Computers in Biology and Medicine 150 (2022) 106145

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

GSAML-DTA: An interpretable drug-target binding affinity prediction model based on graph neural networks with self-attention mechanism and mutual information

Jiaqi Liao [a], Haoyang Chen [a], Lesong Wei [b,**], Leyi Wei [a,*]

[a] School of Software, Shandong University, Jinan, China
[b] Department of Computer Science, University of Tsukuba, Tsukuba, 3058577, Japan

ARTICLE INFO

Keywords:
Drug-target affinity
Graph neural networks
Self-attention mechanism
Mutual information

ABSTRACT

Identifying drug-target affinity (DTA) has great practical importance in the process of designing efficacious drugs for known diseases. Recently, numerous deep learning-based computational methods have been developed to predict drug-target affinity and achieved impressive performance. However, most of them construct the molecule (drug or target) encoder without considering the weights of features of each node (atom or residue). Besides, they generally combine drug and target representations directly, which may contain irrelevant-task information. In this study, we develop GSAML-DTA, an interpretable deep learning framework for DTA prediction. GSAML-DTA integrates a self-attention mechanism and graph neural networks (GNNs) to build representations of drugs and target proteins from the structural information. In addition, mutual information is introduced to filter out redundant information and retain relevant information in the combined representations of drugs and targets. Extensive experimental results demonstrate that GSAML-DTA outperforms state-of-the-art methods for DTA prediction on two benchmark datasets. Furthermore, GSAML-DTA has the interpretation ability to analyze binding atoms and residues, which may be conducive to chemical biology studies from data. Overall, GSAML-DTA can serve as a powerful and interpretable tool suitable for DTA modelling.

1. Introduction

Developing a new drug generally takes more than ten years and costs billions of dollars, and less than 12% of the drugs are approved to enter the market [1,2]. The accuracy assessment of drug-target interaction is a crucial step in the early stage of drug development and uncovering their side effects [3]. Binding affinity is the strength of drug-target interaction, which is usually expressed in different metrics such as inhibition constant (K_i), dissociation constant (K_d), or the half-maximal inhibitory concentration (IC_{50}) [4]. Although wet lab experiments to identify the drug-target binding affinity remain the most reliable and effective methods, they are time-consuming and resource-intensive. To mitigate this issue, numerous computational methods have been proposed to accelerate the speed of new drug development and reduce the cost [5].

The existing computational methods mainly fall into two categories: structure-based methods and structure-free methods. Structure-based methods mainly exploit three-dimensional (3D) structure information of small molecules and proteins to explore potential binding poses at the atom level and identify binding affinities. Molecular docking is one of the well-established structure-based methods that integrate various potential binding poses and scoring functions to minimize the free energy of the pose within binding sites [6,7]. Although these methods have achieved relatively attractive predictive performance and provided reasonable biological interpretation, their coverage is limiteddue to the high computational complexity of solving such 3D structures and the scarcity of small molecules and proteins with known 3D structures.

An alternative to structure-based methods is structure-free methods, including feature-based methods and deep learning methods, which only rely on sequence information and require fewer computational resources. Feature-based methods mainly explore primary sequence information to model the binding affinity. Concretely, they focus on extracting discriminative biological features of a drug-target pair and sending extracted features into a machine/deep learning model, such as Naïve Bayes (NB), logistic regression (LR), deep neural network (DNN),

* Corresponding author.
** Corresponding author.
E-mail addresses: wei.lesong.ww@alumni.tsukuba.ac.jp (L. Wei), weileyi@sdu.edu.cn (L. Wei).

https://doi.org/10.1016/j.compbiomed.2022.106145
Received 10 August 2022; Received in revised form 23 August 2022; Accepted 24 September 2022
Available online 4 October 2022
0010-4825/

and other kernel-based methods, for predicting the binding affinity. For example, Lenselink et al. created and benchmarked a standardized dataset. Based on this dataset, they compared DNN with various traditional classifiers (*e.g.,* NB and LR). It was shown that DNN produced the best results [8]. Rifaioglu et al. integrated multiple protein features, including physicochemical properties and sequential, structural, and evolutionary features, into numerous 2D vectors. They then fed the vectors to state-of-the-art pairwise input hybrid deep neural networks to predict the drug-target interactions [9–11].

Although feature-based methods have a high generalization and sequence sensitivity, they are limited by over-relying on expert knowledge-based hand-crafted feature engineering. Deep learning methods, that is, end-to-end differential models can potentially tackle the above limitations. Indeed, they can automatically learn features and invariances of given data and provide a satisfactory generalization despite a large number of parameters. Inspired by their successful application in various research fields [12,13], numerical deep learning methods are proposed for DTA prediction. For example, Öztürk et al. constructed a deep learning model DeepDTA that employed convolutional neural networks (CNNs) to extract high-latent features of drugs and proteins separately and concatenated the two learned features for final prediction through fully connected layers [14]. Moreover, they proposed another DTA model, WideDTA, which integrated different text-based information to better represent the interaction [15]. DeepCDA [16] proposed a bidirectional attention mechanism to encode the binding strength between each protein substructure-composite substructure pair. And then, a combination of CNN and Long Short Term Memory (LSTM) was built to get good representations of proteins and compounds.

Although CNN-based models have shown satisfactory performance in DTA prediction, these models ignore the structural information. They only use sequences (1-dimensional structure) to represent the input molecules, which may miss the critical spatial information to characterize the intrinsic properties of molecules. To solve this problem, graph neural networks (GNNs), which can extract structural features, are widely used in various DTA prediction models [17–22]. For example, DeepGS [23] first proposed a method to learn the interaction between drugs and targets through the local chemical context and topology structure and then extensive experiments on both large and small benchmark datasets demonstrated the competitiveness and superiority of the proposed DeepGS. GraphDTA [19] represented drug features as graphs and adopted some GNNs, like Graph Convolutional Network (GCN), Graph Attention Network (GAT), and Graph Isomorphic Network (GIN), to extract drug features. The results confirm that deep learning models are beneficial for drug-target binding affinity prediction and representing drugs as graphs is beneficial for model performance improvement. Jiang et al. represented compounds as molecular graphs, utilized contact maps to gain protein graphs through protein sequences, and then built GNN networks to obtain feature representation. The experimental results show that representing proteins through contact maps can improve the prediction performance of the model [24].

