BMC Bioinformatics

RESEARCH

Open Access

SubMDTA: drug target affinity prediction based on substructure extraction and multi-scale features

Shourun Pan[1], Leiming Xia[1], Lei Xu[1] and Zhen Li[1*]

*Correspondence:
lizhen0130@gmail.com

[1] College of Computer Science and Technology, Qingdao University, Qingdao, China

Abstract

Background: Drug–target affinity (DTA) prediction is a critical step in the field of drug discovery. In recent years, deep learning-based methods have emerged for DTA prediction. In order to solve the problem of fusion of substructure information of drug molecular graphs and utilize multi-scale information of protein, a self-supervised pre-training model based on substructure extraction and multi-scale features is proposed in this paper.

Results: For drug molecules, the model obtains substructure information through the method of probability matrix, and the contrastive learning method is implemented on the graph-level representation and subgraph-level representation to pre-train the graph encoder for downstream tasks. For targets, a BiLSTM method that integrates multi-scale features is used to capture long-distance relationships in the amino acid sequence. The experimental results showed that our model achieved better performance for DTA prediction.

Conclusions: The proposed model improves the performance of the DTA prediction, which provides a novel strategy based on substructure extraction and multi-scale features.

Keywords: Drug–target binding affinity, Self-supervised learning, Mutual information, Multi-scale features

Introduction

Drug development is a complex progress involving long research cycles, high costs, and low success rates, which could take several decades and 400–900 million dollars for a new drug from screening small molecules to market approval [1]. In the past few years, the information technology has been widely applied in computer-aided drug design (CADD) methods to accelerate the speed of drug development [2]. The prediction of drug–target binding affinity (DTA) is an important step in drug discovery, which provides information on the strength of interaction between drug molecules and target proteins. Therefore, the development of efficient and accurate algorithm of DTA prediction is of great significance in CADD.

Early computer virtual screening mainly focused on two types of methods: molecular docking [3–5] and ligand-based similarity [6, 7]. The molecular docking technique utilized the three-dimensional structure of protein targets and drug molecules, and the affinity can be predicted by simulating the docking process of proteins and molecules [8, 9]. However, the acquisition of three-dimensional structures is difficult, and large-scale molecular docking process is time-consuming. In contrast to molecular docking, ligand-based methods do not rely on the three-dimensional structure of molecules, which predict DTA by comparing new ligands with known ligands. However, when the number of known ligands is insufficient, the ability of ligand-based approach is limited. In response to these challenges, machine learning methods for DTA prediction [10–12] have been gradually introduced in the virtual screening and improved the performance of DTA prediction. Wang et al. [10] treated the interaction between drugs and targets as a binary classification problem. After extracting chemical descriptors of drugs and protein sequence information, an SVM model was used for prediction. KronRLS [11] used PubChem structure clustering tool [13] and Smith Waterman algorithm [14] to obtain similarity matrices for drugs and proteins, and the Kronecker product of similarity matrices was used to define similarity scores for drug–target pairs. To alleviate the limitation of linear dependence in KronRLS, SimBoost [12] constructed a drug–target similarity network and established a gradient boosting regression tree model for prediction. However, these machine learning methods rely on carefully designed handcrafted features, and the selection of these features depends on specific domain knowledge and experience [15] As deep learning (DL) methods have demonstrated superior learning capabilities over traditional machine learning methods in multiple fields, they have gradually been applied to solve problems in bioinformatics, including the DTA prediction [16–22]. DeepDTA [16] used protein sequence and molecular sequence information in two separate CNN networks. The output feature vectors were concatenated and fed into three fully connected layers to predict binding affinity. DeepCDA [17] combined CNN and LSTM to encode protein sequence and molecular sequence and proposed a bidirectional attention mechanism to predict DTA. FusionDTA [18] replaced the coarse pooling method with a novel multi-head linear attention mechanism to aggregate global information to address the issue of information loss. Additionally, the knowledge distillation was applied to transfer learnable information from a teacher model to a student model to solve the problem of parameter redundancy. As molecule could be represented as a graph, in which chemical atoms and bonds can be represented by nodes and edges. With the rapid development of graph neural networks (GNN), researchers have applied GNN models to DTA prediction. GraphDTA [19] used the topological structure information of molecular graphs and different GNN models for drug representation, while CNN was used to learn protein representation which is similar with DeepDTA. DGraphDTA [20] constructed protein graphs based on protein contact maps for the first time, then the GNN was used to predict DTA through molecular graphs and protein graphs. MGraphDTA [21] constructed a super-deep GNN with 27 graph convolution layers by introducing dense connections to capture both local and global structures of molecules. These methods indicate that deep learning networks can better capture the features of drugs and proteins. Due to the high cost and time consumption of laboratory experiments, the size of training dataset for drug discovery is limited, which may cause overfitting

problems for machine learning methods and affect the generalization of learned features. Self-supervised learning can use unlabeled data for pre-training and transfer the learned model to downstream tasks, which can alleviate the requirement for labeled data. There are also self-supervised learning methods used in drug discovery [23]. InfoGraph [24] maximized the mutual information between graph embedding and substructure embedding at different scales to learn graph representations. MPG [25] compared two half-graphs and distinguished whether they come from the same source as a self-supervised learning strategy. GROVER [26] proposed two pre-training tasks: for the node/edge level task, it randomly masked a local subgraph of the target node/edge and predicted the contextual property; for the graph level task, it extracted the semantic motifs existing in molecular graphs (such as functional groups) and predicted whether these motifs existed for a molecule. However, most existing research integrated all structural features and node attributes of the graph to provide an overview of the graph, ignoring more fine-grained substructure semantics. Proteins are macromolecules composed of amino acids. There are 22 amino acids that make up an organism, which are represented by 22 letters and can be naturally represented as a sequence of letters. Sequence-based DL models can effectively consider the contextual relationships of the sequences. MATT-DTI [27] utilized three convolutional layers as the feature extractor, followed by a max pooling layer. A multi-head attention block was built to model the similarity of drug–target pairs as the interaction information for DTA prediction. TransformerCPI [28] used a one-dimensional convolutional gated convolutional network and gated linear unit instead of the self-attention layer in the Transformer encoder. However, current studies focus only on the single scale of protein sequences, and traditional sequence-based approaches process the whole sequence at once may lead to the loss of local information and neglect multi-scale features of proteins, so how to combine multi-scale information to improve the robust of protein representation is also an open issue. In order to overcome the limitations of existing methods, we propose a novel framework, SubMDTA, a drug target affinity prediction method based on substructure extraction and multi-scale features. For molecules, inspired by Wang et al. [29], a self-supervised learning method based on molecular substructure is proposed for molecular representation. During the pre-training phase, subgraphs are generated to obtain substructure information, and subgraphs are replaced according to their similarity relationships to generate reconstructed graphs. We simultaneously maximize the mutual information between the subgraph and the original graph, as well as between the reconstructed graph and the original graph, to improve the correlation between subgraph-level and graph-level representations. After pre-training, the trained model is fine-tuned in downstream tasks. For proteins, a BiLSTM method that integrates multi-scale information based on n-gram method is proposed for feature extraction. Finally, the drug and protein features are concatenated and fed into a Multilayer Perceptron (MLP) for DTA prediction. We compared our proposed method with several state-of-the-art methods and the experimental results demonstrate that our method significantly outperforms other methods on the Davis [30] and KIBA [31] datasets.

