samples and control the feature learning process, thereby improving the model's generalization ability and prediction accuracy. It predicts drug-target interactions using collaborative matrix factorization techniques.

These matrix factorization-based methods have demonstrated promising performance in DTI prediction, offering valuable insights into drug development and target discovery. By incorporating various data sources and leveraging advanced techniques, these methods contribute to the advancement of computational approaches in the field of DTI prediction.

Neural network learning has been widely applied in computer vision, recommendation systems, social network analysis, and other fields. For example, MCSSC [34] addresses the heterogeneity and structural constraints of different views by constructing a network for each view, which transforms the multi-view clustering problem into a multi-layer network clustering problem, improving the experimental accuracy. Network learning has also been applied in the field of bioinformatics. SLNMF [35] uses a network-based structure learning non-negative matrix factorization to infer cell types and trajectories by utilizing the interactions between cells. It establishes a similarity network based on the relationships between cells and then extracts the latent features of cells using the topological structure of the cell network. TANMF [36] integrates heterogeneous genomic data in the dynamic modules of temporal cancer networks by adding gene attributes to the temporal network. It not only utilizes the topological structure of the network but also integrates multimodal data. In the multi-layer network of LSNMF [37], each layer represents a data source, such as gene expression and DNA methylation. It combines network representation learning and non-negative matrix factorization technology to detect specific layer modules in cancer multi-layer networks. NMF-DEC [38] integrates interactome and transcriptome data to construct a similarity network for gene attributes and studies the relationship between gene attributes and network topology by decomposing the similarity network and interaction network.

Recently, graph neural network (GNN) [39] has shown some success in the field of bioinformatics, offering the unprecedented capability to process non-Euclidean spatial data with remarkable efficiency. By leveraging the intrinsic information embedded in the nodes and their interdependencies, these networks can expertly extract valuable feature information, enabling accurate identification and prediction of vertices or edges. Meanwhile, due to the strong correlation and diversity of biological information data, molecular graphs are very suitable for the representation of biological information, especially the molecular structure and functional relationship between molecules.GNN models are the most effective approaches for identifying molecular features. A GNN learns the network's local and global information by aggregating neighbor information. We will summarize the drug molecule and protein graphs from three different aspects.

1. The graph is the most effective way to describe drug molecule, and we present a summary of drug-target interaction and affinity prediction via graph neural networks.
2. We leverage protein graph, drug molecule graph, and primary protein sequence for description depending on how molecular features are represented.
3. We categorize the graph neural network-based drug-target interaction and affinity prediction into the following three categories: interaction prediction based on molecular drug graph and protein sequence, affinity prediction based on molecular drug graph and protein sequence, and interaction and affinity prediction based on molecular drug graph and protein graph.

2. Representations of drugs and target proteins

2.1. Graph representations of drugs

The drug sequence can represent the information of the drug. The simplified molecular-input line entry system (SMILES) [40] is a common method for drugs that uses one-dimensional ASCII strings to describe the three-dimensional chemical structure of drugs. Recently, many approaches have been derived to apply non-Euclidean-structured data in GNNs [41–45]. Examples include graph convolutional networks (GCNs), graph attention networks (GATs), graph autoencoders (GAEs), graph generative networks, and graph spatial-temporal networks, etc. In bioinformatics, data analysis revealed that the molecular graphs of drug molecules fit very well into GNNs. By using RDKit [46] technology (an open-source chemistry software), the SMILES of a drug can be converted into a molecular graph $G = (V,E)$, where V represents the set of atoms in the drug, i.e., the nodes of the graph and E represents the chemical bonds between atoms, i.e., the graph's edges, in the form of an adjacency matrix. The drug SMILES is converted into a molecular graph, as shown in Fig. 1.

After obtaining the relationship graph of each atom in the drug molecule, it is necessary to assign features to each atom node for further processing. Each node in the graph can be described using a feature vector consisting of five features: atomic symbol, degree of the atom (i.e., the sum of the number of neighboring atoms and adjacent hydrogen atoms), total number of hydrogen atoms, number of implicit hydrogen atoms of the atom, and whether the atom is aromatic. These five features together form a 78-dimensional vector, with each feature occupying a specific number of dimensions as indicated in Table 1. Except for the atomic symbol of the node, the remaining four features are commonly used chemical features of the atom and can be easily obtained using RDKit. The detailed process of obtaining the atom's one-hot feature vector is shown in Table 2, which presents pseudocode.

2.2. Primary sequence representations of proteins

Proteins are composed of amino acids, and the three-dimensional and high-dimensional structures of proteins are not easily accessible and represented. The most common approach in the literature involves the use of the primary protein structure: amino acid sequences. Amino acid sequences exist in text form, which is somewhat similar to the task of natural language processing (NLP) [47]. CNNs [48], long short-term memory (LSTM) [49], gated recurrent units (GRUs) [50], Word2Vec [51,52] and bidirectional encoder representations from transformers (BERT) [53] are popular methods for extracting text characteristics and are suitable for acquiring protein sequence information. The sequence representation of a protein is

shown in Fig. 2.

2.3. Graph representations of proteins

The most popular molecular representations for proteins are primary sequences, but the prediction accuracy would be improved if a protein graph could be employed. The use of a whole protein to form a graph results in a complex graph structure and a lengthy training time, as

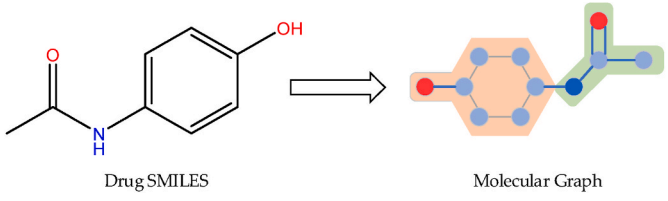

Fig. 1. The SMILES of a drug is converted to the molecular graph(The sequence is CC(=O)NC1 = = CC]C(C]C1)O).

Table 1
The description of atomic features in the drug molecule graph.

Atom Feature	Dimension
Atomic type	44
Degree of the atom	11
Total number of hydrogen atoms	11
Number of implicit hydrogen atoms of the atom	11
Whether the atom is aromatic	1
All	78

proteins contain large numbers of atoms. A contact map, which reflects the high-dimensional structure of a protein, can be utilized to reveal the interactions between protein residues. A protein has only a few hundred residues, which is a very suitable situation for constructing a protein graph. The flow of the protein graph construction process is shown in Fig. 3. Moreover, we also introduced the node features in the protein graph.

Protein structure information encompasses the angles and distances between different residue pairs, and contact maps are a result of structure prediction. Generally, if the distance between two C_b atoms of residues in Euclidean space is less than a certain threshold [54], they are considered in contact with each other. One can use Pconsc4, an open-source and efficient method, to predict contact maps. Additionally, evolutionary scale modeling (ESM) models can also be used to predict contact maps without requiring protein sequence alignment, which can save alignment time. Next, a protein graph can be obtained by using a threshold of 0.5 to generate the corresponding adjacency matrix. In this graph, residues represent nodes, and the relationships between residues are represented by the adjacency matrix. Further processing is needed to obtain the features of each node in the graph. Each residue node feature is determined by the different R functional groups it contains. The features of each node are represented using its hydrophobicity, polarity, charge, and aromaticity, as shown in Table 3.

3. Methods

Drug-target relationships can be classified into two categories: DTIs and DTAs.

