International Journal of Machine Learning and Cybernetics (2024) 15:2709–2718
https://doi.org/10.1007/s13042-023-02058-3

ORIGINAL ARTICLE

Multidta: drug-target binding affinity prediction via representation learning and graph convolutional neural networks

Jiejin Deng[1] · Yijia Zhang[1] · Yaohua Pan[1] · Xiaobo Li[1] · Mingyu Lu[1]

Received: 30 April 2023 / Accepted: 26 November 2023 / Published online: 9 January 2024

Abstract
The prediction of Drug-Target Interactions (DTI) plays a pivotal role in drug repositioning research. While recent years have witnessed the proliferation of neural network-based methods for Drug-Target Affinity (DTA) prediction, existing models predominantly rely on either sequence-based or graph-based approaches to model drug-target pairs. This limitation obstructs models from harnessing more valuable information from various data sources for downstream predictions. To overcome this constraint, this paper introduces an innovative end-to-end learning framework for DTA prediction, named MultiDTA. Firstly, we construct four channels tailored to comprehensively mine representations embedded within drug-target pair sequences and model them through graph structures to learn spatial structural information. Secondly, after capturing latent high-level representations from different data structures across these four channels, we employ an attention mechanism to discern each channel's contributions to downstream tasks. Experimental results demonstrate that our proposed model surpasses sequence-based and graph-based methods, affirming our model's capacity to simultaneously capture high-level representations from multiple data structures. Furthermore, we enhance the model's interpretability by visualizing the contributions of these four channels using the attention mechanism. The code of MultiDTA and the relevant data are available at: https://github.com/dengjiejin/MultiDTA.

Keywords Drug-target affinity prediction · Representation learning · Multi-channel inputs · Graph convolution networks

1 Introduction

Predicting the relationship between drugs and targets is very important for drug knowledge discovery. However, traditional wet lab experiments are inefficient, expensive, and time-consuming [1, 2]. The prediction of drug-target binding affinity based on deep learning can significantly accelerate the efficiency of research and development, which is one of the most important in drug discovery.

Methods for DTA prediction can be categorized into two types: binary classification [3–6], and regression [7–9]. In the past, the DTI modeling task was primarily viewed as a binary classification problem, which overlooked the essential characteristic of protein-ligand interaction. In particular, the binary classification task ignored the binding affinity score to indicate the interaction strength between drugs and targets. The previous work [10] has suggested that there are two problems with binary classification for DTA task: (1) it cannot distinguish unknown values between positive and negative interactions, and (2) the binary relationship is too simplistic to express the strength of the interaction between the drug-target pair. To avoid these problems, we use regression tasks to solve the DTA prediction problem. Currently, DTA predictions are treated as a regression task in most methods. The advantage of treating the prediction of the drug-protein relationship as a regression problem is that it avoids the influence of negative samples on the model, which can provide more practical and valuable information.

✉ Yijia Zhang
zhangyijia@dlmu.edu.cn

✉ Mingyu Lu
lumingyu@dlmu.edu.cn

Jiejin Deng
dengjiejin@dlmu.edu.cn

Yaohua Pan
panyaohua@dlmu.edu.cn

Xiaobo Li
1120210266@dlmu.edu.com

[1] School of Information Science and Technology, Dalian Maritime University, Dalian 116024, Liaoning, China

With the abundant generation of biomedical data in recent years and the continuous breakthroughs in deep learning across various domains, an increasing number of deep learning methods have been applied to DTA prediction. Sequence-based methods [8, 9] have achieved promising results in earlier deep learning approaches. DeepDTA [8] introduced a DTA prediction model based on convolutional neural networks (CNNs). The model leverages the representation learning of drugs and targets to predict binding affinities. As a widely adopted deep learning-based binding affinity prediction model, DeepDTA has demonstrated satisfactory performance. However, it is constrained by the limitations of CNN, particularly in capturing long-range relationships between atomic fragments within drugs when required. Shin et al. [9] introduced a model of a self-attention mechanism model based on unknown embeddings to encode the relationships between all atoms in a compound. However, more than modeling the compounds is required since these existing approaches only label each atom as a corresponding integer according to a dictionary. Although the atoms at specific positions are learned in the simulation of compounds, the correlation between atoms is ignored, and each atom is separated. Zhao et al. [11] proposed a GANsDTA model based on the generative adversarial network (GAN) to learn beneficial patterns in labeled and unlabeled sequences and utilizes convolutional regression to predict binding affinity scores. Zhao et al. [12] proposed an AttentionDTA model based on the attention mechanism. The model uses an attention mechanism to consider which subsequences in proteins are more critical to proteins and which subsequences in drugs are more important to drugs. These methods convert drug compounds and proteins to the corresponding string representation, which is not an efficient way to characterize molecularly.