Materials and methods

The SubMDTA performs DTA prediction by integrating structural information of drug molecules and sequence features of targets, and the general architecture of SubMDTA is shown in Fig. 1. It consists of a pre-training part and a DTA prediction part. In the pre-training part, the drug SMILES (Simplified Molecular Input Line Entry System) [32] strings in the pre-training dataset are first converted into molecular graphs, followed by encoding the graph representations using the GIN [33] network. Then the substructural and reconstruct graphs are extracted. After obtaining two types of features, the mutual information between them and the original graph are maximized. The DTA prediction part uses the trained GIN encoder for molecular representation. For protein sequences, they are firstly embedded by n-gram coding, and fed into BiLSTM to obtain their representations. Finally, the drug representation and the protein representation are concatenated and fed into the fully connected layer to predict the binding affinity.

Datasets

The Davis and KIBA datasets were used to evaluate the performance of the proposed model. The Davis dataset was obtained by selecting certain kinase proteins and their corresponding inhibitors, with binding affinity represented by the dissociation constant K_d, and affinity was processed using Eq. 1. It contains 442 proteins, 68 drugs, and 30,056 drug–target interactions. The average length of the drug SMILES strings is 64, and the average length of the protein sequences is 788. The KIBA dataset includes combined kinase inhibitor biological activities from various sources, such as inhibition constant (K_i), dissociation constant (K_d), or the half-maximal inhibitory concentration (IC_{50}), and predicts biological activity using the KIBA score. It consists of 229

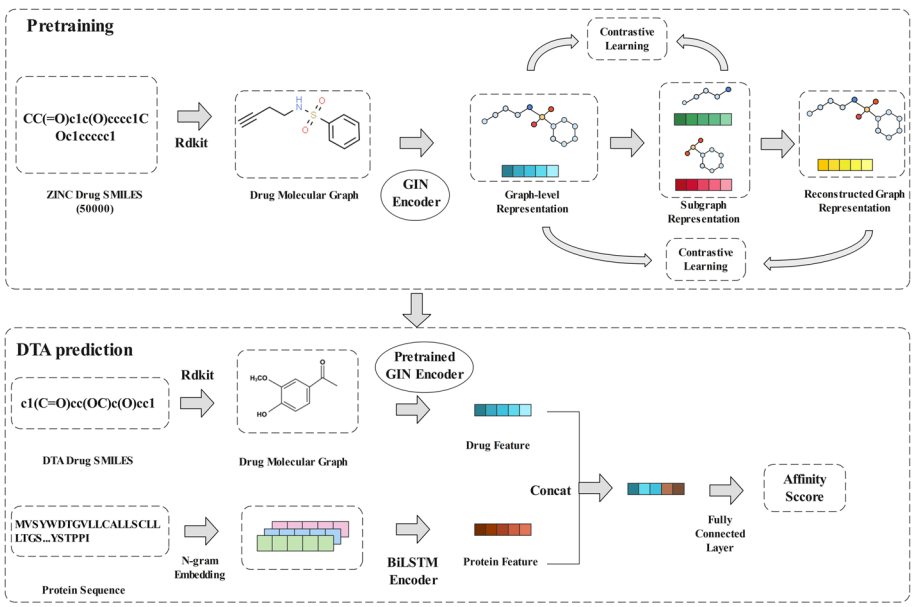


Fig. 1 Overview of SubMDTA

proteins, 2111 drugs, and 118,254 drug–target interactions. The average length of the drug SMILES strings is 58, and the average length of the protein sequences is 728, the detail information of these two datasets is shown in Table 1.

$$pK_d = -\log_{10} \frac{K_d}{10^9} \tag{1}$$

Molecular encoder

For each drug molecule in the experimental dataset, it is represented by its corresponding SMILES. The open source cheminformatics software RDKit [34] is used to convert SMILES string into its corresponding molecular graph. For the node features, we use a set of atomic feature representations adopted from DeepChem [35]. In order to better explore the features of molecule, Graph Isomorphism Network (GIN) is used as the graph encoder in this paper. GIN provides better inductive bias for graph representation learning, which generates node representations by repeatedly aggregating information from the local neighborhood nodes. After each GIN layer, there is a batch normalization layer activated by the ReLU function. Specifically, GIN uses a MLP model to update the node features and its update process can be written as:

$$\mathbf{x}_i^{l+1} = MLP\left((1+\varepsilon)\mathbf{x}_i^l + \sum_{j \in \mathcal{N}_i} \mathbf{x}_j^l\right) \tag{2}$$

where ε is either a learnable parameter or fixed scalar, $\mathbf{x}_i^l$ denotes the node feature of the i-th node in the l-th layer, $\mathcal{N}_i$ are neighborhoods to node i, and $\mathbf{x}_j^l$ denotes the node features of the j-th node in the l-th layer.