3.1. Interaction prediction based on molecular drug graphs and protein sequences

The most typical and effective prediction method at

the moment is to treat drug molecules as graphs. Methods that employ the primary structures of proteins are the most common techniques. Examples include CPGL [49], GanDTI [55], GraphCPI [56], MGraphDTA [57], SAG-DTA [58], WGNN-DTA [59]. GanDTI, GraphCPI, MGraphDTA, SAG-DTA, TransformerCPI, and [60] convert the SMILES of a drug into a molecular graph $G = (V,E)$ via the RDKit technique. The feature representations of molecular graphs are learned by GNNs (e.g., GCNs, GATs, graph isomorphism networks (GIN), etc.). Protein features are processed using different techniques. N-gram [61] segmentation, in which protein sequences are segmented into "biological words" for enhanced semantic analysis, is applied by GanDTI, GraphCPI, and TransformerCPI. GanDTI includes a technique that focuses attention on the protein region that interacts most favorably with the drug of interest. Protein sequences are transformed into matrices by Prot2Vec [62] and then fed into a CNN by GraphCPI for feature training. TransformerCPI provides protein sequences to the coding layer of a transformer so that it can learn hidden representations. SAG-DTA incorporates global pooling and hierarchical pooling, as well as self-attention methods, into the molecular graph feature extraction process to obtain two drug representations with rich drug feature representations. A shallow model is not sufficient for capturing global features. Encouraged by DenseNet [63], MGraphDTA was proposed as a deep network model that uses deep

Table 2
Pseudocode for obtaining atomic feature vectors.

Input : The SMILES sequence of the molecule S.
Output : The atomic feature vector V and the adjacency matrix E of the molecule.

Begin
  Obtain the chemical structure mol of S using the MolFromSmiles() function;
  Obtain the atom set $A_{set} = [a^{(1)}, a^{(2)}, \ldots, a^{(n)}]$ corresponding to the molecule mol using the function GetAtoms();
  For each atom $a^{(i)}$ in A_{set} do
    Use the GetSymbol() function to obtain the atomic symbol $f_1^{(i)}$ of $a^{(i)}$;
    Use the GetDegree() function to obtain the degree $f_2^{(i)}$ of $a^{(i)}$;
    Use the GetTotalNumHs() function to obtain the total number of hydrogen atoms $f_3^{(i)}$ of $a^{(i)}$;
    Use the GetImplicitValence() function to obtain the implicit valence $f_4^{(i)}$ of $a^{(i)}$;
    If $a^{(i)}$ is aromatic do
      set $f_5^{(i)} = 1$;
    Else
      set $f_5^{(i)} = 0$;
    End if
    Concatenate the five vectors to obtain the feature vector $V^{(i)}$ for each atom;
    Add $V^{(i)}$ to V.
  End for
  Obtain the edge set $B_{set} = [b^{(1)}, b^{(2)}, \ldots, b^{(m)}]$ corresponding to the molecule mol using the function GetBonds();
  For each bond $b^{(i)}$ in B_{set} do
  Get the two end atoms $e_1^{(i)}$ and $e_2^{(i)}$ of the bond $b^{(i)}$ through the function GetEndAtomIdx();
    Add $e_1^{(i)}$ and $e_2^{(i)}$ to E.
  End for
End

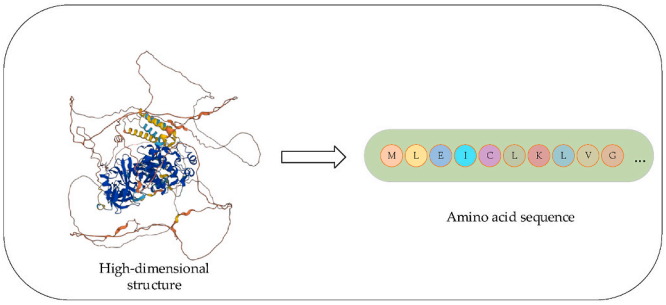

Fig. 2. Sequence representation of a protein. The protein used in this study is Tyrosine-protein kinase ABL1 (P00519), with an amino acid sequence length of 1130. The amino acid sequence shown in the figure represents the first ten characters of the protein sequence.

network learning features for both drug molecules and protein sequences. Models that exhibit outstanding performance on small datasets may not perform well on large datasets due to the lack of available protein resources. As a result, [60] conducted transformer-based pretraining on a large number of proteins.

CPGL and [64] use subgraphs to enhance the representation learning process. The presence of fewer types of atoms and chemical bonds makes it possible to overlook crucial information when learning features. *R*-radius subgraphs [65], where the graph's vertices aggregate the data of all neighboring vertices and edges within a radius *r*, have been introduced as a solution to this issue. Among them, *r* is a parameter, as shown in Fig. 4., where $r = 1$ denotes the first-order neighbor of vertex *i* (such as vertex *j*) and $r = 2$ represents the second-order neighbor of vertex *i* (such as vertex *k*). Sequences based on primary structures are used to represent proteins. To maintain spatially similar amino acids at the same distance in a text sequence, CPGL utilizes LSTM to process amino acid sequence information. LSTM solves the problem of long-term dependence on lengthier sequences, making it ideal for text sequences of amino acids. [64] adopted *N*-gram segmentation and added an attention mechanism to learn feature representations.

3.2. Affinity prediction based on molecular drug graphs and protein sequences

The presence of an interaction between drug and target can be instructive for experts but is still overly simplistic. On the other hand, the ability to forecast particular drug-target association values can be more helpful to drug development team. This task is referred to as "drug target affinity prediction".

DTA prediction has recently observed significant advances as well. The molecular graph of a drug and the primary structure of a target sequence are both described in this section. The transformation of drug strings into molecular graphs remains a common operation, and the majority of proteins involve sequence text information. GraphDTA [66] employs molecular graphs under GNNs (GCNs, GATs, or GINs) to obtain high-level features. For proteins, CNNs are adopted to learn their hidden features. Many methods have been derived based on this idea. For instance, DeepGLSTM [67] and GDGRU-DTA [50] both apply LSTM to learn hidden features for protein processing, and LSTM is more applicable to long sequences of text information. For the processing of molecular graphs, DeepGLSTM combines multiple GCNs to learn multiple drug features, and GDGRU-DTA not only employs GNNs but also utilizes

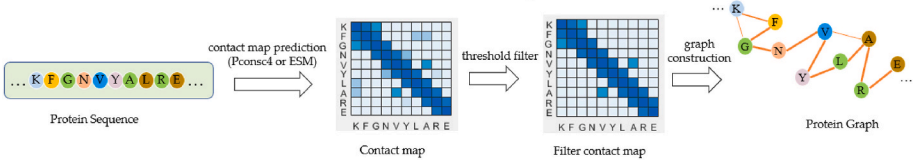

Fig. 3. Construction of a protein graph. The protein sequence is processed by the open-source method Pconsc4 or the ESM model to predict the contact map. Each value in the contact map matrix represents the contact degree between two residues. Then, the Filter contact map matrix is obtained by using a threshold (0.5), which meets the input requirements of the graph neural network. Finally, the protein graph is obtained, and the connected relationships between each node in the graph can be inferred from the Filter contact map.

Table 3
Residue features in protein graph.