The sequence-based approach described above produces promising results, but it ignores key 3D structural information of the molecules. In recent studies, graph neural networks (GNNs) have been used to address this problem in drug-target interaction (DTI) prediction. Nguyen et al. [13] introduced the GraphDTA model based on GNN for DTA prediction, which constructs a graph with atoms as nodes and bonds as edges to describe drug molecules. However, GraphDTA only considers the topology of the molecule, while the structure of the protein molecule was ignored. Cheng et al. [14] and Li et al. [15] employed graph attention networks(GATs) to extract drug features. Compared to sequence-based methods, graph-based methods offer the advantage of capturing the three-dimensional structural information of drug molecules, leading to richer information and successful learning of drug molecule representations. However, these graph-based methods, while capturing the three-dimensional structural information of drugs, tend to overlook the unique features hidden within the sequences. Therefore, we believe that instead of segregating sequence-based modeling methods and graph-based modeling methods, they should be integrated comprehensively to learn richer representations.

To address the aforementioned limitations, this paper introduces a novel end-to-end DTA prediction model, MultiDTA, which extensively considers the sequence and spatial topology information of drugs and proteins. Firstly, we establish four channels to comprehensively mine the embedded representations within drug-target sequences and model them through graph structures to learn spatial structural information. Secondly, after capturing potential high-level representations from the different data structures of these four channels, we employ an attention mechanism to identify the contributions of each channel to downstream tasks. Diverging from existing deep learning-based DTA models, we simultaneously leverage the two-dimensional and three-dimensional spatial structures of drugs and proteins. This enables us to extract local chemical context information of drugs and proteins, along with their spatial topology information, thereby enhancing prediction accuracy. In summary, the primary contributions of this paper are as follows:

- We propose a four-channel DTA prediction input model that utilizes both drug and protein sequences and graph structures as inputs, harnessing the two-dimensional and three-dimensional spatial structures of drugs and proteins for DTA prediction.
- We have designed an attention-based fusion mechanism that empowers the model with the capability to learn the contributions of the four channels.
- Extensive experiments on two real datasets demonstrate the superiority of MultiDTA. Additionally, we conducted rigorous ablation experiments to validate the effectiveness of each proposed channel and provided interpretability analysis of the contribution of each channel through attention visualization.

2 Methodology

In this Section, we describe our proposed four-channel approach named MultiDTA, which comprises five key components. The first two components use convolution neural network(CNN) and long short-term memory (LSTM) to learn low-dimensional vector representations of drug and protein sequences, respectively. The other two components