Above all, most of the existing deep learning methods fail to consider the contribution of each drug atom and protein residue to the binding affinity and ignore the information hidden in different layers, which will lead to partial information loss during the feature learning process and cause poor prediction performance. Moreover, when concatenating the learned features of drugs and proteins directly, it may introduce much task-irrelevant information without further optimization. To overcome the above limitations, here we propose GSAML-DTA, an interpretable deep learning framework for predicting drug-target binding affinity. First, we construct drug graphs and protein graphs from drug SMILES (Simplified Molecular Input Line Entry System) strings and protein contact maps, respectively. Next, a hybrid network GAT-GCN with a self-attention mechanism is designed to extract layer-wise structural information from drug and protein graphs. The extracted layer-wise features of the drug and target are fused separately, and then fused features are concatenated to obtain a combined representation of a drug-target pair. Finally, the mutual information principle is applied to the combined representation, and the output is fed into fully connected layers to predict binding affinity. Through comprehensive evaluation on two benchmark datasets, we demonstrate that GSAML-DTA outperforms state-of-the-art methods. Additionally, our model can be employed to identify the important binding atoms and residues that contribute most to DTA prediction, thus providing biological interpretability.

2. Materials and methods

2.1. Datasets

To perform head-to-head comparisons of GSAML-DTA to existing machine/deep learning-based methods, we evaluate our model on two publicly available DTA datasets, Davis dataset [25] and KIBA dataset [26]. The Davis dataset consists of 442 proteins and 68 compounds forming 30056 drug-target pairs, in which the binding affinity is measured by kinase dissociation constant (K_d) values. The higher value of K_d represent lower binding strength of a drug-target pair. These data are selected from the kinase protein family. Following the previous study [14,15,27], the K_d value is transformed into log space as follows:

$$pK_d = -\log_{10}\frac{K_d}{10^9}. \quad (1)$$

The KIBA dataset adopts KIBA scores as drug-target binding affinities, which are obtained by integrating kinase inhibitor bioactivities from different sources such as K_i, K_d, and IC_{50} [4]. The higher binding strength between the drug and target corresponds to a lower KIBA score. It contains 229 proteins and 2111 compounds forming 118254 drug-target pairs. In these two datasets, the drug SMILES strings are collected from the PubChem compound database [28] and protein sequences are collected from the UniProt protein database [29]. The statistics of these two datasets are provided in Table 1. Note that due to the computer memory limitation, a long protein sequence and its related pairs are deleted from the KIBA dataset. Similar to DGraphDTA [24], each dataset is randomly divided into two parts (training set and testing set) with the ratio 5:1 for model performance evaluation. For the fair comparison, 5-fold cross validation is applied to training set and the average score is reported as the final performance.

2.2. Model architecture

In this section, we introduce the details of GSAML-DTA. The overall architecture of GSAML-DTA is shown in Fig. 1. In the first step, the drug SMILES strings are transformed into molecular graphs, and simultaneously the protein graphs are constructed based on contact maps, which are predicted from raw protein sequences. Then, a hybrid network GAT-GCN with a self-attention mechanism is utilized to extract the latent features from molecular graphs and protein graphs, respectively. Subsequently, we concatenate the features of the drugs and proteins and introduce mutual information to optimize the combined features for obtaining more comprehensive and effective feature representations. Finally, the optimized representation is fed into fully connected layers to predict the binding affinity.

2.3. Graph representation of drug molecules

In the datasets, each sample consists of a drug molecule and a target

Table 1
Statistics of the two datasets.

Number	Dataset	Proteins	Compounds	Binding entities
1	Davis	442	68	30056
2	KIBA	229	2111	118254

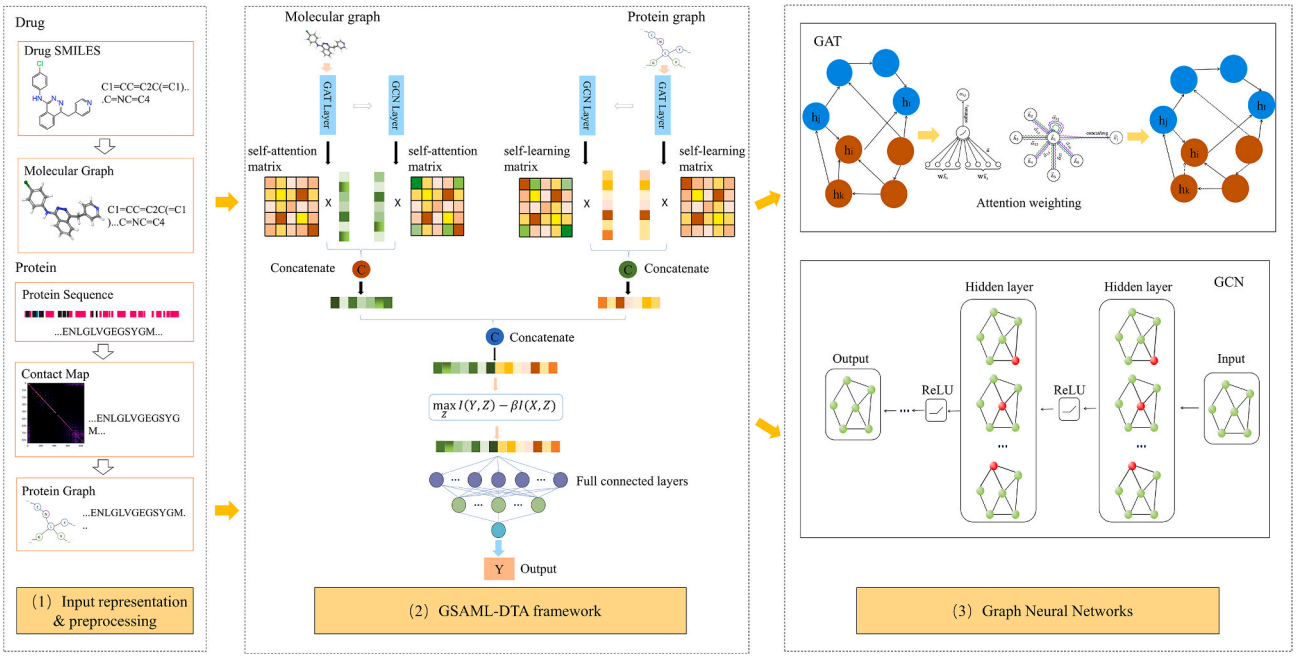

Fig. 1. The framework of the proposed GSAML-DTA model. Firstly, the drug SMILES sequences and protein sequences are transformed into molecular graphs and protein graphs, respectively. Then, the molecular and protein graphs are sent into a hybrid network GAT-GCN with the self-attention mechanism to learn the representations of the drug and protein. Next, mutual information is adopted to optimize the combined representation of a drug-target pair. Finally, the optimized representation is fed into fully connected layers to predict the affinity.

protein. For drug molecule, it is represented by short ASCII string SMILES that depicts the structure of chemical species. We employ RDKit [30] to convert the drug SMILES string into a molecular graph, in which nodes denote atoms and edges denote bonds. Following the model GraphDTA, we adopt a set of atomic features as the initial drug molecular graph feature representation. More detailed information about atomic features is illustrated in Table S1.