Node feature	Dimension
One-hot encoding of residue	21
PSSM	21
Whether is aliphatic	1
Whether is aromatic	1
Whether is polar neutral	1
Whether is acidic charged	1
Whether is basic charged	1
Residue weight	1
Dissociation constant for the –COOH group	1
Dissociation constant for the –NH3 group	1
Dissociation constant for any other group in the molecule	1
The pH at the isoelectric point	1
Hydrophobicity(ph = 2)	1
Hydrophobicity(ph = 7)	1
All	54

bidirectional GRUs [68] to extract drug features. The GRU variation of LSTM processes the same task more quickly than LSTM because it has a more straightforward internal structure. Therefore, GDGRU-DTA adopts GRUs to extract the features of proteins. Motivated by Word2Vec, [51] employed a CNN to extract protein features and *N*-gram segmentation to convert protein sequences into meaningful "biological words". Since the contextual information contained in amino acid and drug sequences are not taken into consideration, DeepGS [69] utilizes Prot2Vec [62] and Smi2Vec [70], respectively, to characterize these two types of sequence information, whereas local information is extracted via a CNN and a bidirectional GRU. After that, the features obtained through the molecular drug graph are combined with this information to make predictions. One-hot protein coding reduces the correlations between atoms, so [71] introduced the term frequency-inverse document frequency (TF-IDF) [47] algorithm from NLP to enhance proteins. The GraphDTA method has been implemented to build drug molecule graphs. Two graphs are utilized by EmbedDTI [72] to describe the chemical structures of drugs and to infer their global and local structures. The protein features are continually learned using a CNN. MGraphDTA [57] explores deeper networks even when shallow models offer superior outcomes. The authors applied an ultradeep GCN with 27 layers, inspired by DenseNet, to capture the chemical structures of drugs. The structures of protein sequences are detected by CNNs with three receptive fields possessing different sizes.

3.3. Interaction and affinity prediction based on molecular drug graphs and protein graphs

Despite the protein-based primary sequences that have been produced, there is still room for improvement. The best protein representations are high-dimensional structures that contain richer structural information. However, these structures cannot be suitably acquired. A two-dimensional graph structure can be employed to describe the three-dimensional structure of a protein. Given the lengths of the input protein sequences, directly translating them to graph structures would raise the training cost. A graph composed of residues is more appropriate because proteins only contain a few hundred residues each. This process is shown in Fig. 3.

Recently, some authors have predicted the relationships between drugs and targets using protein graphs instead of the primary structures of proteins. For example, protein sequences are processed by DGraphDTA [73] to yield a contact map, which is used to build a graph. Utilizing RDKit, the drug molecules are transformed from SMILES to molecule graphs. These two molecular graphs are then loaded into a GNN to enable feature learning. To enrich the learned features while obtaining drug and protein graphs, X-DPI [74] employs Mol2vec [75] for drugs to acquire drug-drug relationships. It then applies a pre-trained

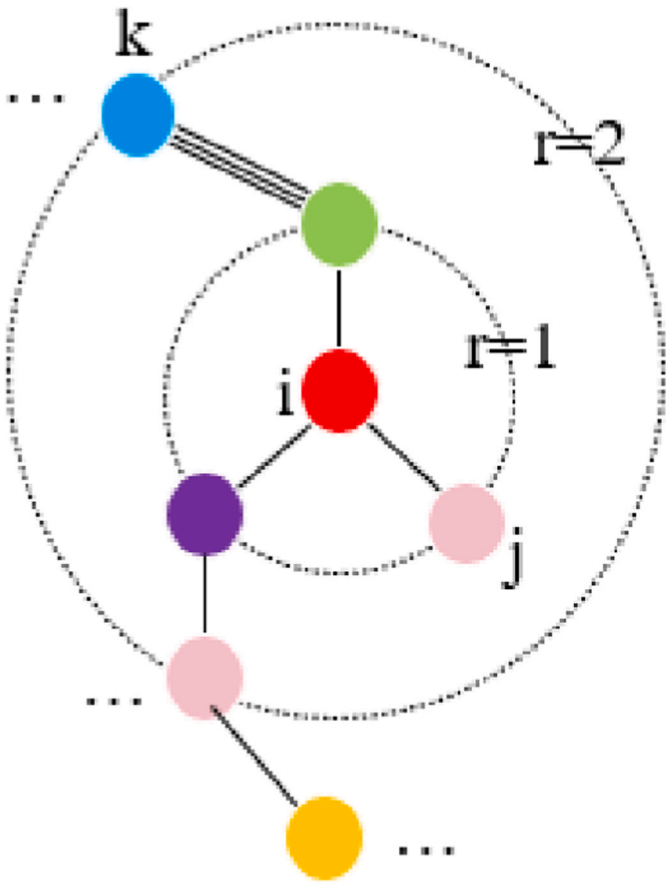

Fig. 4. *r*-radius subgraphs. It is a form of the drug molecular graph. Vertex *i* aggregates distinct neighbor information based on the value of *r*. When *r* is 1 or 2, respectively, vertices *j* and *k* are the neighbors that need to be aggregated for vertices *i*. Circle: carbon atom.

BERT [76] model to learn embedded representations from a large volume of unlabeled protein sequences provided by TAPE [77]. DGraphDTA requires extensive database scanning because of its lengthy processing time, particularly during the sequence alignment step, which greatly lowers the resulting prediction accuracy. Therefore, WGNN-DTA [59] draws on evolutionary scale modeling (ESM) [78] and extracts a protein graph from the model; this approach can efficiently complete the prediction process and produce better results. DGraphDTA ignores the target representation changes induced by DTIs, as it is built by training with a restricted number of proteins. To address this issue, [79] proposed a novel model called graph early fusion affinity (GEFA), which transfers a drug molecule graph to a protein graph for learning and prediction via an attention mechanism.

We provide a summary of the approaches for solving the DTI prediction and DTA prediction tasks based on GNNs, the details of which are displayed in Table 4.

Protein sequence is a linear polypeptide chain composed of amino acids, providing primary structure information of protein molecules. Protein graph is a graphical representation of the high-dimensional structure of proteins, providing visual information about protein structure. The sequence-based method can process the entire sequence and learn its hidden features from a large amount of sequence information. It does not require a deep understanding of the structural features and topology of proteins, so it is relatively less time-consuming and faster in feature engineering and model prediction modules. Due to the use of a large amount of known protein interaction information, its prediction accuracy is acceptable. However, it does not consider the structural

Table 4
Summary of DTI and DTA prediction methods based on GNNs.

Types	Methods	Input drug representation	Input protein representation	Drug feature learning	Protein feature learning	Prediction
A	GanDTI	Molecular graph	Protein sequence	Residual GNN	Attention module	MLP
	GraphCPI	Molecular graph	Protein sequence	GNN (GCN\GAT\GIN)	Prot2Vec [62] + CNN	FC
	MGraphDTA	Molecular graph	Protein sequence	MGNN	MCNN	MLP
	SAG-DTA	Molecular graph	Amino acid sequence	GNN with SAG pooling layers	CNN	FC
	TransformerCPI	Drug molecular	Amino acid sequence	GCN	CNN with gated linear units	Transformer
	[60]	Molecular graph	Protein sequence	GraphNet [85]	Transformer + CNN	FC
	CPGL	R-radius subgraph	Protein sequence	GAT	Bi-LSTM	FC with attention
	[64]	R-radius subgraph	Protein sequence	GNN	CNN	Attention + Softmax
B	GraphDTA	Molecular graph	Protein sequence	GNN (GCN\GAT\GIN\GAT-GCN)	CNN	FC
	DeepGLSTM	Molecular graph	Amino acid sequence	Multiblock GCN	Bi-LSTM	FC
	GDGRU-DTA	Molecular graph	Protein sequence	GatedGraph [86] + TransformerConv [87]	GRU + BiGRU	FC
	[51]	Molecular graph	Amino acid sequence	GCN	2D CNN with Word2Vec [88]	FC
	DeepGS	SMILES sequence + Molecular graph	Amino acid sequence	BiGRU with Smi2Vec [70] + GAT	CNN with Prot2Vec	FC
	[71]	Molecular graph	Protein sequence	GNN (GCN\GAT\GIN\GAT-GCN)	CNN with TF-IDF	FC
	EmbedDTI	Atom graph and substructure graph	Protein sequence	GCN with attention module	CNN	FC
C	DGraphDTA	Molecular graph	Protein graph	GNN (GCN\GAT)	GNN (GCN\GAT)	FC
	WGNN-DTA	Molecular graph	Weighted protein graph	GNN (GCN\GAT)	GNN (GCN\GAT)	FC
	GEFA	Molecular graph	Protein graph + Protein sequence embedding	GCN with residual blocks	GCN with residual blocks	Linear layers
	X-DPI	Molecular graph + Mol2vec embedding	Protein graph + TAPE embedding	GCN encoder	GCN encoder	Transformer decoder