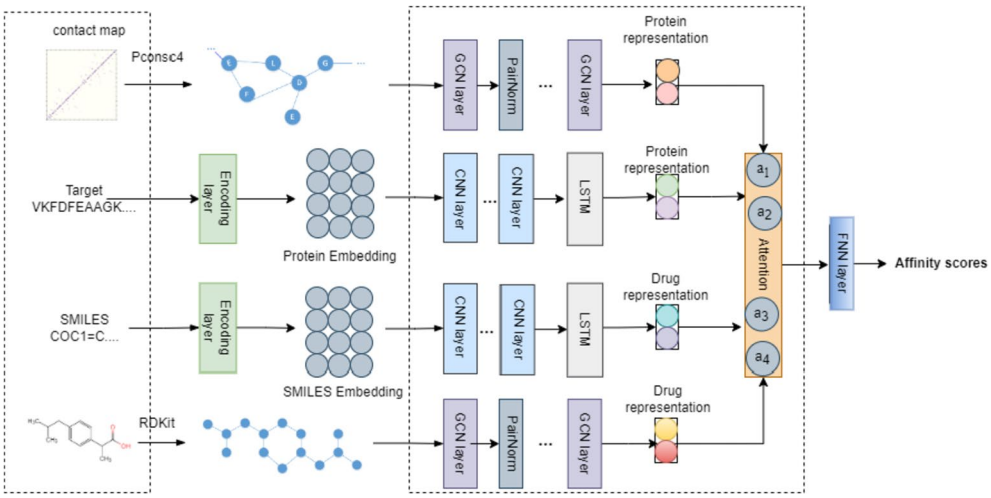

Fig. 1 Overview of MultiDTA. MultiDTA is an integrated framework composed of five components, including four channels and an attention mechanism. The first two channels utilize Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to learn low-dimensional vector representations of drug and protein sequences, respectively. The other two components employ Graph Convolutional Networks (GCNs) to capture the topological structural information of drugs and proteins. The features from these final four channels are fused together using an attention mechanism, assigning varying weights to represent the distinct contributions of each channel

employ a graph convolution network(GCN) to learn the topological structure information of drugs and proteins. The final component is used to predict the interaction between the drug and the target. Figure 1 shows the overview of the MultiDTA. It takes the symbolic sequences of proteins and drugs and the molecular structures as multi-channel inputs. Its output is the affinity of the drug for the target. The following is a detailed introduction to our model.

2.1 Representation learning of drug sequence

In this section, we utilize the Simplified Molecular Line-Entry System (SMILES) format, a commonly used computer-readable format [16], to represent drugs and other molecules. We take the SMILES string of a drug as input and learn representations embedded within the sequence. For example, the SMILES string of the drug "COC1=C" shown in Fig. 1. The SMILES is a specification in the form of a line maker describing the structure of a compound, which is a sequence of atoms and covalent pieces. We consider both atoms and covalent bonds as symbolic markers for ease of representation. According to [8], the SMILES sequence consists of 64 characters. In our model, we use integer/label encoding, integers as categories to represent inputs. Each label is represented by the corresponding integer ('C': 1, '=': 2, 'N': 3, etc.). For example, the sequence representation of SMILES is given below: [C N = C = O] = [1 3 2 1 2 5]. Similar to [8], we set a fixed length to obtain effective representation since each drug has a different length. Specifically, we set the maximum length of SMILES to 100. The sequence with a length greater than the maximum length will be truncated, and the sequence with a shorter length will be filled with zero. Such a length setting can reduce computational complexity and retain enough effective information.

After the above preprocessing of the string, we input the processed data fed into the embedding layer and output the two-dimensional embedding matrix. Then, we utilize a combination module of CNN and LSTM to extract two-dimensional features for drugs. Each filter in the CNN convolves with the input of that layer to encode the local knowledge of the small receptive domain. Because the drug sequence can be regarded as a time series, recurrent neural networks such as LSTM can extract the sequence information better. The sequence input to CNN layers first, then input to LSTM to extract its middle layer features, which reduce the training difficulty of LSTM. The CNN-LSTM combination has been successfully applied to several fields [17–19]. The complete pseudo-code of the representation learning of drug sequence algorithm is illustrated in Algorithm 1.

Algorithm 1 An algorithm for representation learning of drug sequence

Input: Initialized embedding of SMILES emb_0, The number of 1DCNN layers M, Constant λ, Layer normalization function LN
$emb_{in} \leftarrow Conv(emb_0)$
for m in M **do**
    $emb_{out} \leftarrow LN(Conv(emb_{in}) + emb_{in} * \lambda)$
    $emb_{in} \leftarrow LSTM(emb_{out})$
end for
Output: Latent high-level representations of drug sequences emb_{in}

2.2 Representation learning of target sequence

Similar to the representation learning for molecular sequences, we take the amino acid sequence of proteins as input and extract hidden deep representations from it. For protein sequences, there are 25 different categories. For example: ['A': 1, 'E': 4, etc.]. The protein code of "AACGFED" is given below: [A A C G F E D] = [1 1 2 6 5 4 3]. Same as the drug sequence. Each protein has a different length, and we set a fixed length to obtain a valid representation. The maximum length of the protein is set to 1000.

To extract protein sequence features, we follow the same process as with drugs. Firstly, the protein sequence passes through the embedding layer to obtain the initial embedding. Then we utilize the combined module of CNN and LSTM to extract two-dimensional features for proteins.