2.4. Graph representation of target proteins

Similarly, a target protein can be described as a graph of integrations between residues. Subsequently, graph learning algorithms can be adopted to extract structural information. To this end, we employ the Pconsc4 [31], an open-source protein structure prediction tool, to generate target protein graphs for further mining intrinsic structural information hidden in protein sequences. However, the input of PSSM is the result of protein sequence alignment, and to improve computational accuracy and efficiency, HHblits [32] is adopted for protein sequence alignment. After that, Pconsc4 can convert the results of the protein sequence alignment into contact maps, that are, residue-residue interaction matrixes, in which the value indicates the Euclidean distance between two residues. If the Euclidean distance between a residue pair is less than a certain threshold, there is a contact between them [33]. Similar to the study DGraphDTA, the threshold is set to 0.5 and a group of residue descriptors are utilized as initial features of each residue in the protein graph. More details of the descriptors are displayed in Table S2.

2.5. Representation learning on graphs

The CNN framework has shown remarkable performance in processing regular Euclidean data like text and images. However, it cannot be applied to non-Euclidean data such as molecular graphs. To solve this limitation, the GNN network is designed to directly operate graphs and extract their structural information. At present, GNN has evolved many powerful variants, such as GCN and GAT, which are effective in learning graph feature representation. In this study, we adopt GAT and GCN to capture intrinsic structural information of both drugs and targets.

To extract the structure characteristics of drugs and proteins, we construct the model GSAML-DTA. It passes the input drug features through the GAT layer and GCN layer, and the output features of the two layers are multiplied by the self-attention matrix, respectively. The obtained two features are concatenated as the drug features. Similarly, we can also learn protein features. After the two parts of features are concatenated, the mutual information principle is applied to remove the noise of the combined features. Finally, the optimized features are fed into the full connection layers to predict the affinity.

2.5.1. Graph attention network

The GAT model, an attention-based architecture, is initially proposed to tackle the problem of node classification of graph-structured data [34]. The model introduces a self-attention strategy to learn each node representation by attending to its neighbors. Mathematically, given an input graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of N nodes, each node is embedded by a d-dimensional vector, and $\mathcal{E}$ is the set of edges and is represented as an adjacency matrix $A \in \mathbb{R}^{N\times N}$that describes the graph structure. The input of the GAT layer is a set of node features, $x = \{\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_N}\}$, $\overrightarrow{x_i} \in \mathbb{R}^d$. To obtain node features with sufficient expressive power, a linear transformation is applied to each node by a learnable weight matrix $W \in \mathbb{R}^{d' \times d}$, where d' is the feature dimension of output nodes. For the node i, the attention coefficient between it and its neighbor node j is calculated based their features:

$$e_{ij} = a\left(W\overrightarrow{x_i}, W\overrightarrow{x_j}\right). \tag{2}$$

This value expresses the importance of node j to node i. To be comparable easily across different neighbor nodes, a softmax function is applied to normalize these attention coefficients:

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(e_{ik}\right)}, \tag{3}$$

where $\mathcal{N}_i$ is a set of neighborhoods of node i in the graph. Then, the final output features of each node are obtained by computing a linear com-

bination of the features corresponding to their normalized attention coefficients:

$$\overrightarrow{x_i'} = \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}W\overrightarrow{x_j},\right), \tag{4}$$

where $\sigma(\bullet)$ is the non-linear activation function, which is the ReLU activation function in our model.

2.5.2. Graph convolutional network

To further learn the local topological structure of the graph, we put a GCN layer on the top of the GAT layer. The GCN takes the node feature matrix X and the adjacency matrix A as inputs. The GCN operator $f_{GCN}(\bullet)$ over the graph is defined as:

$$H^{(l+1)} = f_{GCN}\left(H^{(l)}, A\right) = \sigma\left(\widehat{D}^{-\frac{1}{2}}\widehat{A}\widehat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right), \tag{5}$$

where $\widehat{A} = A + I$ is the adjacency matrix with self-loop in each node; I is the identity matrix; $\widehat{D} \in \mathbb{R}^{N\times N}$ is graph diagonal degree matrix of $\widehat{A}$; $\sigma(\bullet)$ is the nonlinear activation function which is ReLU in our model; $W^{(l)}$ is the trainable weight parameter at the l-th layer; $H^{(l)}$ is the node feature matrix at the l-th layer. In our model, $H^{(0)} = X$ that is the output of the GAT layer.

2.5.3. Self-attention mechanism

Consider that the contribution of the features learned by the GAT network and GCN network to the final affinity prediction is different. Following the attention mechanism in GAT, we generate a set of weight matrices to learn the weights of the features obtained from each GCN network. Furthermore, the weighted features are concatenated together as the feature representation of the drug (target) molecule. The formula is as follows :

$$H_s = W_1h_1 \| W_2h_2, \tag{6}$$

where W_1 and W_2 are two attention weight matrices, h_1 is the output of the GAT network, and h_2 is the output of the GCN network. H_s represents the combined features of the two layers.

2.5.4. Mutual information

After obtaining the weighted feature representations of the drug and target separately, we concatenate them as the representation of a drug-target pair. However, directly concatenating different features may occur the problem of curse of dimensionality and introduce redundant information, lead to high computational complexity and prediction performance drop. To learn a more compact and accurate feature representation of a drug-target pair, the mutual information principle is adopted to remove irrelevant information and keep task-relevant information as much as possible [35]. Therefore, our objective is as follows:

$$\max_Z I(Y,Z) - \beta I(X,Z), \tag{7}$$

where β is a trade-off parameter that control the balance between complexity and accuracy of the learned features, and the function I is utilized to measure the mutual information between two random variables A and B as follows:

$$I(A,B) = \int da\, db\, p(a,b) log\frac{p(a,b)}{p(a)p(b)}, \tag{8}$$

where a and b are examples of random variables A and B separately, $p(a,b)$ is the joint distribution, and $p(a)$ and $p(b)$ are all the marginal distributions.

2.5.5. Drug-target binding affinity prediction

In our work, the drug-target binding affinity prediction problem is considered as a regression task. Therefore, we use the embedding $\widehat{C_{ij}}$ optimized by mutual information for final affinity prediction through two full connection layers:

$$y_{i,j} = FC\left(W_1, W_2, \widehat{C_{ij}}\right), \tag{9}$$

where $y_{i,j}$ is the predicted affinity value, W_1 and W_2 are the weight metrics of the two full connection layers.