Abbreviations: DTI — drug-target interaction; DTA — drug-target affinity; GNN — graph neural network; MLP — multilayer perceptron; GCN — graph convolutional network; GAT — graph attention network; GIN — graph isomorphism network; CNN — convolutional neural network; FC — fully connected layer; MGNN — multiscale graph neural network; MCNN — multiscale convolutional neural network; SAG — self-attention graph; Bi-LSTM — bidirectional long short-term memory; GRU — gated recurrent unit; BiGRU — bidirectional gated recurrent unit; TF-IDF — term frequency-inverse document frequency.

information of proteins and cannot solve the problem of similarity between sequences, such as homologous proteins leading to biased predictions. The graph-based method considers the topological and functional information of the graph, can mine the complex features of protein structure, and improves the prediction accuracy with high reliability. However, it cannot process each amino acid in the protein molecule as a node of the graph, resulting in a very large structure of the graph that requires more computational resources and time to construct and analyze the protein graph. For unknown or rare protein structures and functions, its model prediction effectiveness may be limited. In conclusion, both sequence-based and graph-based methods have their own advantages and should be selected according to the characteristics of the research question and data features.

4. Experiments

4.1. Datasets

We evaluate several datasets for both DTI prediction and DTA prediction. The former leverages the human and C. elegans datasets, and the latter utilizes the Davis and KIBA datasets. We summarize these datasets, and their statistical information is shown in Table 5.

Table 5
Statistics of the datasets.

	Drugs	Targets	Interactions
DTA			
Davis	68	442	30 056
KIBA	2111	229	118 254
DTI			
Human	2726	2001	6728
C. elegans	1767	1876	7786

Abbreviations: DTI — drug-target interaction; DTA — drug-target affinity.

Davis dataset: The Davis dataset [80] offers 30 056 samples of interaction pairs between 68 drugs and 442 targets. To measure how well a drug binds to a target, the K_d value (kinase dissociation constant) can be utilized. The range of this value is from 0.016 to 10,000. Given that this value range is too large to have an impact on the experimental accuracy, the K_d value is replaced with a logarithmic function (pK_d) which is calculated as described in Eq. (1).

$$pK_d = -log_{10}\left(\frac{K_d}{10^9}\right) \tag{1}$$

KIBA dataset: The KIBA dataset [81] measures the relationships between drugs and targets through KIBA scores, which are calculated by heterogeneous information sources, such as K_i (the inhibition constant), K_d, and IC_{50} (the half-maximal inhibitory concentration). The data, with KIBA scores ranging from 0.0 to 17.2, include 2111 drugs, 229 targets, and 118 254 interaction pairs.

Human and C. elegans datasets: [82] provided highly confident negative samples for both datasets. The datasets are split into balanced and unbalanced samples with ratios of 1:1 and 1:3, respectively, for positive and negative samples (in accordance with [64]). The human dataset consists of 2726 drug molecules, 2001 targets, and 6728 interaction samples. The 7786 interaction samples in the C. elegans dataset include 1767 drugs and 1876 targets.

4.2. Evaluation metrics

We evaluate regression tasks employing the three most popular evaluation indicators, namely, the mean squared error (MSE), concor-

dance index (CI) and r_m^2 index.

MSE: This is an evaluation index that measures the difference between the model's predicted value and the true value. The calculation method is shown in Eq. (2).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{2}$$

CI: The CI is applied to assess how well different models can perform discrimination [83]. The closer the CI value, which ranges from 0 to 1, is to 1, the better the model fit is. The calculation form is shown in Eq. (3).

$$CI = \frac{1}{Z}\sum_{\sigma_i > \sigma_j}\varphi(b_i - b_j) \tag{3}$$

In the equation, the affinity value of σ_i is larger than that of σ_j, and b_i and b_j are the predicted values corresponding to σ_i and σ_j, respectively. Z is a normalized constant, $\varphi(x)$ is a function, and the calculation method of this function is as described in Eq. (4).

$$\varphi(x) = \begin{cases} 0, if\ x < 0 \\ 0.5, if\ x = 0 \\ 1, if\ x > 0. \end{cases} \tag{4}$$

r_m^2 index: The r_m^2 index [84] was used in DeepDTA, and it is also an evaluation index for regression tasks. It can be used to evaluate the external predictive performance of QSAR models. The calculation method is as described in Eq. (5).

$$r_m{}^2 = r^2 \times \left(1 - \sqrt{r^2 - r_0{}^2}\right) \tag{5}$$

where the squares of the correlation coefficients with and without the intercept are denoted as r^2 and $r_0{}^2$, respectively.

We evaluate the model performance on the classification task with four metrics: the **area under the receiver operating characteristic curve (AUC), precision, recall, and F1 score**.

4.3. Results

The positive samples used in the experiments were known drug-target pairs that exhibit interactions or have high affinity. The negative samples consisted of drug-target pairs that do not exhibit interactions or have low affinity. The data used in the experiments were partitioned in the same manner. We divided the data into training, validation, and test sets to evaluate the ability of DTI prediction. The training set was used for optimizing the model's parameters and training process, the validation set was used for tuning the model's hyperparameters and selecting the best model, and the test set was used to assess the final performance of the model.

We experiment with several datasets for the DTI prediction and DTA prediction tasks. On the human and C. elegans datasets, we employ the AUC, precision, recall, and F1 score metrics to evaluate the DTI prediction performance. We deploy the MSE, CI, and r_m^2 evaluation indicators on the Davis and KIBA datasets to estimate DTA accuracy. The source papers serve as the basis for the datasets, model architecture (including the network parameters), and data processing approach employed in the experiment. Please see the corresponding articles for more information.

When compared to sequence-based representations, molecular graph representations can provide more semantic drug information. The primary sequence is applied to describe proteins. Table 6 and Table 7, which are the results of testing on the human and C. elegans datasets, respectively, show the obtained interaction prediction results based on molecular drug graph and protein sequence. Several methods listed in the table obtain better experimental outcomes. The prediction results in terms of the AUC, precision, recall, and F1 score metrics for the other methods on the human and C. elegans datasets are all greater than 0.9, except for the recall scores GraphCPI and [64] on the human dataset, which are 0.890 and 0.883, respectively. This shows that both the drug molecular graph and the protein sequence are fully captured.

Table 6
The interaction prediction based on the drug molecule graph and protein sequence on human dataset.

Methods	AUC	Precision	Recall	F1 score
GanDTI	0.982	0.943	0.941	0.942
GraphCPI[a]	0.973	0.940	0.890	–
MGraphDTA	0.983	0.934	0.947	0.943
SAG-DTA[a] (HierPool)	0.985	0.945	0.933	–
SAG-DTA[a] (GlobPool)	0.984	0.946	0.931	–
TransformerCPI	0.974	0.928	0.939	0.933
[60]	0.979	0.939	0.955	0.947
CPGL	0.981	0.935	0.929	0.932
[64]	0.955	0.958	0.883	0.919

[a] *Represents that the experimental results are acquired from the source paper. Abbreviations: AUC — area under the receiver operating characteristic curve.*

Table 7
The interaction prediction based on the drug molecule graph and protein sequence on the *C. elegans* dataset.