Multiple GIN layers could aggregate information of node from its multi-hop neighbors, and the information embedded in the representations of different hops will gradually change from local information to global information. After L layers of GIN, a list of node representations $\left\{\mathbf{x}_i^0, \mathbf{x}_i^1, \ldots, \mathbf{x}_i^L\right\}$ is generated. To avoid loss of node information, a convolution kernel of size (L, 1) called **Conv** is used to aggregate node representations at different layers as Eq. 3, thus local and global information can be combined.

$$x_i^G = \mathbf{Conv}\left(\left[\mathbf{x}_i^0, \mathbf{x}_i^1, \ldots, \mathbf{x}_i^L\right]\right) \tag{3}$$

After obtaining the final node embeddings containing information at different levels of the graph, the obtained embeddings are aggregated into fixed-length graph-level representations using a read-out function. In this paper, we use a global summation pooling function which we called as **GlobalAddPool** to read out the representation $h(G)$ of the nodes as Eq. 4:

Table 1 Summary of the benchmark datasets

Dataset	Proteins	Compounds	Interactions	Train	Test
Davis	442	68	30,056	25,046	5010
KIBA	229	2111	118,254	98,545	19,709

$$h(G) = \textbf{GlobalAddPool}\left(X^G\right) \tag{4}$$

where the X^G represent the node feature matrix. It returns batched graph-level output by aggregating node features across the node dimension, thus ensuring that the global representation of graph is more comprehensive.

Contrastive learning method for molecular representation

Inspired by the mutual information-based contrastive learning algorithm [36, 37], maximizing the mutual information of molecular graphs can obtain more feature representations. The overall framework of our approach is shown in Fig. 2. The drug molecule graph acquires the original features after GIN encoding, followed by substructure extraction. For the original graph, its original feature is selected to form a positive sample pair with each subgraph representations, and the subgraphs of other graphs in the same batch form negative sample pairs. In order to capture the inherent relations between graphs, subgraphs are ranked according to similarity and half of them are replaced to obtain the reconstructed graph. The original graph with its reconstructed graph constitutes a positive sample, and with the reconstructed graph within the same batch constitutes a negative sample.

Subgraph-level contrastive learning

In this paper, a subgraph's generation method [29] is utilized in contrastive learning. After obtaining the node feature matrix X^G, it is transformed by linearly function with the learnable matrix W and the row-by-row **Softmax** function is used to obtain a probability matrix A as Eq. 5. A_{ij} denotes the probability of the i-th node in the j-th subgraph. The **Softmax** function exponentiates the input vectors and sums them to obtain a scalar. The exponent value of each element is then divided by this scalar to obtain the normalized probability value.

$$A = \textbf{Softmax}\left(X^G \cdot W\right) \tag{5}$$

Based on the probability matrix A, we can divide the original graph into two subgraphs by a pre-defined probability 0.5. After T rounds of splitting, we obtain $S = 2^T$ subgraphs.

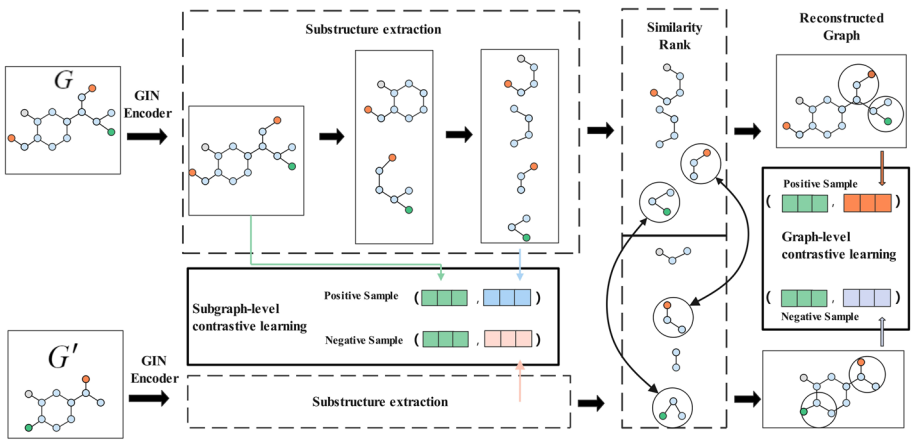


Fig. 2 Contrastive learning method for molecular representation

The node representations of each subgraph are denoted as X^{G_i}, $i = 1, 2, \ldots, S$. Here, we adopt the same pooling function as Eq. 4 to obtain the graph-level representation. After the reading out function, we can obtain the subgraph representation $h(G_i)$ as Eq. 6:

$$h(G_i) = \textbf{GlobalAddPool}\left(X^{G_i}\right), i = 1, 2, \ldots, S \tag{6}$$

Mutual information (MI) is an indicator to quantify the relationship between two random variables. Let ϕ represent the parameters of the graph neural network, and a discriminator $T_\omega : h_\phi(G) \times h_\phi(G_i)$ which takes as input a subgraph/graph embedding pair and determines whether they come from the same graph is used:

$$\hat{\phi}, \hat{\omega} = \arg\max_{\phi,\omega} \sum_{G \in \mathbf{G}} \frac{1}{|\mathbf{G}|} I_{\phi,\omega}\left(h_\phi(G); h_\phi(G_i)\right) \tag{7}$$

where $I_{\phi,\omega}\left(h_\phi(G); h_\phi(G_i)\right)$ is a mutual information estimator modeled by the discriminator and parameterized by the neural network.

We use the Jensen–Shannon (JS) mutual information estimator [38] on local/global pairs to maximize the mutual information on a given subgraph/graph embedding as Eq. 8. The JS mutual information estimator is approximately monotonic with respect to the KL scatter (the traditional definition of mutual information), but it is more stable and can provide better results [39].

$$\begin{aligned} &I^{JSD}_{\phi,\omega}\left(h_\phi(G); h_\phi(G_i)\right) = \\ &\mathbf{E}_P(-sp(-T_\omega(h_\phi(G), h_\phi(G_i)))) - \mathbf{E}_P(sp(T_\omega(h_\phi(G), h_\phi(G_i'))))\end{aligned} \tag{8}$$

where $P = p\left(h_\phi(G), h_\phi(G_i)\right)$ is the joint distribution of the global graph representation and the subgraph representation, and $Q = p\left(h_\phi(G)\right)p\left(h_\phi(G_i)\right)$ denotes the product of marginal distributions of two embeddings. In contrastive learning, Q denotes the distribution of positive pairs, P denotes the distribution of negative pairs, and $sp(x) = log\left(1 + e^x\right)$ is the softplus function.

Graph-level contrastive learning

The reconstructed graph generation method is based on the strategy of similar subgraph substitution. To better capture the structural information of the graph, given the generated subgraph of a certain original graph G_i, we compute its cosine similarity to the generated subgraphs of other original graphs G_i in the same batch as Eq. 9:

$$similarity = \cos(\theta) = \frac{h(G_i) \cdot h(G_i')}{\left\|h(G_i)\right\| \left\|h(G_i')\right\|} \tag{9}$$

After ranking, half of the original subgraphs are replaced according to the similarity values, and finally aggregated and assembled into a reconstructed graph using a convolution kernel of size $(S, 1)$.