2.3 Representation learning of drug topology

Due to the fact that drug molecules exist in the real world as three-dimensional structures, representing them using sequences is not their natural form of representation. In this section, we employ a graph-based approach to model drug molecules, aiming to capture the three-dimensional structural information of these molecules and enhance the representation of drug-related information. We use SMILES strings as input and, with the assistance of the RDKit tool [20], construct molecular graphs based on the drug's SMILES string. Subsequently, we leverage graph neural networks to learn the topological structure information of drug molecules.

We first use the RDKit tool to convert SMILES strings into molecular graph $G = (V, E)$, in which the node $v_i \in V$ represents the i-th atom, and the edge $e_{ij} \in E$ represents the chemical bond between the i-th and the j-th atoms. In order to guarantee that the graph can fully consider the features of nodes in the convolution process, self-loops were added to improve the characteristic performance of drug molecules. After the SMILES string is converted into graphics, it can be applied to downstream task analysis using geometric depth learning technology. For GCN, the propagation formula between layers is as follows:

$$H^{l+1} = f(H^l, A) = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^l W^l) \tag{1}$$

where $\tilde{A} = A + I$, the A denotes the adjacency matrix, and the I denotes the identity matrix. The $\tilde{D}$ is the degree matrix of $\tilde{A}$, the H denotes the characteristics of each layer, and the W denotes the weight matrix.

In order to get a vector representation of the drug. The pooling operation is applied in the last layer of GCN to convert the molecular graph into a vector:

$$P_d = POOLING(H^l) \tag{2}$$

where P_d is the vector representation of the drug molecular graph.

We introduced PariNorm [21] to address the potential oversmoothing issue that may arise from excessive layer stacking during training. We add PairNorm regularization to each layer after propagation, except for the last layer. PairNorm is a normalization of the output of the graph convolution. After PairNorm processing, because the total distance between nodes remains unchanged, nodes belonging to the same category are smoother after the convolution layer, making the distance between nodes smaller, while the distance between nodes belonging to different classes is more extensive, thus avoiding the problem of over-smoothing.

The core idea of PairNorm is to keep the difference between all node features output by GNN at each layer as a constant to prevent the features of all nodes from becoming consistent. Defined $TPSD(X) = \sum_{i,j\in[n]} \|x_i - x_j\|_2^2$. $\dot{X}$ is the PariNorm output, and the goal of PairNorm is to ensure that $TPSD(\dot{X}) = TPSD(X)$:

$$\sum_{(i,j)\in\varepsilon} \|\dot{x}_i - \dot{x}_j\|_2^2 + \sum_{(i,j)\notin\varepsilon} \|\dot{x}_i - \dot{x}_j\|_2^2 = \sum_{(i,j)\in\varepsilon} \|x_i - x_j\|_2^2 + \sum_{(i,j)\notin\varepsilon} \|x_i - x_j\|_2^2 \tag{3}$$

Because convolution tends to produce similar features for neighboring nodes, the first term on the left side of the equation must not exceed the first term on the right. Consequently, when establishing the control equation, the second term on the left side will be no smaller than the second term on the right side. As a result, this equation ensures that nodes sharing similarities have comparable characteristics and prevents the feature differences among distant nodes from becoming overly small.

2.4 Presentation learning of target topology

We employ Graph Neural Networks (GNNs) to extract latent feature vectors for proteins, necessitating the construction of the protein's graph structure. Similar to handling drug molecules, the steps for protein feature extraction include building a graph representation for the protein and then utilizing GNNs to extract spatial structural information from the protein. To construct the protein's graph structure, we utilize Pconsc4. Pconsc4 can predict a protein's contact map, which is a fast, convenient, open-source, and effective method for obtaining the protein's graphical structure. The output of Pconsc4 is the probability of contact between residues, and we typically use a threshold of 0.5 to obtain a contact map of size (L, L), where L represents the number of nodes (residues). This contact map accurately corresponds to the protein's adjacency matrix, which effectively encodes the protein's spatial information. This is crucial for understanding the binding affinity between proteins and molecules.

Once the protein is transformed into a graph-based representation, we utilize multiple Graph Convolutional Network (GCN) layers to extract the protein's spatial topological information, ultimately obtaining the protein's three-dimensional feature information. These pieces of information are vital for predicting interactions between drug molecules and proteins in downstream tasks.