Methods	AUC	Precision	Recall	F1 score
GanDTI	0.984	0.943	0.947	0.945
GraphCPI[a]	0.989	0.937	0.955	–
MGraphDTA	0.988	0.961	0.961	0.961
TransformerCPI	0.972	0.943	0.930	0.937
[60]	0.991	0.955	0.972	0.964
CPGL	0.988	0.956	0.958	0.957
[64]	0.958	0.927	0.908	0.917

[a] *Represents that the experimental results are acquired from the source paper. Abbreviations: AUC — area under the receiver operating characteristic curve.*

The obtained affinity prediction results based on the molecular drug graphs and protein sequences from the Davis and KIBA datasets are shown in Tables 8 and 9, respectively. On various datasets, different methodologies produce various outcomes. For example, on the Davis and KIBA datasets, GraphDTA produces MSE, CI, and r_m^2 values of 0.251, 0.881, and 0.676, and 0.135, 0.894, and 0.772, respectively.

One way to represent molecules is through graphs, which can include more information than sequences. Complete protein sequence representation via graphs is challenging and increases the required training time and memory overhead. Better prediction performance is attained by employing a contact map as the protein graph, which depicts the relationships between protein residues. The results obtained for the DTI and DTA prediction tasks based on drug molecule and protein graphs are shown in Table 10. The MSE, CI, and r_m^2 indicators yielded by the

Table 8
The affinity prediction based on the drug molecule graph and protein sequence on the Davis dataset.

Methods	MSE	CI	r_m^2
GraphDTA	0.251	0.881	0.676
DeepGLSTM	0.250	0.891	0.681
GDGRU-DTA[a] (Transformer-BiGRU)	0.214	0.902	0.697
GDGRU-DTA[a] (GatedGraph-BiGRU)	0.214	0.904	0.708
[51]	0.225	0.892	–
DeepGS	0.252	0.882	0.686
[71]	0.220	0.899	–
EmbedDTI	0.402	0.835	0.495

[a] *Represents that the experimental results are acquired from the source paper. Abbreviations: MSE — mean squared error; CI — concordance index. The r_m^2 index was used in Deep-DTA, which can be used to evaluate the external predictive performance of quantitative structure–activity relationship models.*

Table 9
The affinity prediction based on the drug molecule graph and protein sequence on the KIBA dataset.

Methods	MSE	CI	r_m^2
GraphDTA	0.135	0.894	0.772
DeepGLSTM[a]	0.133	0.897	0.792
GDGRU-DTA[a] (Transformer-BiGRU)	0.134	0.894	0.780
GDGRU-DTA[a] (GatedGraph-BiGRU)	0.137	0.892	0.775
[51] [a]	0.137	0.895	–
DeepGS[a]	0.193	0.860	0.684
[71] [a]	0.126	0.901	–
EmbedDTI[a]	0.133	0.897	–

[a] *Represents that the experimental results are acquired from the source paper. Abbreviations: MSE — mean squared error; CI — concordance index. The r_m^2 index was used in Deep-DTA, which can be used to evaluate the external predictive performance of quantitative structure–activity relationship models.*

Table 10
The interaction and affinity prediction performance achieved based on the drug molecule graphs and protein sequences of several datasets.

Methods	Datasets	MSE	CI	r_m^2
DGraphDTA	Davis	0.240	0.890	0.659
	KIBA	0.147	0.891	0.767
GEFA[a] (Warm start setting)	Davis	0.228	0.893	–
X-DPI[a]	Davis	0.474	–	–
WGNN-DTA	Davis	0.215	0.900	0.711
	KIBA	0.140	0.899	0.749
		AUC	Precision	Recall
	Human	0.978	0.926	0.960
	C.elegans	0.989	0.969	0.954

[a] *Represents that the experimental results are acquired from the source paper. Abbreviations: MSE — mean squared error; CI — concordance index. The r_m^2 index was used in Deep-DTA, which can be used to evaluate the external predictive performance of quantitative structure–activity relationship models.*

DGraphDTA and GEFA methods on the Davis and KIBA datasets are promising. WGNN-DTA is almost better than the affinity prediction and interaction prediction methods based on drug molecule graphs and protein sequences on the Davis and C. elegans datasets, respectively. Moreover, it behaves as well as the other methods on the KIBA and human datasets. The experimental results once more show that protein graphs play an important role in the prediction of DTIs and DTAs.

5. Discussion and conclusion

We summarize DTI prediction and DTA prediction based on GNNs. Drug molecule graphs, primary protein sequences, and protein graphs are classified for interaction prediction based on the drug molecule graphs and protein sequences, affinity prediction based on the drug molecule graphs and protein sequences, and overall prediction based on the drug molecule graphs and protein graphs. On several fundamental datasets, representative approaches are selected for experimental validation. The outcomes demonstrate the superiority of molecular graph-based representations over sequence-based representations. Molecular graphs have access to greater semantic information for feature learning than molecular sequences. The interactions between atoms are more accurately reflected in the topologies of molecular networks.

In contrast to molecular graphs, although molecular sequences can gather sequence information for feature learning, the structural information contained in molecular graphs is ignored, which affects the accuracy of prediction. Drug character sequences are currently replaced by molecular graphs, which include a wealth of feature information. Many target proteins lack 3D structure information, making it impossible to exploit their high-dimensional properties. Additionally, because target protein sequences are rather long, utilizing graphs to represent them causes the training time to increase. A protein graph cannot be used to depict a whole protein sequence in practice. However, with the emergence of AlphaFold, a breakthrough in protein structure prediction, high-quality predictions of protein structures are now available. This development has significantly impacted the field of protein structure prediction, as it allows for the approximation of 3D structures for many target proteins that previously lacked structural information. Therefore, a contact map, which is the output of the protein structure prediction process, is introduced, and the adjacency matrix of the protein residue pairings is described by this map. The protein graph constructed from the contact map is sent to a neural network to learn features.

Predicting the interactions between drug molecules and protein sequences, as well as predicting the affinity between them, is a common computational method in drug development. A typical application scenario of using these methods based on drug molecule graphs and protein sequences is drug discovery and design. In drug discovery, researchers usually look for specific proteins as drug targets. By using this prediction method, the interactions between the drug targets can be determined, and the binding site and interaction type of the drug molecule can be identified. These prediction results can be used to design new drug molecules, optimize existing ones, or evaluate the interaction abilities of different drug molecules for better disease treatment. A typical application scenario of using these methods based on drug molecule graphs and protein sequences for affinity prediction is drug screening and optimization. By exploring the affinity between drug molecules and specific proteins, drug molecules with high affinity can be selected and evaluated for their activity, selectivity, and toxicity. This method focuses on evaluating and optimizing existing drug molecules to improve their efficacy and safety. Protein graphs can help researchers understand the structure and function of proteins. In summary, DTI and DTA have important applications in drug development. They can help drug researchers quickly and effectively understand the mechanisms of interaction between drug molecules and proteins at different stages of drug development.

There is still room for improvement even though molecular graph representations are richer than molecular sequences. Protein contact map creation is a time-consuming process. In future work, protein graphs can be constructed with a sped-up process. For unknown protein structure information, researchers should find a way to identify the relationships between drugs and targets.

Furthermore, GNNs have proven to be effective in capturing and leveraging the relationships between drugs and proteins, providing strong support for biological interpretability in DTI and DTA applications. GNNs can model the structural relationships, functional regions, and multimodal features between drugs and proteins, generating interpretable predictions and visualizations. These capabilities contribute to uncovering the mechanisms of drug-protein interactions, the functional substructures, and the importance of drug features, offering deeper insights and guidance for drug design and development.

In terms of structural feature modeling, GNNs can represent the structures of drugs and proteins as nodes and edges in a graph. By learning the interactions and information propagation among the nodes, GNNs can extract structural features of drugs and proteins and explain their importance in interaction prediction. This helps to reveal the mechanisms and structural basis of drug-protein interactions.