Finally, the mean squared error (MSE) is adopted as the loss function as follows:

$$L = \frac{1}{n}\sum\left(y_{i,j} - \widehat{y}_{i,j}\right)^2, \tag{10}$$

where $\widehat{y}_{i,j}$ is the truth affinity value of the drug-target pair (d_i,t_j), and n is the sample size. According to the loss function, we optimize the mapping function $\Theta(\omega) : (\mathcal{N}, \mathcal{G}_{d_i}, \mathcal{G}_{t_j}) \longrightarrow \mathbf{y}_{i,j}$ and find the optimal trainable parameter ω.

Algorithm1. The GSAML Algorithm

3. Results and discussion

3.1. Performance evaluation metrics

To assess the performance of the proposed GSAML-DTA, we adopt three commonly used statistical metrics: Concordance Index (CI) [36], Mean Squared Error (MSE), and r_m^2 [37]. CI is mainly employed to assess the difference between the predicted value and the actual value as follows:

$$\mathrm{CI} = \frac{1}{Z}\sum_{d_x-d_y} h\left(b_x - b_y\right), \tag{11}$$

$$h(x) = \begin{cases} 1, if\ x > 0 \\ 0.5, if\ x = 0 \\ 0, if\ x < 0 \end{cases}, \tag{12}$$

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, Z is the normalization constant, and $h(x)$ is the step function as shown in equation (12).

MSE is a statistical measure that directly evaluates error. Assuming there are n estimated samples and their corresponding actual values, MSE is expressed as the expected value of the squared loss:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2, \tag{13}$$

where p_i is the estimated value of the *ith* sample, y_i is the actual value of the *ith* sample.

The r_m^2 is an indicator proposed in DeepDTA. If a variable is substantial, then r_m^2 means how close it is to the mean next time. The specific calculation formula is following:

$$r_m^2 = r^2 \times \left(1 - \sqrt{r^2 - r_0^2}\right), \tag{14}$$

where r^2 is the squared correlation coefficient with an intercept, and r_0^2 is the squared correlation coefficient without an intercept.

3.2. The performance of GSAML-DTA

To evaluate the performance of our proposed model, we compare our model with existing methods, including KronRLS [25], SimBoost [26], DeepDTA, WideDTA, MT-DTI [38], DeepCDA, MATT_ DTI [39], GraphDTA and DGraphDTA. From Table 2, we can see that in the Davis

Table 2
The performance of GSAML-DTA and baseline models on the Davis dataset.

Model	CI(std)	MSE	r_m^2(std)
KronRLS	0.871(±0.001)	0.379	0.407(±0.005)
Simboost	0.872(±0.002)	0.282	0.664(±0.006)
DeepDTA	0.878(±0.004)	0.261	0.630(±0.017)
WideDTA	0.886(±0.003)	0.262	–
MT-DTI	0.887(±0.003)	0.245	0.665(±0.014)
DeepCDA	0.891(±0.003)	0.248	0.649(±0.009)
MATT_DTI	0.891(±0.002)	0.227	0.683(±0.017)
GraphDTA	0.893(±0.001)	0.229	–
DGraphDTA	**0.904(± 0.001)**	0.202	0.700(±0.015)
GSAML-DTA	0.896(±0.001)	**0.201**	**0.718(± 0.004)**

– These results are not reported from original properties.

dataset, our method significantly outperforms other methods in terms of MSE, r_m^2. In particular, GSAML-DTA achieves an MSE of 0.201 and a r_m^2 of 0.718, which yield relative improvements over runner-up method DGraphDTA of 0.1% and 1.8%, respectively. GSAML-DTA does have a lower CI of 0.8% than DGraphDTA, while the differences between them are small.

From Table 3, except for DGraphDTA, GSAML-DTA achieves the best performance amongst other existing methods with a CI of 0.900, an MSE of 0.132 and a r_m^2 of 0.800, which are 0.9–11.8%, 0.7–30.9% and 4.4–45.8% higher than other competing methods. When compared to DGraphDTA, GSAML-DTA achieves a higher r_m^2 of 0.800 (a relative increase 1.4%), while the CI and MSE are slightly worse than DGraphDTA.

The results above indicate that our proposed method can be considered as an accurate and efficient tool for DTA prediction. For the superiority of our model, there are main three reasons: (i) Compared to traditional methods that only extract sequential information from raw protein sequences, our method employs contact maps to represent proteins which will contain more spatial structure information; (ii) To obtain a more discriminative protein (compound) representation, we utilize two weight matrixes to optimize the two features from GAT and GCN networks, respectively, and then concatenate these two optimized features as the protein (compound) representation; (iii) After concatenating the protein and compound features, mutual information is introduced to filter out superfluous information and preserve task-related information as much as possible for obtaining a more compact and accuracy representation. Therefore, our model can integrate the intrinsic information of protein and compound into a more comprehensive representation, which plays a vital role in building an accuracy and robust model.

3.3. Ablation study

We design a set of ablation experiments to identify the contributions of the self-attention mechanism, fusion features, and mutual information. The methods used in the ablation experiments are as follows:

Table 3
The performance of GSAML-DTA and baseline models on the KIBA dataset.

Model	CI(std)	MSE	r_m^2(std)
KronRLS	0.782(±0.001)	0.441	0.342(±0.001)
Simboost	0.836(±0.001)	0.222	0.629(±0.007)
DeepDTA	0.863(±0.002)	0.194	0.673(±0.009)
WideDTA	0.875(±0.001)	0.179	–
MT-DTI	0.882(±0.001)	0.152	0.738(±0.006)
DeepCDA	0.889(±0.002)	0.176	0.682(±0.008)
MATT_DTI	0.889(±0.001)	0.150	0.756(±0.011)
GraphDTA	0.891(±0.002)	0.139	–
DGraphDTA	**0.904(± 0.001)**	**0.126**	0.786(±0.011)
GSAML-DTA	0.900(±0.004)	0.132	**0.800(± 0.004)**

– These results are not reported from original properties.

(1) **GSAML-DTA with only GAT-GCN hybrid network (only GHN)** is the most basic GAT-GCN hybrid network. The raw sequences pass through the GAT and GCN networks successively and then directly enter the fully connected layer for prediction.
(2) **GSAML-DTA without self-attention mechanism (w/o SAM)** does not employ self-attention mechanism to optimize the concatenation of features learned by multi-layer networks and does not adopt mutual information, which directly concatenates the features learned by each layer of the network without weighting.
(3) **GSAML-DTA without mutual information (w/o MI)** does not use mutual information to filter out the irrelevant information. The features acquired by the networks are directly imported into the fully connected layer for prediction.

Fig. 2 shows the comparison results between GSAML-DTA and its three variants on two benchmark datasets. Overall, the proposed GSAML-DTA significantly outperforms its variants, which demonstrates the effectiveness of the GSAML-DTA model architecture. Specifically, the performance of GSAML-DTA (w/o SAM) is better than GSAML-DTA (only GHN), which confirms the effectiveness of the feature concatenation method. In addition, GSAML-DTA (w/o SAM) performance is worse than GSAML-DTA (w/o MI) since it directly concatenates the features from GAT and GCN networks, which results in more task-irrelated information hidden in the combined features. It suggests the significance of the self-attention mechanism component for prediction. Besides, GSAML-DTA performs better than GSAML-DTA (w/o MI), which shows that the mutual information principle is beneficial to filter out task-irrelated information and retain effective information.