For the reconstructed representation $h(\hat{G})$, the global feature $h(G)$ of its original graph is selected to form a positive sample pair, and the negative sample pair constitute the reconstructed graph $h(\hat{G}')$ of other graphs in the same batch. We use the same mutual

information calculation method to maximize the mutual information between the original and reconstructed graphs, denoted as $I^{JSD}_{\phi,\omega}(h_\phi(G); h_\phi(\hat{G}))$ as Eqs. 10 and 11:

$$\hat{\phi}, \hat{\omega} = \arg\max_{\phi,\omega} \sum_{G \in \mathbf{G}} \frac{1}{|\mathbf{G}|} I_{\phi,\omega}(h_\phi(G); h_\phi(\hat{G})) \tag{10}$$

$$\begin{aligned} &I^{JSD}_{\phi,\omega}(h_\phi(G); h_\phi(\hat{G})) = \\ &\mathbf{E}_P(-sp(-T_\omega(h_\phi(G), h_\phi(\hat{G}))))-\mathbf{E}_P(sp(T_\omega(h_\phi(G), h_\phi(\hat{G}'))))\end{aligned} \tag{11}$$

The final loss is the sum of two mutual information losses:

$$Loss_{\phi,\omega} = I^{JSD}_{\phi,\omega}\left(h_\phi(G); h_\phi(G_i)\right) + I^{JSD}_{\phi,\omega}(h_\phi(G); h_\phi(\hat{G})) \tag{12}$$

To enhance the generalization of the self-supervised learning features, 50,000 molecules are randomly selected from the ZINC database for pre-training the self-supervised model, and a high-quality molecular encoder is obtained from learning rich molecular structure and semantic information in unlabeled molecular data.

Protein representation

For each protein in the experimental dataset, the protein sequence is obtained from the UniProt database through its gene name. The sequence is a string of ASCII characters representing amino acids. The n-gram [40] is used to define the "words" in the amino acid sequence, and the protein sequence is split into multiple overlapping n-gram amino acid word. Depending on the permutations and combinations, there are 22^n n-gram words. However, if the n-gram syntax number is too large, the word frequency may be too low. Taking n = 3 as an example, given a protein sequence $S = s_1 s_2 s_3 \ldots s_{|s|}$, $|S|$ represents the length of protein sequence, we divide it into n-gram words:

$$[s_1; s_2; s_3], [s_2; s_3; s_4], \ldots, \left[s_{|S|-2}; s_{|S|-1}; s_{|S|}\right] \tag{13}$$

We use the symbol $s_{i:i+2}$ to represent the protein word $[s_i; s_{i+1}; s_{i+2}]$, and then encode the word using the Eq. 14.

$$c_i = \mathbf{Embedding}(s_{i:i+2}) \tag{14}$$

where the **Embedding** function initializes the weight from the standard normal distribution according to the input vocabulary size and embedding dimension, and outputs the word vector corresponding to the vocabulary index in the weight.

In this work, inspired by MGraphDTA [21], we set n = 2, 3, 4 to encode protein respectively in order to detect the local residue patterns of proteins at different scales. Finally we get three types of embedding c_i^2, c_i^3, c_i^4. For protein sequence, sequence-based models are the optimal choice for feature extraction. Long short-term memory network (LSTM) [41] is a DL model to overcome the gradient disappearance problem to process sequence data. The main idea is to introduce an adaptive gating mechanism which determines the extent to which the LSTM unit maintains its previous state and remembers the extracted features of the current data input.

Bi-directional LSTM (BiLSTM) [42] is a variant of LSTM that combines the outputs of two LSTMs, one processing sequences from left to right and the other from right to left, to capture long-term dependencies and contextual relationships. Since each amino acid residue in the sequence information of the protein has interrelationship with residues in the both directions, the BiLSTM is more suitable to process protein sequence, which is defined as Eq. 15:

$$\begin{aligned} \overrightarrow{h_i} &= \overrightarrow{LSTM}(c_i, h_{i-1}) \\ \overleftarrow{h_i} &= \overleftarrow{LSTM}(c_i, h_{i+1}) \\ h_i &= \overrightarrow{h_i} \parallel \overleftarrow{h_i} \end{aligned} \tag{15}$$

where $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ denote the hidden states of the time step computed from left-to-right and right-to-left, respectively, and h_i denotes the global representation of the t-th time step stitched together by them.

The word vector c_i^2, c_i^3, c_i^4 are fed into the BiLSTM layer to capture the dependencies between characters in the sequence. After the max-pooling layer, the three features are concatenated together to obtain the final protein representation. The BiLSTM framework is shown in Fig. 3.

DTA prediction

In this paper, we treat the drug–target binding affinity prediction task as a regression task. With the representation learned from the previous sections, we can integrate all the information from the drug and target to predict the DTA value. As shown in Fig. 4, drug representation and protein representation are concatenated together, which is fed into two dense fully connected layers to predict the DTA value. Besides, the ReLU is used as the activation function for increasing the nonlinear relationship. Given the set of drug–target pairs and the ground-truth labels, we use the mean squared error (MSE) as the loss function.

Results and discussion

Metrics

The DTA prediction is regarded as a regression problem and our model was evaluated using three metrics including mean squared error (MSE), concordance index (CI), and regression toward the mean (r_m^2 index). MSE calculates difference between the predicted and actual values through the function of squared loss as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \tag{16}$$

where $\hat{y}_i$ is the predicted value, y_i is the true value, and n is the number of drug–target pairs. CI is used to measure whether the predicted DTA values of two random drug–target pairs are predicted in the same order as their true values:

$$CI = \frac{1}{Z}\sum_{d_x > d_y} h\left(b_x - b_y\right) \tag{17}$$

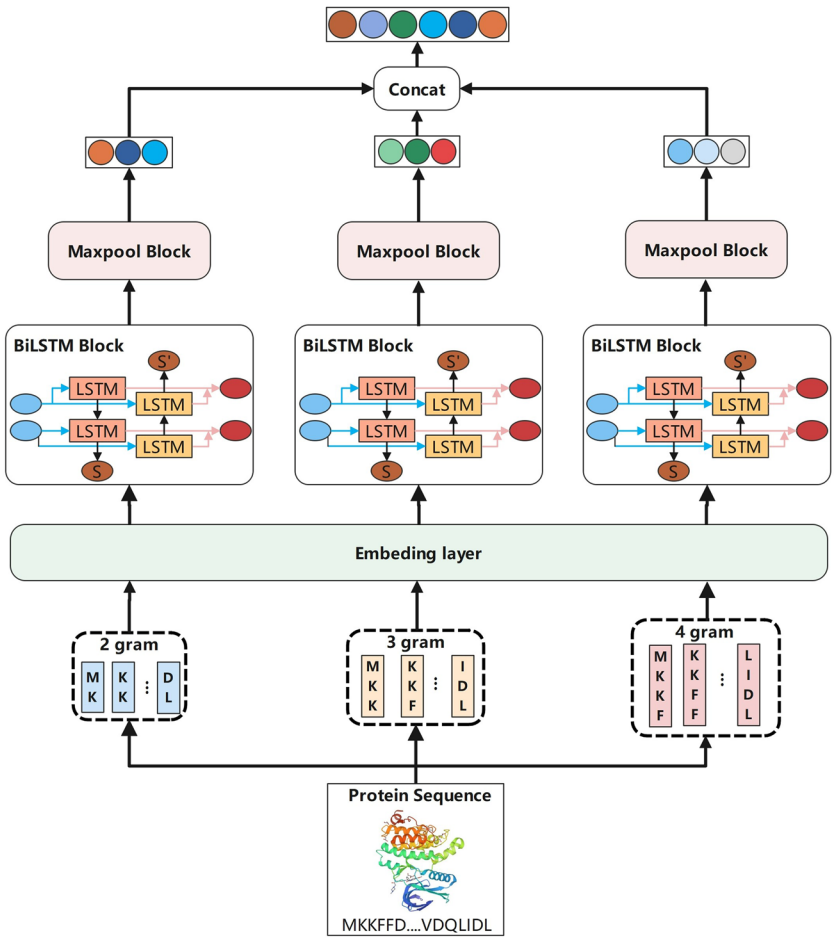


Fig. 3 Multi-scale protein representation method

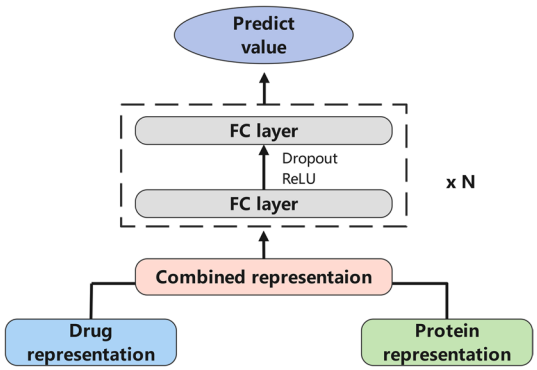


Fig. 4 The prediction part of the model

$$h(x) = \begin{cases} 1, & if\ x > 0 \\ 0.5, & if\ x = 0 \\ 0, & if\ x < 0 \end{cases} \tag{18}$$

where b_x is the predicted value of the larger affinity d_x, b_y is the predicted value of the smaller affinity d_y, $h(x)$ is the step function. Z is the normalization constant which indicates the number of drug–target pairs.

r_m^2 is used to evaluate the external predictive performance of the model as follows:

$$r_m^2 = r^2 \times \left(1 - \sqrt{r^2 - r_0^2}\right) \tag{19}$$

where r^2 and r_0^2 are the squared correlation coefficients between the true and predicted values with and without intercepts, respectively.

Comparison with existing methods

To evaluate the performance of our model, we compared the model with other methods for DTA prediction, including KronRLS [11], SimBoost [12], DeepDTA [16], WideDTA [43], MATT-DTI [26], DeepGS [44], AttentionDTA [45], GraphDTA [18], and DeepGLSTM [46]. Table 2 shows the performance of different models based on MSE, CI, and r_m^2 metrics on the Davis dataset. On the Davis dataset, our method significantly outperformed the other methods in terms of MSE (0.218) and r_m^2 (0.719), which are 4.8% and 4.8% better than the previous optimal method, respectively. The CI of SubMDTA was very close to the best method DeepGLSTM by 0.001.

Moreover, we evaluated our model on KIBA dataset. As shown in Table 2, SubMDTA achieved the best performance among existing methods with MSE of 0.129, CI of 0.898, and r_m^2 of 0.793, where the MSE was 3% higher than the previous best method. The above results show that the proposed method can be considered as an accurate and effective tool for DTA prediction. Compared with other models, the superiority of our model can be summarized for two reasons: (i) to obtain more discriminative molecular representations, we utilized the local and global information of molecule through a pre-training task, which can focus on the structural features of molecular graph; (ii) compared with the conventional embedding method of protein sequence, our method used multiple

Table 2 Prediction performance on Davis dataset

Model	MSE	CI	r_m^2
KronRLS	0.379	0.871	0.407
SimBoost	0.282	0.872	0.644
DeepDTA	0.261	0.878	0.630
WideDTA	0.262	0.886	0.633
MATT_DTI	0.229	0.890	0.682
DeepGS	0.252	0.882	0.686
AttentionDTA	0.245	0.887	0.657
GraphDTA	0.229	0.893	0.649
DeepGLSTM	0.232	0.895	0.680
SubMDTA	0.218	0.894	0.719

n-gram sequence representations containing multi-level information. Thus, our model can integrate the intrinsic information of compounds and protein sequences into a more comprehensive representation, which is helpful to improve the accuracy and robustness of the model.

In addition, we evaluated our model on KIBA dataset. As shown in Table 3, SubMDTA achieved the best performance among existing methods with MSE of 0.129, CI of 0.898, and r_m^2 of 0.793, where the MSE was 3% higher than the previous best method.

The above results show that the proposed method can be considered as an accurate and effective tool for DTA prediction. Compared with other models, the superiority of our model can be summarized for two reasons: (i) to obtain more discriminative molecular representations, we utilized the local and global information of molecule through a pre-training task, which can focus on the structural features of molecular graph; (ii) compared with the conventional embedding method of protein sequence, our method used multiple n-gram sequence representations containing multi-level information. Thus, our model can integrate the intrinsic information of compounds and protein sequences into a more comprehensive representation, which is helpful to improve the accuracy and robustness of the model.