For protein substructure and functional analysis, GNNs can model the substructures of proteins and provide explanations for different functional regions in predictions. By modeling the topological structures, domain structures, or other specific features of proteins, GNNs can assist in identifying and interpreting functional regions relevant to drug binding, thereby offering explanatory capabilities for substructures and functions.

Regarding the multimodal representation of drugs, GNNs support the

modeling of diverse features and their fusion in a graph network. These multimodal features can include molecular structures, chemical properties, drug metabolism information, and more. By learning the multimodal representations of drugs, GNNs can provide explanations for the importance of different features, aiding in the understanding of the mechanisms underlying drug-protein interactions.

Moreover, the application of GNNs in DTI or DTA prediction can yield visual and interpretable results. By visualizing the relationship graphs between drugs and proteins, the weights of important nodes and edges, as well as the features associated with the prediction results, GNNs offer an intuitive and interpretable way to explain the prediction outcomes and the decision-making process of the prediction model.

And, The diversity of data is also reflected in molecular graphs, where richer node information leads to more effective predictions. Single-mode data models may have certain limitations, but combining multimodal data can provide more comprehensive information. Gene expression data can provide information on protein expression levels, cell states, and molecular affinity-related information. Molecular graphs describe the chemical structural characteristics between molecules, and fusing these two types of data can help better understand the interactions between drug molecules and proteins. Multimodal data for molecules is not limited to gene expression; drug molecule graphs can also carry information on toxicity, side effects, and protein graphs can include functional area data, among others. All these multimodal data can contribute to model predictions.

In developing drug-target interaction prediction models, uncertainties such as data noise, missing data, and unknown structure are also considered. These uncertainties may lower prediction results and reduce model interpretability. Bayesian networks and probabilistic models can be used to improve data reliability in response to these issues. Model interpretability is also a key concern, as it is important for guiding experimental design, accelerating new drug development, reducing unnecessary experiments, and lowering research and development costs. The visualization method of the attention mechanism can observe the contribution of nodes in the graph to experimental results. In the future, researchers should design a better and more interpretable model to predict DTI and DTA by deepening their understanding of model interpretability. Furthermore, after learning drug and target features, a better binding strategy can be adopted instead of simple connections.

Declaration of competing interest

The authors declare that they have no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (2022YFE0112200), the Key-Area Research and Development of Guangdong Province (2022A0505050014, 2022B1111050002), and the National Natural Science Foundation of China (U21A20520, 62172112).

References

[1] T.T. Ashburn, K.B. Thor, Drug repositioning: identifying and developing new uses for existing drugs, Nat. Rev. Drug Discov. 3 (8) (Aug. 2004) 673–683.
[2] A.D. Roses, Pharmacogenetics in drug discovery and development: a translational perspective, Nat. Rev. Drug Discov. 7 (10) (Oct. 2008) 807–817.
[3] J.A. DiMasi, H.G. Grabowski, R.W. Hansen, Innovation in the pharmaceutical industry: new estimates of R&D costs, J. Health Econ. 47 (May. 2016) 20–33.
[4] A. Mullard, New drugs cost US$2.6 billion to develop, Nat. Rev. Drug Discov. 13 (12) (Dec. 2014), 12.
[5] J.C. Pereira, E.R. Caffarena, C.N. dos Santos, Boosting docking-based virtual screening with deep learning, J. Chem. Inf. Model. 56 (12) (Dec. 2016) 2495–2506.
[6] S.M. Strittmatter, Overcoming drug development bottlenecks with repurposing: old drugs learn new tricks, Nat. Med. 20 (6) (Jun. 2014) 590–591.
[7] S. Pathak, X. Cai, Ensemble learning algorithm for drug-target interaction prediction, in: 2017 IEEE 7th International Conference on Computational Advances in Bio and Medical Sciences, ICCABS, Orlando, FL, USA, Oct. 2017, 1–1.
[8] M. Deshpande, M. Kuramochi, N. Wale, G. Karypis, Frequent substructure-based approaches for classifying chemical compounds, IEEE Trans. Knowl. Data Eng. 17 (8) (Aug. 2005) 1036–1050.
[9] M.J. Keiser, B.L. Roth, B.N. Armbruster, P. Ernsberger, J.J. Irwin, B.K. Shoichet, Relating protein pharmacology by ligand chemistry, Nat. Biotechnol. 25 (2) (Feb. 2007) 197–206.
[10] D.T. Ahneman, J.G. Estrada, S. Lin, S.D. Dreher, A.G. Doyle, Predicting reaction performance in C-N cross-coupling using machine learning, Science 360 (6385) (Apr. 2018) 186–190.
[11] K.V. Balakin, S.E. Tkachenko, S.A. Lang, I. Okun, A.A. Ivashchenko, N.P. Savchuk, Property-based design of GPCR-targeted library, J. Chem. Inf. Comput. Sci. 42 (6) (Nov. 2002) 1332–1342.
[12] F. Napolitano, Y. Zhao, V.M. Moreira, R. Tagliaferri, J. Kere, M. D'Amato, D. Greco, Drug repositioning: a machine-learning approach through data integration, J. Cheminf. 5 (1) (Dec. 2013) 30.
[13] Z. Liu, H. Fang, K. Reagan, X. Xu, D.L. Mendrick, W. Slikker, W. Tong, In silico drug repositioning: what we need to know, Drug Discov. Today 18 (3–4) (Feb. 2013) 110–115.
[14] A.C. Cheng, R.G. Coleman, K.T. Smyth, Q. Cao, P. Soulard, D.R. Caffrey, A. C. Salzberg, E.S. Huang, Structure-based maximal affinity model predicts small-molecule druggability, Nat. Biotechnol. 25 (1) (Jan. 2007) 71–75.
[15] P.T. Lang, S.R. Brozell, S. Mukherjee, E.F. Pettersen, E.C. Meng, V. Thomas, R. C. Rizzo, D.A. Case, T.L. James, I.D. Kuntz, DOCK 6: combining techniques to model RNA-small molecule complexes, RNA 15 (6) (Jun. 2009) 1219–1230.
[16] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, A. J. Olson, AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility, J. Comput. Chem. 30 (16) (Dec. 2009) 2785–2791.
[17] W. Filgueira de Azevedo, G.C. dos Santos, D.M. dos Santos, J.R. Olivieri, F. Canduri, R.G. Silva, L.A. Basso, G. Renard, I.O. da Fonseca, M.A. Mendes, M. S. Palma, D.S. Santos, Docking and small angle X-ray scattering studies of purine nucleoside phosphorylase, Biochem. Biophys. Res. Commun. 309 (4) (Oct. 2003) 923–928.
[18] N.M.B. Levin, V.O. Pintro, M.B. de Avila, B.B. de Mattos, W.F. De Azevedo, Understanding the structural basis for inhibition of cyclin-dependent kinases. new pieces in the molecular puzzle, Curr. Drug Targets 18 (9) (Jun. 2017) 1104–1111.
[19] L. Chen, X. Tan, D. Wang, F. Zhong, X. Liu, T. Yang, X. Luo, K. Chen, H. Jiang, M. Zheng, TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments, Bioinformatics 36 (16) (Aug. 2020) 4406–4414.
[20] F. Wan, L. Hong, A. Xiao, T. Jiang, J. Zeng, NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions, Bioinformatics 35 (1) (Jan. 2019) 104–111.
[21] S. Hu, C. Zhang, P. Chen, P. Gu, J. Zhang, B. Wang, Predicting drug-target interactions from drug structure and protein sequence using novel convolutional neural networks, BMC Bioinf. 20 (Suppl 25) (Dec. 2019) 689.
[22] H. Öztürk, A. Özgür, E. Ozkirimli, DeepDTA: deep drug–target binding affinity prediction, Bioinformatics 34 (17) (Sep. 2018) i821–i829.
[23] H. Öztürk, E. Ozkirimli, A. Özgür, WideDTA: Prediction of Drug-Target Binding Affinity, *The Institute of Electronics, Information and Communication Engineers*, TOKYO , JAPAN, Feb. 2019 [Online], https://ui.adsabs.harvard.edu/abs/2019arXiv190204166O. (Accessed 23 November 2022).
[24] T. Li, X.-M. Zhao, L. Li, Co-VAE: drug-target binding affinity prediction by co-regularized variational autoencoders, IEEE Trans. Pattern Anal. Mach. Intell. 44 (12) (Dec. 2022) 8861–8873.
[25] M. Yazdani-Jahromi, N. Yousefi, A. Tayebi, E. Kolanthai, C.J. Neal, S. Seal, O. O. Garibay, AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification, Briefings Bioinf. 23 (4) (Jul. 2022) bbac272.
[26] P. Bai, F. Miljković, B. John, H. Lu, Interpretable bilinear attention network with domain adaptation improves drug–target prediction, Nat. Mach. Intell. 5 (2) (Feb. 2023) 126–136.
[27] Q. Zhao, H. Zhao, K. Zheng, J. Wang, HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism, Bioinformatics 38 (3) (Jan. 2022) 655–662.
[28] W. Yuan, G. Chen, C.Y.-C. Chen, FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction, Briefings Bioinf. 23 (1) (Jan. 2022) bbab506.
[29] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, J. Zeng, DTINet:a network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information, Nat. Commun. 8 (1) (Dec. 2017) 573.
[30] Y. Liu, M. Wu, C. Miao, P. Zhao, X.-L. Li, Neighborhood regularized logistic matrix factorization for drug-target interaction prediction, PLoS Comput. Biol. 12 (2) (Feb. 2016), e1004760.
[31] T. Ban, M. Ohue, Y. Akiyama, NRLMFβ: beta-distribution-rescored neighborhood regularized logistic matrix factorization for improving the performance of drug–target interaction prediction, Biochem. Biophys. Rep. 18 (Jul. 2019), 100615.