3.4. Performances of different GNN models

It is critical to construct effective GNN networks to extract the discriminative characteristics of drugs and targets for improving the prediction accuracy of DTA. According to experience, it is often difficult for a single-layer network to obtain enough information compared with a multi-layer network, while too many layers may introduce too much noise. Therefore, this experiment is only for the two-layer network. Here, we adopt different schemes to combine two types of graph network architectures (GCN and GAT) and implement their performance comparison. Detailed network information and results are shown in Table 4 and Fig. 3, respectively. From Fig. 3(a), it is obvious to see that when GAT and GCN are employed to extract the features of drugs and targets in turn, the model achieves the best performance, which generates an MSE of 0.201, a CI of 0.896, and a r_m^2 of 0.718. Although when using the GCN-GAT and the GAT-GCN hybrid networks to extract the features of drugs and proteins, respectively, the r_m^2 of the model can reach 0.755, which is higher than GSAML-DTA, but the gap is small. Besides, we can see that in Fig. 3(b), GSAML-DTA achieves better performance than other methods in terms of MSE, CI, and r_m^2, which are 0.132, 0.900, and 0.800, respectively. Given the above analysis, the GAT-GCN network for characterizing drugs and proteins is finally adopted.

3.5. Model interpretability based on the compound

Machine learning is often regarded as a black box model due to that it is challenging to locate and analyze which features are essential. The lack of interpretability limits the further application of deep learning methods, especially in computer-aided drug discovery. To explore whether our models can detect essential substructures responsible for specific toxicity or not (also called structural alerts) [40,41], we invoke Grad-AAM [22] to visualize the atomic importance of molecules. The following three groups of tests were designed:

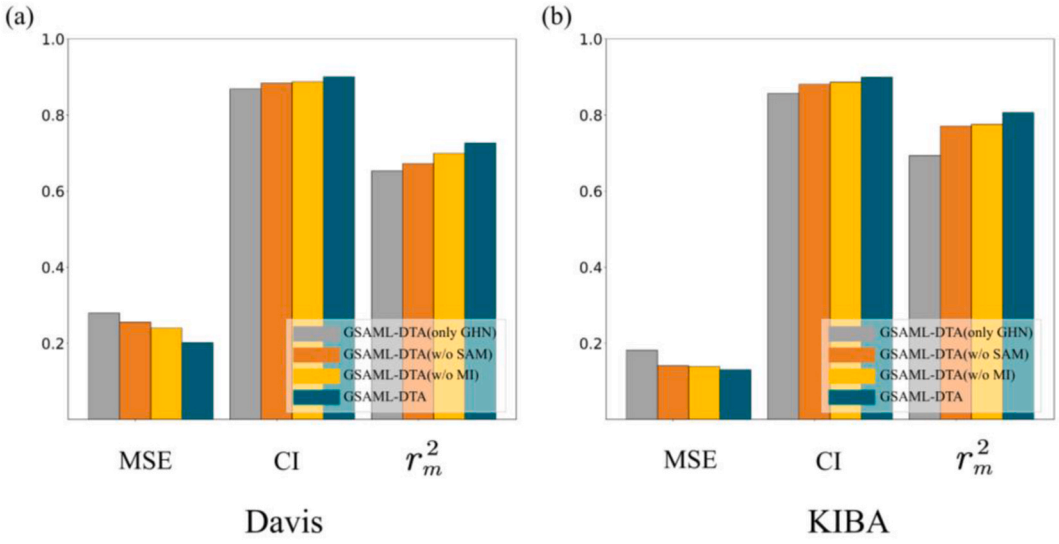

Fig. 2. Investigate the individual contributions of the self-attention mechanism, fusion features and mutual information on Davis and KIBA datasets.

Table 4
Combinations of different GNN models.

Scheme	Compounds	Proteins
1	2GAT	2GAT
2	2GCN	2GCN
3	GCN + GAT	CAT + GCN
4	CAT + GCN	GCN + GAT
5	GAT + GCN	GAT + GCN

(1) **GSAML-DTA**
(2) **GAT-GCN** adopts traditional GAT-GCN hybrid network instead of GSAML-DTA network.
(3) **GAT-GAT** introduces a two-layer GAT network instead of GSAML-DTA network.

Fig. 3 shows that GSAML-DTA performs better than the model with only a GAT-GCN or GAT-GAT networks in terms of MSE and CI, which confirms the superiority of GSAML-DTA. Fig. 4 shows the visualization results of some molecules (the deeper the color, the more important structure). According to previous studies [42–46], epoxide [45], fatty acid [43,46], sulfonate [42], and aromatic nitroso [44] are all fundamental structures of specific pathology. As shown in Fig. 4(a)–(d), we found that in GSAML-DTA, Grad-AAM did give higher weights to these structures, while in GAT-GAT, these significant weights were lower, which showed that our model could indeed learn the key structures for