Comparison with different drug molecular representations

The complex structure of drug molecules is difficult to directly obtain its features, so special representation methods are required. We validated graph-based representation methods and molecular fingerprint methods. SubMDTA first converts the smiles string of the drug molecule into a molecular graph, and then uses one-hot encoding to obtain the features of the drug molecule according to the atomic attributes. Molecular fingerprint is a method of converting a molecular structure into a binary or sparse vector representation, where each bit or feature represents a specific substructure or chemical property of the molecule. In this section, the Morgan fingerprint [47] and the MACCS fingerprint [48] were used for comparison. SubMDTA, Morgan, and MACCS achieved MSE of 0.218, 0.221, and 0.222, respectively. It can be seen from Fig. 5 that SubMDTA finally obtained the best results among three, which may be related to the fact that the graph-based method can better capture the detailed structure of molecules.

Table 3 Prediction performance on KIBA dataset

Model	MSE	CI	r_m^2
KronRLS	0.411	0.782	0.342
SimBoost	0.222	0.836	0.629
DeepDTA	0.194	0.863	0.673
WideDTA	0.179	0.875	0.675
MATT_DTI	0.150	0.889	0.756
DeepGS	0.193	0.860	0.684
AttentionDTA	0.162	0.882	0.735
GraphDTA	0.147	0.889	0.674
DeepGLSTM	0.133	0.897	0.792
SubMDTA	0.129	0.898	0.793

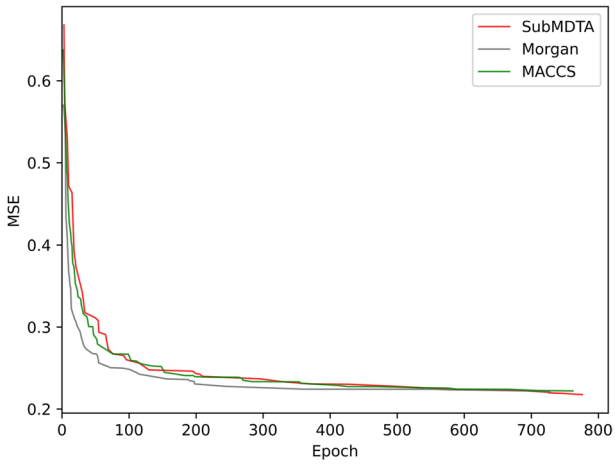
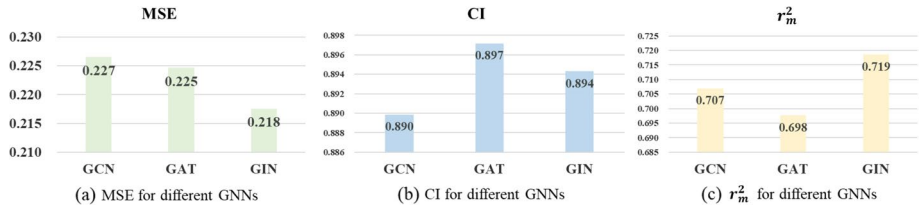


Fig. 5 Performances of different molecular representation methods



(a) MSE for different GNNs (b) CI for different GNNs (c) r_m^2 for different GNNs

Fig. 6 Performances of different GNNs

The construction of effective GNN networks for extracting discriminative features of drugs is essential to improve the prediction accuracy of DTA. Empirically, it is often difficult to obtain sufficient information from single-layer networks compared with multilayer networks, and too many layers may result in the problem of over-smoothing. Therefore, a four-layer GNN network was used in the proposed method. We tried three types of GNN architectures (GCN, GAT, and GIN) for performance comparison. It is obvious from the Fig. 6a and the Fig. 6c that the GIN model achieves an MSE of 0.218 and fran r_m^2 of 0.719, which is the best performance. As shown in Fig. 6b, the CI of the GAT model achieves 0.897, which is higher than 0.894 of GIN, but the difference is not obvious. This may be because that GIN can capture local features in the graph while retaining global information, thus improving its characterization ability.

Comparison with different protein representations

For protein feature representation, we propose a method based on n-gram multi-scale features fusion. Thus, we explored the effects of different protein sequence embedding methods, which are one-hot coding, 2-gram, 3-gram, 4-gram coding and n-gram fusion coding methods, and the experimental results are shown in Fig. 7. One-hot coding achieved an MSE of 0.234, CI of 0.892, and r_m^2 of 0.695. Compared with one-hot encoding, n-gram encoding provided better representations by capturing multiple characters in the sequence, and the MSE reached 0.226, 0.225, and 0.223 using 2-gram, 3-gram, and 4-gram, respectively.

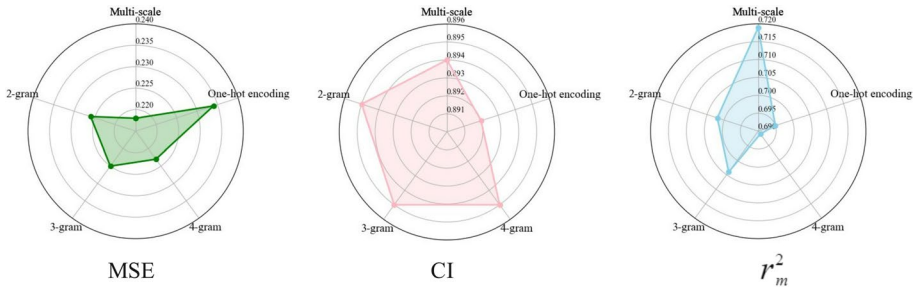


Fig. 7 Performances of different protein feature extraction methods

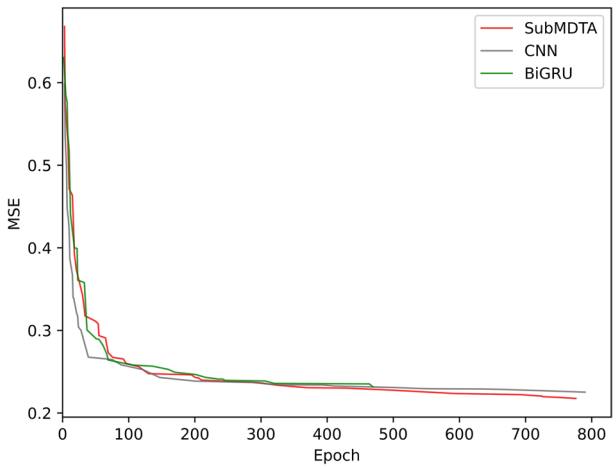


Fig. 8 Performances of different protein feature extraction methods

The performance of multi-scale representations was the best among them. This is because that the whole protein sequence contains many subsequences or structural domains, and the introduction of multi-scale features could capture more amino acid combinations and result in a better performance.