[32] M. Gönen, Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization, Bioinformatics 28 (18) (Sep. 2012) 2304–2310.

[33] L.-Y. Xia, Z.-Y. Yang, H. Zhang, Y. Liang, Improved prediction of drug–target interactions using self-paced learning with collaborative matrix factorization, J. Chem. Inf. Model. 59 (7) (Jul. 2019) 3340–3351.

[34] X. Gao, X. Ma, W. Zhang, J. Huang, H. Li, Y. Li, J. Cui, Multi-View clustering with self-representation and structural constraint, IEEE Transact. Big Data 8 (4) (Aug. 2022) 882–893.

[35] W. Wu, X. Ma, Network-based structural learning nonnegative matrix factorization algorithm for clustering of scRNA-seq data, IEEE ACM Trans. Comput. Biol. Bioinf 20 (1) (Jan. 2023) 566–575.

[36] D. Li, S. Zhang, X. Ma, Dynamic module detection in temporal attributed networks of cancers, IEEE ACM Trans. Comput. Biol. Bioinf 19 (4) (Jul. 2022) 2219–2230.

[37] X. Ma, W. Zhao, W. Wu, Layer-specific modules detection in cancer multi-layer networks, IEEE ACM Trans. Comput. Biol. Bioinf 20 (2) (Mar. 2023) 1170–1179.

[38] Z. Huang, Y. Wang, X. Ma, Clustering of cancer attributed networks by dynamically and jointly factorizing multi-layer graphs, IEEE ACM Trans. Comput. Biol. Bioinf 19 (5) (Sep. 2022) 2737–2748.

[39] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: a review of methods and applications, AI Open 1 (Jan. 2020) 57–81.

[40] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1) (Feb. 1988) 31–36.

[41] M. Sun, S. Zhao, C. Gilvary, O. Elemento, J. Zhou, F. Wang, Graph convolutional networks for computational drug development and discovery, Briefings Bioinf. 21 (3) (May 2020) 919–935.

[42] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P.S. Yu, A comprehensive survey on graph neural networks, IEEE Transact. Neural Networks Learn. Syst. 32 (1) (Jan. 2021) 4–24.

[43] X. Yue, Z. Wang, J. Huang, S. Parthasarathy, S. Moosavinasab, Y. Huang, S.M. Lin, W. Zhang, P. Zhang, H. Sun, Graph embedding on biomedical networks: methods, applications and evaluations, Bioinformatics 36 (4) (Feb. 2020) 1241–1251.

[44] J. Lim, S. Ryu, K. Park, Y.J. Choe, J. Ham, W.Y. Kim, Predicting drug-target interaction using a novel graph neural network with 3D structure-embedded graph representation, J. Chem. Inf. Model. 59 (9) (Sep. 2019) 3981–3988.

[45] Z. Liu, Q. Chen, W. Lan, H. Pan, X. Hao, S. Pan, GADTI: graph autoencoder approach for DTI prediction from heterogeneous network, Front. Genet. 12 (Apr. 2021), 650821.

[46] T.D. Crawford, M. Ruchardt, G. Torrance, RDKit: a Comprehensive Toolkit for Drug Discovery, vol. 46, Academic Press, 2011, pp. 15–28 [Online], http://rdkit.org/. (Accessed 5 March 2023).

[47] J. Kaur, J. Saini, Designing punjabi poetry classifiers using machine learning and different textual features, Int. Arab J. Inf. Technol. (2019) 38–44.

[48] Q. Feng, E.V. Dueva, A. Cherkasov, M. Ester, PADME: a deep learning-based framework for drug-target interaction prediction, Comput. Res. Repository abs/1807.09741 (2018) [Online], http://arxiv.org/abs/1807.09741. (Accessed 23 November 2022).

[49] M. Zhao, M. Yuan, Y. Yang, S.X. Xu, CPGL: prediction of compound-protein interaction by integrating graph attention network with long short-term memory neural network, IEEE ACM Trans. Comput. Biol. Bioinf (2022) 2022, 04.19.488691, Apr.

[50] L. Zhijian, J. Shaohua, L. Yigao, G. Min, GDGRU-DTA: predicting drug-target binding affinity based on GNN and double GRU, in: Data Mining and Machine Learning, Apr. 2022, pp. 25–37.

[51] M. Xia, J. Hu, X. Zhang, X. Lin, Drug-target binding affinity prediction based on graph neural networks and word2vec, in: D.-S. Huang, K.-H. Jo, J. Jing, P. Premaratne, V. Bevilacqua, A. Hussain (Eds.), Intelligent Computing Theories and Application, Lecture Notes in Computer Science, vol. 13394, Springer International Publishing, Cham, 2022, pp. 496–506.

[52] B.-W. Zhao, Z.-H. You, L. Hu, Z.-H. Guo, L. Wang, Z.-H. Chen, L. Wong, A novel method to predict drug-target interactions based on large-scale graph representation learning, Cancers 13 (9) (Apr. 2021) 2111.

[53] M. Lennox, N. Robertson, B. Devereux, Modelling drug-target binding affinity using a BERT based graph neural network, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), vol. 2021, Nov. 2021, pp. 4348–4353.

[54] Q. Wu, Z. Peng, I. Anishchenko, Q. Cong, D. Baker, J. Yang, Protein contact prediction using metagenome sequence data and residual neural networks, Bioinformatics 36 (1) (Jan. 2020) 41–48.

[55] S. Wang, P. Shan, Y. Zhao, L. Zuo, GanDTI: a multi-task neural network for drug-target interaction prediction, Comput. Biol. Chem. 92 (Jun. 2021), 107476.