2.5 Attention-based feature fusion

In this study, we consider drug-target binding affinity prediction as a regression task. We take the features learned from the four channels, process them through the attention layer, feed them into the fully connected layer, and finally output the predicted binding affinity values. Since different inputs represent different degrees of the contribution of the extracted information to the final output results, we combine attention to the model to further improve the model's predictive performance. The weight of each part is obtained by normalizing the representation information of the four features via Softmax, as follows:

$$x_i = \frac{exp(p_i)}{\sum_{k=1}^{4} exp(p_i)} \tag{4}$$

where the p_i denote feature representation. The higher x_i, the greater contribution to the binding affinity prediction result.

By learning these weights, we can fuse the four features representation to get the final embedding representation Z as follows:

$$Z = \sum_{i=1}^{4} x_i \cdot P_i \tag{5}$$

2.6 Drug-target binding affinity prediction

In the study, the ultimate output is the binding affinity $\hat{y}$ predicted by the model between drug-target pairs. The representations from the four channels, which are fused by the attention mechanism in Sect. 2.5, are combined into the downstream fully connected layer.

$$\hat{y} = \delta(W_{out}Z + b_{out}) \tag{6}$$

where δ is the sigmoid function, W_{out} and b_{out} are the learnable parameters, and $\hat{y}$ is the predicted label. We employed the mean squared error (MSE) as our loss function, defined as follows:

$$Loss_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{7}$$

where y_i is the ground truth score of i-th drug-target pair, and n is the sample size.

3 Experiments

3.1 Datasets

We evaluated our model using two different datasets, the Kinase dataset Davis [22] and KIBA [23] datasets. The Davis dataset includes selective assays for kinase protein families and their respective dissociation constants (K_d) for related inhibitors. The KIBA dataset originates from a method called KIBA,

Table 1 Sumary of the benchmark datasets

Datasets	Proteins	Compounds	Interactions
Davis	442	68	30,056
KIBA	229	2111	118,254

Table 2 Hyperparameters of MultiDTA

Parameter	Setting
Max length (drug)	100
Max length (protein)	1000
Embedding size	128
Batch size	256
Optimizer	Adam
Learning rate (lr)	0.001
Epoch	2000
Layer of GCN	3
Layer of CNN	2,3,4
Input size of LSTM	114
Hidden size of LSTM	114
Activation function	ReLU

which combines the biological activity of kinase inhibitors from different sources (such as K_i, K_d, and IC_{50}). According to previous work [7, 10], they are widely used as a standard dataset for binding affinity prediction. Table 1 shows the specific details of these two datasets.

3.2 Setting

The MultiDTA is built by PyTorch, an open-source machine learning framework. To improve the accuracy and reliability of the experimental evaluation results, we utilized a five-fold cross-validation approach. This involved dividing the dataset into five evenly sized subsets, with each subset used for both training and testing. By performing training and testing on each subset, we were able to obtain a more comprehensive assessment of our model's performance. The final experimental results represent the average of 10 trials, which were conducted to minimize the impact of variations in dataset partitioning. Table 2 provides details of the settings. With a relatively small range of parameter adjustments, we obtained the best performance of the framework. Our experiments were run on the Windows platform Intel(R)Core(TM)i5-10400F CPU@2.90 GHz and NVIDIA GeForce RTX 3080(10 GB).

3.3 Evaluation metrices

We use several metrics to evaluate the performance of our model, and these are calculated as follows.

- **Concordance Index (CI)**: CI is mainly used to calculate and analyze the difference between the predicted and true values, calculated by Equation (8):

$$CI = \frac{1}{Z} \sum_{d_x > d_y} h(b_x - b_y) \tag{8}$$

here b_x and b_y denote the predicted values of larger affinity d_x and smaller affinity d_y, respectively. Z is the normalization constant, $h(x)$ is the leap function [10]. As Equation (9):

$$h(x) = \begin{cases} 0 & x < 0 \\ 0.5 & x = 0 \\ 1 & x > 0 \end{cases} \tag{9}$$

- **Mean Squared Error (MSE)**: the mean squared error is a common measure of the difference between the predicted and true values. It represents the average difference between the predicted and actual output values. A smaller mean squared error means that the predicted value of the sample is closer to the true value, as in Equation (10):

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (p_i - y_i)^2 \tag{10}$$

the p_i denote predicted value and the y_i denote lable value.