For protein feature extraction methods, we choose convolutional neural network (CNN) and bidirectional gated recurrent unit (BiGRU) as comparison methods. CNN extracts features from input data through convolution operations. BiGRU is a variant of recurrent neural network which consists of two GRUs for forward and backward processing. CNN and BiGRU achieved MSE of 0.225, and 0.232, respectively. SubMDTA achieved MSE of 0.219, which increased by 3.1% and 6.0%. As can be seen from Fig. 8, SubMDTA obtained the best MSE result, which proves the superiority of SubMDTA in processing protein sequence data.

Ablation study

To verify the effectiveness of the proposed model, we designed and conducted ablation experiments to determine the contributions of different factors of the model.

In the proposed model, maximizing the mutual information between the graph and the subgraph representations in the SSL task is helpful to preserve substructure information. In order to demonstrate the advantages of substructures, we designed three variants

Table 4 Performances of different molecular training tasks

Model	MSE	CI	r_m^2
SubMDTA-a(no pretraining)	0.224	0.894	0.714
SubMDTA-b (only subgraph-level pretraining)	0.225	0.897	0.703
SubMDTA-c (only graph-level pretraining)	0.230	0.895	0.697
SubMDTA (both subgraph-level and graph-level pretraining)	0.218	0.894	0.719

Table 5 Compound ranking based on the predicted affinities of SubMDTA

DrugBank ID	Rank
DB00678	21
DB01029	24
DB00275	28
DB00966	30
DB00177	42
DB00796	51
DB08822	74
DB00876	224
DB11842	1130

SubMDTA-a, SubMDTA-b and SubMDTA-c to evaluate the importance of the pre-training task module. As shown in Table 4, SubMDTA-a obtained an MSE of 0.224. The introduction of contrastive learning improved the MSE to 0.225 and 0.230 by SubMDTA-b and SubMDTA-c, respectively. The MSE of SubMDTA which combined these two methods reached 0.218. This may be related to the fact that using one type of mutual information alone cannot obtain the comprehensive features. Meanwhile, maximizing the mutual information between the graph representation and the reconstructed graph representation can enable the embedding to focus on the global features of the graph.

Case study

In order to verify the robustness of proposed method, we applied approved drugs targeting the Type-1 angiotensin II receptor in DrugBank for a case study. According to similar steps to MSF-DTA [49], after training SubMDTA on the Davis dataset, we predicted the affinities between the receptor and 1781 available small molecule drugs. Among them, 9 out of 1781 drugs are known to bind this receptor. To ensure a fair comparison, this receptor never appeared in the Davis dataset. The predicted affinities between the nine drugs and the receptor are listed in descending order, as shown in Table 5. It can be seen that according to the sorting results of SubMDTA, 8 drugs are ranked in the top 13 % of 1781 drugs, and 7 drugs appear in the top 4 %. These results suggest that SubMDTA can identify novel target-protein interacting drugs well and has the potential to be developed as a predictive tool.

Conclusion

In this paper, we present a new model SubMDTA using self-supervised learning and multi-scale features for DTA prediction. The drug representations are extracted by contrastive learning methods between graph-level and subgraph representations and between graph-level and reconstructed graph representations, which is refined by downstream task. In addition, multi-scale sequence features were fused to learn protein representations, which captured long distance and multiple relationships in amino acid sequences. The experimental results proved that our method outperformed existing methods. In our future work, we will take account into the progresses made in heterogeneous information networks [50] and incorporate them to enhance the prediction ability of our models.

Abbreviations

DTA	Drug–target affinity
DL	Deep learning
SMILES	Simplified molecular input line entry system
BiLSTM	Bi-directional long short-term memory
LSTM	Long short-term memory
GCN	Graph convolutional networks
GAT	Graph attention networks
GIN	Graph isomorphism networks
SSL	Self-supervised learning

Acknowledgements

Not applicable.

Author contributions

Conceptualization, SP and ZL; methodology, SP; validation, LX; dataset, LX; writing-original draft preparation, SP; writing-review and editing, ZL All authors read and approved the final manuscript.

Funding

This work has been supported by Shandong Key Science and Technology Innovation Project [2021CXGC011003] and Qingdao Key Technology Research and Industrialization Projects[22-3-2-qljh-8-gx]

Availability of data and materials

The code and data are provided at https://github.com/1q84er/SubMDTA

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 11 May 2023 Accepted: 31 August 2023
Published online: 07 September 2023

References

1. Vermaas JV, Sedova A, Baker MB, Boehm S, Rogers DM, Larkin J, Glaser J, Smith MD, Hernandez O, Smith JC. Supercomputing pipelines search for therapeutics against covid-19. Comput Sci Eng. 2020;23(1):7–16.
2. Lin X, Li X, Lin X. A review on applications of computational methods in drug screening and design. Molecules. 2020;25(6):1375.
3. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. Autodock4 and autodocktools4: automated docking with selective receptor flexibility. J Comput Chem. 2009;30(16):2785–91.
4. Trott O, Olson AJ. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455–61.