[56] Z. Quan, Y. Guo, X. Lin, Z.-J. Wang, X. Zeng, GraphCPI: graph neural representation learning for compound-protein interaction, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine, 2019, pp. 717–722. San Diego, CA, USA, Nov.

[57] Z. Yang, W. Zhong, L. Zhao, C. Yu-Chian Chen, MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction, Chem. Sci. 13 (3) (2022) 816–833.

[58] S. Zhang, M. Jiang, S. Wang, X. Wang, Z. Wei, Z. Li, SAG-DTA: prediction of drug–target affinity using self-attention graph network, Int. J. Mol. Sci. 22 (16) (Aug. 2021) 8993.

[59] M. Jiang, S. Wang, S. Zhang, W. Zhou, Y. Zhang, Z. Li, WGNN-DTA: sequence-based drug-target affinity prediction using weighted graph neural networks, BMC Genom. 23 (1) (Dec. 2022) 449.

[60] Qh Kim, J.-H. Ko, S. Kim, N. Park, W. Jhe, Bayesian neural network with pretrained protein embedding enhances prediction accuracy of drug-protein interaction, Bioinformatics 37 (20) (Oct. 2021) 3428–3435.

[61] Q.-W. Dong, X.-L. Wang, L. Lin, Application of latent semantic analysis to protein remote homology detection, Bioinformatics 22 (3) (Feb. 2006) 285–290.

[62] E. Asgari, M.R.K. Mofrad, Continuous distributed representation of biological sequences for deep proteomics and genomics, PLoS One 10 (11) (2015), e0141287.

[63] G. Li, M. Mueller, G. Qian, I.C. Delgadillo Perez, A. Abualshour, A.K. Thabet, B. Ghanem, DeepGCNs: Making GCNs Go as Deep as CNNs, IEEE Transactions on Pattern Analysis and Machine Intelligence, Apr. 2021, 1–1.

[64] M. Tsubaki, K. Tomii, J. Sese, Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences, Bioinformatics 35 (2) (Jan. 2019) 309–318.

[65] F. Costa, K.D. Grave, Fast neighborhood subgraph pairwise distance kernel, in: Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, Jun. 2010, pp. 255–262. (Accessed 23 November 2022).

[66] T. Nguyen, H. Le, T.P. Quinn, T. Nguyen, T.D. Le, S. Venkatesh, GraphDTA: predicting drug-target binding affinity with graph neural networks, Bioinformatics 37 (8) (May 2021) 1140–1147.

[67] S. Mukherjee, M. Ghosh, P. Basuchowdhuri, DeepGLSTM: deep graph convolutional network and LSTM based approach for predicting drug-target binding affinity, arXiv (Jan. 18, 2022) [Online], http://arxiv.org/abs/2201.06872. (Accessed 5 October 2022).

[68] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Gated feedback recurrent neural networks, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning, vol. 37, Jul. 2015, pp. 2067–2075. Lille, France.

[69] X. Lin, K. Zhao, T. Xiao, Z. Quan, Z.-J. Wang, P.S. Yu, DeepGS: deep representation learning of graphs and sequences for drug-target binding affinity prediction, in: ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020 vol. 325, Santiago de Compostela, Spain, 2020, pp. 1301–1308. August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020).

[70] Z. Quan, X. Lin, Z.-J. Wang, Y. Liu, F. Wang, K. Li, A system for learning atoms based on long short-term memory recurrent neural networks, in: IEEE International Conference on Bioinformatics and Biomedicine, 2018, pp. 728–733. December 3-6, 2018, Madrid, Spain.

[71] X. Wang, Y. Liu, F. Lu, H. Li, P. Gao, D. Wei, Dipeptide frequency of word frequency and graph convolutional networks for DTA prediction, Front. Bioeng. Biotechnol. 8 (Apr. 2020) 267.

[72] Y. Jin, J. Lu, R. Shi, Y. Yang, EmbedDTI: enhancing the molecular representations via sequence embedding and graph convolutional network for the prediction of drug-target interaction, Biomolecules 11 (12) (1783) 2021.

[73] M. Jiang, Z. Li, S. Zhang, S. Wang, X. Wang, Q. Yuan, Z. Wei, DGraphDTA: drug–target affinity prediction using graph neural network and contact maps, RSC Adv. 10 (35) (2020) 20701–20712.

[74] P. Wang, S. Zheng, Y. Jiang, C. Li, J. Liu, C. Wen, A. Patronov, D. Qian, H. Chen, Y. Yang, X-DPI: a Structure-Aware Multi-Modal Deep Learning Model for Drug-Protein Interactions Prediction, Cold Spring Harbor Laboratory, Jun. 2021.

[75] S. Jaeger, S. Fulle, S. Turk, Mol2vec: unsupervised machine learning approach with chemical intuition, J. Chem. Inf. Model. 58 (1) (Jan. 2018) 27–35.

[76] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, 2019, pp. 4171–4186. June 2-7, 2019, Minneapolis, MN, USA.

[77] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, X. Chen, J. Canny, P. Abbeel, Y. S. Song, Evaluating protein transfer learning with TAPE, Adv. Neural Inf. Process. Syst. 32 (Dec. 2019) 9689–9701.

[78] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, A. Rives, Transformer protein language models are unsupervised structure learners, in: 9th International Conference on Learning Representations, May 3-7, 2021, Vienna, Austria, 2021.

[79] T.M. Nguyen, T. Nguyen, T.M. Le, T. Tran, GEFA: early fusion approach in drug-target affinity prediction, IEEE ACM Trans. Comput. Biol. Bioinf 19 (2) (Mar. 2022) 718–728.

[80] M.I. Davis, J.P. Hunt, S. Herrgard, P. Ciceri, L.M. Wodicka, G. Pallares, M. Hocker, D.K. Treiber, P.P. Zarrinkar, Comprehensive analysis of kinase inhibitor selectivity, Nat. Biotechnol. 29 (11) (2011) 1046–1051, Nov.

[81] J. Tang, A. Szwajda, S. Shakyawar, T. Xu, P. Hintsanen, K. Wennerberg, T. Aittokallio, Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis, J. Chem. Inf. Model. 54 (3) (Mar. 2014) 735–743.

[82] H. Liu, J. Sun, J. Guan, J. Zheng, S. Zhou, Improving compound-protein interaction prediction by building up highly credible negative samples, Bioinformatics 31 (12) (Jun. 2015) i221–i229.

[83] M. Gonen, G. Heller, Concordance probability and discriminatory power in proportional hazards regression, Biometrika 92 (4) (2005) 965–970.

[84] K. Roy, P. Chakraborty, I. Mitra, P.K. Ojha, S. Kar, R.N. Das, Some case studies on application of ' r_m ~2 ' metrics for judging quality of quantitative structure-activity relationship predictions: emphasis on scaling of response data, J. Comput. Chem. 34 (12) (May 2013) 1071–1082.

[85] P. Battaglia, R. Pascanu, M. Lai, D.J. Rezende, K. kavukcuoglu, Interaction networks for learning about objects, relations and physics, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, Dec. 2016, pp. 4509–4517.

[86] Y. Li, R. Zemel, M. Brockschmidt, D. Tarlow, in: "Gated Graph Sequence Neural Networks," *Presented at the Proceedings of the 32nd International Conference on International Conference on Machine Learning*, Lille, France, Jul. 2015, pp. 2067–2075 [Online], http://arxiv.org/abs/1511.05493. (Accessed 28 November 2022).

[87] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, Y. Sun, in: "Masked Label Prediction: Unified Message Passing Model for Semi-supervised Classification," *Presented at the 29th International Joint Conference on Artificial Intelligence*, Yokohama, Japan, vol. 2, Aug. 2021, pp. 1548–1554.

[88] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems, vol. 2, Red Hook, NY, USA, Dec. 2013, pp. 3111–3119.