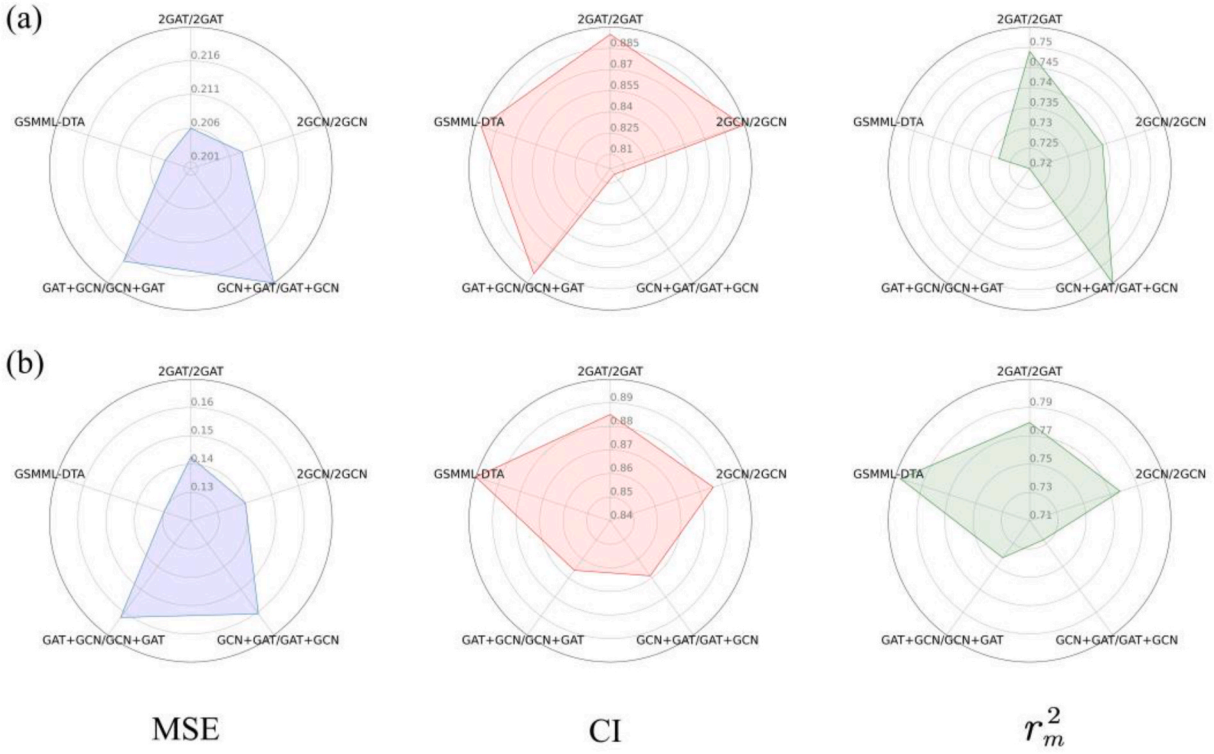

Fig. 3. Performances of different GNN models on benchmark datasets. (a) Results on Davis dataset, (b) Results on KIBA dataset.

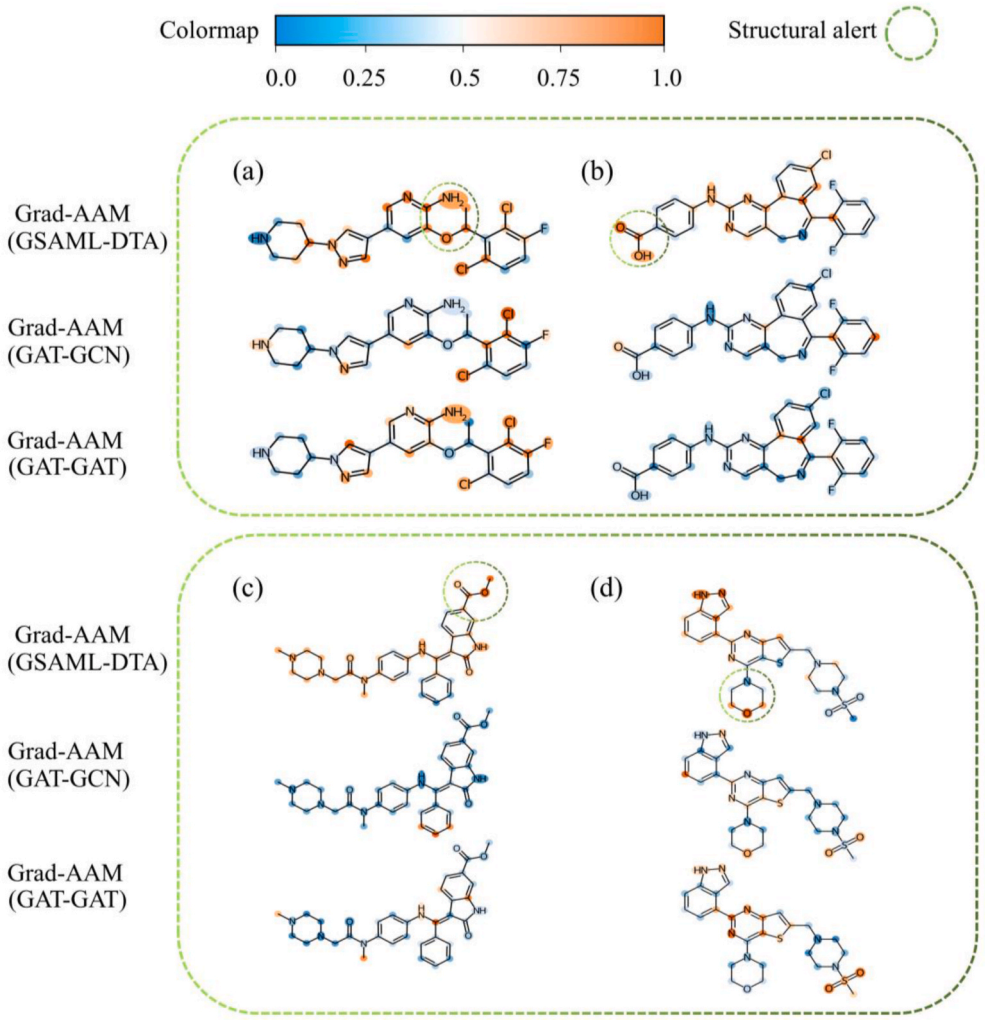

Fig. 4. Atom significance uncovered by Grad-AAM (GSAML-DTA), Grad-AAM (GAT-GCN) and Grad-AAM (GAT-GAT), and graph attention in structural alerts of (a) and (d) epoxides, (b) fatty acids, (c) fatty acid salts.

affinity prediction. Moreover, the GAT-GAT network performed slightly better than the GAT-GCN network, and in Fig. 4(a) and (b), GAT-GAT also paid more attention to the important structures. In addition, Grad-AAM also showed that our model can obtain not only key local information but also global structural information, as shown in Fig. 4(c) and (d). Therefore, the experimental results show that GSAML-DTA pays more attention to critical structures and achieves a higher performance.

3.6. Model interpretation based on the protein

To further explain the effectiveness of GSAML-DTA, we use the attention weight learned by GSAML-DTA to analyze the interactions between drug compounds and the target proteins, which play key roles in the binding pocket. To visualize the main interaction regions, we first learn and obtain the attention weight matrix of drug compounds and target proteins through the GAT network and then sort the weights to select the interaction sites with greater attention. Fig. 5 shows an example of weight visualization of the proposed GSAML-DTA.

We select benzoic acid molecule and cdkl2 (PDB: 4aaa) protein for interactive visual analysis. The results show that the weight of drug compounds ranges from 0.76E-1 to 1.51. As shown in Fig. 5, we rank them according to attention and mark some atoms larger than 1.2 in red. The protein weight ranges from 1.38E-1 to 3.00 are obtained by our model and the main amino acid regions are residues 113–330. The peak is LEU-142. Most of the residues with large weight fall in the binding pocket, which indicates that our model can accurately predict the potential binding sites. It also shows that our model can accurately capture the important structure for predicting the interactions between proteins and drugs. While some basic binding residues were not detected, some erroneous binding sites were highlighted, suggesting that some essential residues may indirectly affect binding. Overall, our model can focus on most residues in the binding pocket, demonstrating that our model can capture crucial structural features that play vital roles in drug-target binding.