5. John S, Thangapandian S, Sakkiah S, Lee KW. Potent bace-1 inhibitor design using pharmacophore modeling, in silico screening and molecular docking studies. BMC Bioinform. 2011;12(1):1–11.
6. Schuffenhauer A, Floersheim P, Acklin P, Jacoby E. Similarity metrics for ligands reflecting the similarity of the target proteins. J Chem Inf Comput Sci. 2003;43(2):391–405.
7. Klabunde T. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. Br J Pharmacol. 2007;152(1):5–7.
8. Shaik NA, Hakeem KR, Banaganapalli B, Elango R. Essentials of bioinformatics, vol. i. Cham: Springer International Publishing; 2019.
9. Yang C, Chen EA, Zhang Y. Protein-ligand docking in the machine-learning era. Molecules. 2022;27(14):4568.
10. Wang F, Liu D, Wang H, Luo C, Zheng M, Liu H, Zhu W, Luo X, Zhang J, Jiang H. Computational screening for active compounds targeting protein sequences: methodology and experimental validation. J Chem Inf Model. 2011;51(11):2821–8.
11. Pahikkala T, Airola A, Pietilä S, Shakyawar S, Szwajda A, Tang J, Aittokallio T. Toward more realistic drug–target interaction predictions. Brief Bioinform. 2015;16(2):325–37.
12. He T, Heidemeyer M, Ban F, Cherkasov A, Ester M. Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. J Cheminform. 2017;9(1):1–14.
13. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. Pubchem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. 2009;37(suppl_2):623–33.
14. Smith TF, Waterman MS, et al. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7.
15. Wu Y, Gao M, Zeng M, Zhang J, Li M. Bridgedpi: a novel graph neural network for predicting drug-protein interactions. Bioinformatics. 2022;38(9):2571–8.
16. Öztürk H, Özgür A, Ozkirimli E. Deepdta: deep drug–target binding affinity prediction. Bioinformatics. 2018;34(17):821–9.
17. Abbasi K, Razzaghi P, Poso A, Amanlou M, Ghasemi JB, Masoudi-Nejad A. Deepcda: deep cross-domain compound-protein affinity prediction through LSTM and convolutional neural networks. Bioinformatics. 2020;36(17):4633–42.
18. Yuan W, Chen G, Chen CYC. Fusiondta attention-based feature polymerizer and knowledge distillation for drug–target binding affinity prediction. Brief Bioinform. 2022;23(1):506.
19. Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. Graphdta: predicting drug–target binding affinity with graph neural networks. Bioinformatics. 2021;37(8):1140–7.
20. 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(35):20701–12.
21. Yang Z, Zhong W, Zhao L, Chen CY-C. Mgraphdta: deep multiscale graph neural network for explainable drug–target binding affinity prediction. Chem Sci. 2022;13(3):816–33.
22. Lin S, Shi C, Chen J. Generalizeddta: combining pre-training and multi-task learning to predict drug–target binding affinity for unknown drug discovery. BMC Bioinform. 2022;23(1):1–17.
23. Li Z, Jiang M, Wang S, Zhang S. Deep learning methods for molecular representation and property prediction. Drug Discov Today. 2022. https://doi.org/10.1016/j.drudis.2022.103373.
24. Sun F-Y, Hoffmann J, Verma V, Tang J. Infograph: unsupervised and semi-supervised graph-level representation learning via mutual information maximization. arXiv preprint arXiv:1908.01000 2019.
25. Li P, Wang J, Qiao Y, Chen H, Yu Y, Yao X, Gao P, Xie G, Song S. An effective self-supervised framework for learning expressive molecular global representations to drug discovery. Brief Bioinform. 2021;22(6):109.
26. Rong Y, Bian Y, Xu T, Xie W, Wei Y, Huang W, Huang J. Self-supervised graph transformer on large-scale molecular data. Adv Neural Inf Process Syst. 2020;33:12559–71.
27. Zeng Y, Chen X, Luo Y, Li X, Peng D. Deep drug–target binding affinity prediction with multiple attention blocks. Brief Bioinform. 2021;22(5):117.
28. Chen L, Tan X, Wang D, Zhong F, Liu X, Yang T, Luo X, Chen K, Jiang H, Zheng M. Transformercpi: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics. 2020;36(16):4406–14.
29. Wang C, Liu Z. Learning graph representation by aggregating subgraphs via mutual information maximization. arXiv preprint arXiv:2103.13125 2021.
30. 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(11):1046–51.
31. Tang J, Szwajda A, Shakyawar S, Xu T, Hintsanen P, Wennerberg K, Aittokallio T. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J Chem Inf Model. 2014;54(3):735–43.
32. Weininger D. Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28(1):31–6.
33. Xu K, Hu W, Leskovec J, Jegelka S. How powerful are graph neural networks? arXiv preprint arXiv:1810.00826 2018.
34. Bento AP, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, Bellis LJ, De Veij M, Leach AR. An open source chemical structure curation pipeline using RDKit. J Cheminform. 2020;12:1–16.
35. Ramsundar B, Eastman P, Walters P, Pande V. Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more. O'Reilly Media; 2019.
36. Velickovic P, Fedus W, Hamilton WL, Liò P, Bengio Y, Hjelm RD. Deep graph infomax ICLR (Poster). 2019;2(3):4.
37. Park C, Han J, Yu H. Deep multiplex graph infomax: attentive multiplex network embedding using global information. Knowl-Based Syst. 2020;197:105861.
38. Nowozin S, Cseke B, Tomioka R. f-gan: Training generative neural samplers using variational divergence minimization. Adv Neural Inf Process Syst 2016;29.
39. Hjelm RD, Fedorov A, Lavoie-Marchildon S, Grewal K, Bachman P, Trischler A, Bengio Y. Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670 2018.
40. Dong Q-W, Wang X-L, Lin L. Application of latent semantic analysis to protein remote homology detection. Bioinformatics. 2006;22(3):285–90.
41. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.

42. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans Sign Process. 1997;45(11):2673–81.
43. Öztürk H, Ozkirimli E, Özgür A. Widedta: prediction of drug–target binding affinity. arXiv preprint arXiv:1902.04166 2019.
44. Lin X. Deepgs: Deep representation learning of graphs and sequences for drug–target binding affinity prediction. arXiv preprint arXiv:2003.13902 2020.
45. Zhao Q, Xiao F, Yang M, Li Y, Wang J. Attentiondta: prediction of drug–target binding affinity using attention model. In: 2019 IEEE international conference on bioinformatics and biomedicine (BIBM), 2019; IEEE, pp. 64–69.
46. Mukherjee S, Ghosh M, Basuchowdhuri P. Deepglstm: deep graph convolutional network and lstm based approach for predicting drug–target binding affinity. In: Proceedings of the 2022 SIAM international conference on data mining (SDM), 2022; SIAM, 729–737.
47. Zhao B-W, You Z-H, Hu L, Guo Z-H, Wang L, Chen Z-H, Wong L. A novel method to predict drug–target interactions based on large-scale graph representation learning. Cancers. 2021;13(9):2111.
48. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of mdl keys for use in drug discovery. J Chem Inf Comput Sci. 2002;42(6):1273–80.
49. Ma W, Zhang S, Li Z, Jiang M, Wang S, Guo N, Li Y, Bi X, Jiang H, Wei Z. Predicting drug–target affinity by learning protein knowledge from biological networks. IEEE J Biomed Health Inform. 2023;27(4):2128–37.
50. Zhao B-W, Wang L, Hu P-W, Wong L, Su X-R, Wang B-Q, You Z-H, Hu L. Fusing higher and lower-order biological information for drug repositioning via graph representation learning. IEEE Trans Emerg Topics Comput. 2023. https://doi.org/10.1109/TETC.2023.3239949.

Publisher's Note