4. Conclusion

In this study, we propose a novel deep-learning model, GSAML-DTA, to predict binding affinities of drug-target pairs, which is a crucial step for rapid virtual drug screening and drug development. We first generate graphs of the drug and target, and then employ a self-attention mechanism and a hybrid graph neural network GAT-GCN to extract structural information of them. Subsequently, to learn an informative representation of the drug-target pair, mutual information is applied to the combined features for removing irrelevant elements and retaining relevant information as much as possible. A series of evaluation experiments are conducted on two benchmark datasets and the results show that the proposed method achieves a better predictive performance than the state-of-the-art methods. Moreover, the model has been shown to identify the essential binding structures and residues in the drugs and

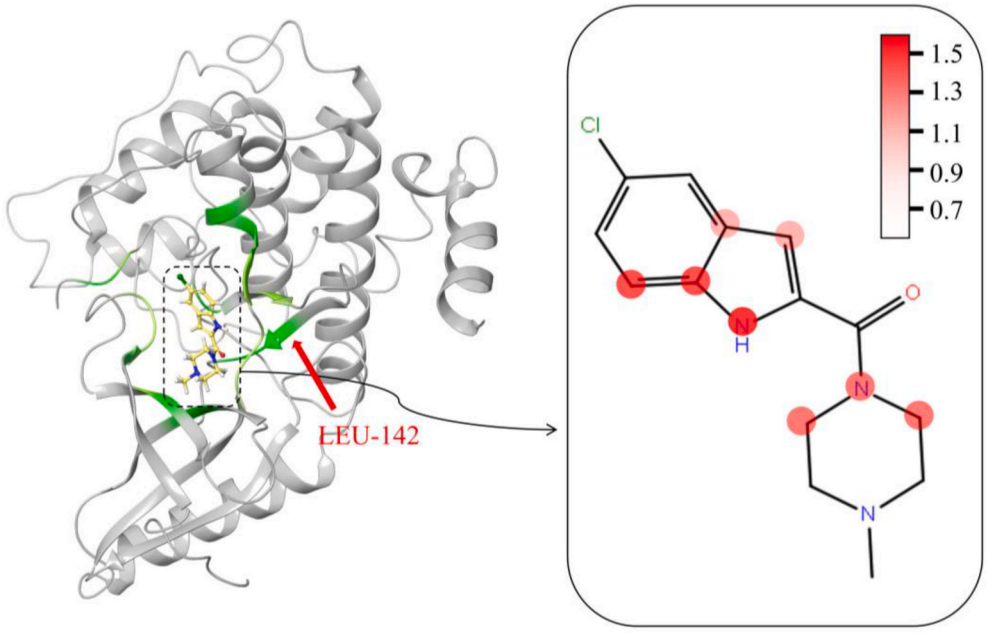

Fig. 5. An example of the weight visualization of the proposed model. The cdkl2 (PDB: 4aaa) is shown with a cartoon, while the benzoic acid molecule is shown with a stick. The green highlights in the left part are the highly focused positions of the protein and focused drug atoms in the binding pocket, and the deeper the color, the greater the attention weight. And the red highlights in the right region are the area where the molecular attention is relatively large.

targets, respectively. These advantages demonstrate that GSAML-DTA not only improves the predictive ability of DTA prediction, but also provides biological insights for understanding the potential drug-target binding mechanism.

Although GSAML-DTA has improved the performance of the DTA prediction, there are still several drawbacks in the current work. Firstly, this work only utilizes two-layer GNNs to extract structural features of molecules, which may be insufficient to learn the global information of a graph. We will consider building a deep framework of GNN for capturing more comprehensive topological information of a graph in future work. In addition, it is worth noting that GSAML-DTA can focus on specific 'important' sites in the drug or target molecule by the attention mechanism. However, it hits a roadblock in automatically identifying the accurate interactive sites between drugs and targets without extra information. In the future, we will focus on constructing an end-to-end deep learning architecture to predict the drug-target binding sites in the drugs and targets based on only primary sequences.

Funding

This study was supported by the Natural Science Foundation of China (No. 62071278).

Declaration of competing interest

There is no competing financial interest to declare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.106145.

References

[1] D.J. Newman, G.M. Cragg, Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019, Journal of Natural Products 83 (2020) 770–803.
[2] T. Takebe, R. Imai, S.J.C. Ono, t. science, The current status of drug discovery and development as originated in United States academia, the influence of industrial and academic collaboration on drug discovery and development 11 (2018) 597–606.
[3] T. Zhao, Y. Hu, L.R. Valsdottir, T. Zang, J. Peng, Identifying drug-target interactions based on graph convolutional network and deep neural network, Briefings in Bioinformatics 22 (2021) 2141–2150.
[4] 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,, Journal of Chemical Information and Modeling 54 (2014) 735–743.
[5] W. Xue, P. Wang, G. Tu, F. Yang, G. Zheng, X. Li, X. Li, Y. Chen, X. Yao, F. Zhu, Computational identification of the binding mechanism of a triple reuptake inhibitor amitifadine for the treatment of major depressive disorder, Phys. Chem. Chem. Phys. 20 (2018) 6606–6616.
[6] 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.J.R. Kuntz, Dock 6, Combining techniques to model RNA–small molecule complexes 15 (2009) 1219–1230.
[7] 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, Journal of Computational Chemistry 30 (2009) 2785–2791.
[8] E.B. Lenselink, N. ten Dijke, B. Bongers, G. Papadatos, H.W.T. van Vlijmen, W. Kowalczyk, A.P. Ijzerman, G.J.P. van Westen, Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set, Journal of Cheminformatics 9 (2017).
[9] A.S. Rifaioglu, R. Cetin Atalay, D. Cansen Kahraman, T. Doğan, M. Martin, V.J. B. Atalay, MDeePred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in, drug discovery 37 (2021) 693–704.
[10] Q. Yang, B. Li, S. Chen, J. Tang, Y. Li, Y. Li, S. Zhang, C. Shi, Y. Zhang, M. Mou, W. Xue, F. Zhu, MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis, J. Proteonomics 232 (2021), 104023.
[11] J. Hong, Y. Luo, Y. Zhang, J. Ying, W. Xue, T. Xie, L. Tao, F. Zhu, Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning, Briefings Bioinf. 21 (2020) 1437–1447.
[12] W. Xia, L. Zheng, J. Fang, F. Li, Y. Zhou, Z. Zeng, B. Zhang, Z. Li, H. Li, F. Zhu, PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods, Comput. Biol. Med. 145 (2022), 105465.
[13] J. Hong, Y. Luo, M. Mou, J. Fu, Y. Zhang, W. Xue, T. Xie, L. Tao, Y. Lou, F. Zhu, Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery, Briefings Bioinf. 21 (2020) 1825–1836.
[14] H. Öztürk, A. Özgür, E.J.B. Ozkirimli, DeepDTA: deep drug–target binding affinity prediction 34 (2018) i821–i829.
[15] H. Öztürk, E. Ozkirimli, A. Özgür, WideDTA: prediction of drug-target binding affinity, arXiv (2019 Feb 4) preprint arXiv:1902.04166.
[16] K. Abbasi, P. Razzaghi, A. Poso, M. Amanlou, J.B. Ghasemi, A.J.B. Masoudi-Nejad, DeepCDA: deep cross-domain compound–protein affinity prediction through, LSTM and convolutional neural networks 36 (2020) 4633–4642.

[17] M. Karimi, D. Wu, Z. Wang, Y.J.B. Shen, DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks 35 (2019) 3329–3338.
[18] M. Karimi, D. Wu, Z. Wang, Y. Shen, Explainable Deep Relational Networks for Predicting Compound- Protein Affinities and Contacts, Journal of Chemical Information and Modeling 61 (2021) 46–66.
[19] T. Nguyen, H. Le, T.P. Quinn, T. Nguyen, T.D. Le, S.J.B. Venkatesh, GraphDTA, Predicting drug–target binding affinity with graph neural networks 37 (2021) 1140–1147.
[20] W. Torng, R.B. Altman, Graph Convolutional Neural Networks for Predicting Drug-Target Interactions, Journal of Chemical Information and Modeling 59 (2019) 4131–4149.
[21] M. Tsubaki, K. Tomii, J. Sese, Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences, Bioinformatics 35 (2019) 309–318.
[22] Z. Yang, W. Zhong, L. Zhao, C.Y.-C.J. Chen, MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction, Chemical Science (2022).
[23] X.J.a.p.a. Lin, Deepgs: deep representation learning of graphs and sequences for drug-target binding affinity prediction, 2020.
[24] M. Jiang, Z. Li, S. Zhang, S. Wang, X. Wang, Q. Yuan, Z.J.R.A. Wei, Drug–target affinity prediction using graph neural network and contact maps 10 (2020) 20701–20712.
[25] T. Pahikkala, A. Airola, S. Pietila, S. Shakyawar, A. Szwajda, J. Tang, T. Aittokallio, Toward more realistic drug-target interaction predictions, Briefings in Bioinformatics 16 (2015) 325–337.
[26] T. He, M. Heidemeyer, F. Ban, A. Cherkasov, M. Ester, SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines, Journal of Cheminformatics 9 (2017).
[27] T. Nguyen, H. Le, S.J.B. Venkatesh, GraphDTA: prediction of drug–target binding affinity using graph convolutional networks, 2019, 684662.
[28] E.E. Bolton, Y. Wang, P.A. Thiessen, S.H. Bryant, PubChem: Integrated Platform of Small Molecules and Biological Activities, Annual Reports in Computational Chemistry, 2008, pp. 217–241. Elsevier.
[29] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M.J.N.a.r. Magrane, UniProt: the universal protein knowledgebase 32 (2004) D115–D119.
[30] A.P. Bento, A. Hersey, E. Felix, G. Landrum, A. Gaulton, F. Atkinson, L.J. Bellis, M. De Veij, A.R. Leach, An open source chemical structure curation pipeline using RDKit, Journal of Cheminformatics 12 (2020).
[31] M. Michel, D. Menéndez Hurtado, A.J.B. Elofsson, PconsC4: fast, accurate and hassle-free contact predictions 35 (2019) 2677–2679.
[32] M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S.J. Haunsberger, J.J. Söding, HH-suite3 for fast remote homology detection and deep protein annotation, BMC Bioinformatics 20 (2019) 1–15.
[33] Q. Wu, Z. Peng, I. Anishchenko, Q. Cong, D. Baker, J.J.B. Yang, Protein contact prediction using metagenome sequence data and residual neural networks 36, 2020, pp. 41–48.
[34] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y.J.a.p.a. Bengio, Graph attention networks, 2017.
[35] L. Wei, X. Ye, T. Sakurai, Z. Mu, L.J.B. Wei, ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning, 2022.
[36] M. Gönen, G.J.B. Heller, Concordance probability and discriminatory power in proportional hazards regression 92 (2005) 965–970.
[37] 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, Journal of Computational Chemistry 34 (2013) 1071–1082.
[38] B. Shin, S. Park, K. Kang, J.C. Ho, Self-attention Based Molecule Representation for Predicting Drug-Target Interaction, Machine Learning for Healthcare Conference, PMLR, 2019, pp. 230–248.
[39] Y. Zeng, X. Chen, Y. Luo, X. Li, D.J. Peng, Deep drug-target binding affinity prediction with multiple attention blocks, Briefings in Bioinformatics 22 (2021) bbab117.
[40] Z. Wu, D. Jiang, J. Wang, C.-Y. Hsieh, D. Cao, T.J. Hou, Mining toxicity information from large amounts of toxicity data, J. Med. Chem. 64 (2021) 6924–6936.
[41] A. Mukherjee, A. Su, K. Rajan, Deep Learning Model for Identifying Critical Structural Motifs in Potential Endocrine Disruptors, Journal of Chemical Information and Modeling 61 (2021) 2187–2197.
[42] M. Barratt, D. Basketter, M. Chamberlain, G. Admans, J.J.T.i.v. Langowski, An expert system rulebase for identifying contact allergens 8 (1994) 1053–1060.
[43] A.S. Kalgutkar, J.R. Soglia, Minimising the potential for metabolic activation in drug discovery, Expert opinion on drug metabolism & toxicology 1 (2005) 91–142.
[44] J. Kazius, R. McGuire, R. Bursi, Derivation and validation of toxicophores for mutagenicity prediction, Journal of Medicinal Chemistry 48 (2005) 312–320.
[45] M.P. Payne, P.T. Walsh, STRUCTURE-ACTIVITY-RELATIONSHIPS FOR SKIN SENSITIZATION POTENTIAL - DEVELOPMENT OF STRUCTURAL ALERTS FOR USE IN KNOWLEDGE-BASED TOXICITY PREDICTION SYSTEMS, Journal of Chemical Information and Computer Sciences 34 (1994) 154–161.
[46] M.M. Shahzad, J.M. Arevalo, G.N. Armaiz-Pena, C. Lu, R.L. Stone, M. Moreno-Smith, M. Nishimura, J.-W. Lee, N.B. Jennings, J.J. Bottsford-Miller, Stress effects on FosB-and interleukin-8 (IL8)-driven ovarian cancer growth and metastasis, J Biol Chem. 285 (2010) 35462–35470.