- **R squared** [24, 25]: The R squared is a statistical measure used to evaluate the goodness of fit of a regression model. It represents the percentage of variance in the dependent variable that can be explained by the model and ranges from 0 to 1. When the R squared value is 1, it indicates a perfect fit of the model to the data, while a value of 0 means the model is unable to explain any variation in the dependent variable. Generally, a higher R squared value indicates a better fit of the model. Mathematically it can be expressed by Eq.n (11):

$$r_m^2 = r^2 * (1 - \sqrt{r^2 - r_0^2}) \tag{11}$$

here r^2 and r_0^2 are the values of the squared correlation coefficient between the observed and predicted values with and without intercept, respectively. A model is only an acceptable model if the model's value on the test set is greater than 0.5.

3.4 Baseline methods

KronRLS [10]: The model uses traditional machine learning methods based on Kronecker regular least squares.

SimBoost [7]: The model constructs three features and trains the gradient enhancer model to represent the non-

Table 3 The comparative results with the baseline on both the Davis and KIBA datasets

Method	Davis			KIBA		
	CI	MSE	r_m^2	CI	MSE	r_m^2
KronRLS	0.871	0.379	0.407	0.782	0.411	0.342
SimBoost	0.872	0.282	0.644	0.836	0.222	0.629
DeepDTA	0.878	0.261	0.630	0.863	0.194	0.673
MT-DTI	0.887	0.245	0.665	0.882	0.220	0.584
GANsDTA	0.881	0.276	0.653	0.886	0.224	**0.775**
DeepGS	0.882	0.252	0.686	0.860	0.193	0.684
MultiDTA	**0.893**	**0.231**	**0.694**	**0.890**	**0.156**	0.761

The bold values represent the highest value in each column

linear association between the input features and the bound affinity.

DeepDTA [8]: The model trains two three-layer CNN compound and protein sequence label/one-hot encoding to predict the DTA task. Their CNN model consists of two independent CNN blocks that learn features from a smile string containing compound and protein sequence, respectively. Drug and target representations are concatenated and passed to a fully connected layer for DTA prediction.

GANsDTA [11]: The model presents a GAN-based semi-supervised method to estimate drug-target binding affinity while efficiently learning useful features from labeled unlabeled data. The GAN was used to learn representations from raw sequence data of proteins and drugs and use convolutional regression in predicting affinity.

MT-DTI [9]: The model uses pre-training and proposes a new molecular representation of the self-attention mechanism to predict DTI.

DeepGS [26]: The model uses the GAT and BiGRU to obtain the drug's molecular map topology and local chemical context, respectively. Moreover, it encodes amino acid and SMILES sequences using advanced embedding techniques.

4 Results and discussion

4.1 Comparative experiments with the baseline

We compare our proposed MultiDTA with the state-of-the-art model for DTA prediction and perform comparative experiments under the same conditions. In Table 3, we give the *CI*, *MSE* and r_m^2 scores on the Davis and KIBA datasets.

4.1.1 Comparison of results

It can be seen from Table 3 that although the DeepGS model has achieved good results, the model we extracted is completely superior to the DeepGS in three indicators. Specifically, we can see that the CI, MSE, and r_m^2 of MultiDTA are 0.893, 0.231, and 0.694, respectively, on the dataset of Davis, which is higher than other prediction methods. On the KIBA dataset, our model is also competitive. The CI of MultiDTA reaches 0.0890, and the MSE of MultiDTA reaches 0.156, which is still higher than all the baselines. But, our r_m^2 is 0.761, only 0.014 lower than the best model value.

Table 4 Results of ablation experiments on the Davis dataset

	MSE	*CI*	R_m^2
MultiDTA-drug_seq	0.233	**0.893**	**0.694**
MultiDTA-protein_seq	0.232	0.889	0.670
MultiDTA-drug_gra	0.245	0.886	0.652
MultiDTA-protein_gra	0.254	0.883	0.656
MultiDTA	**0.231**	**0.893**	**0.694**

The bold values represent the highest value in each column

Table 5 Results of ablation experiments on the KIBA dataset

	MSE	*CI*	R_m^2
MultiDTA-drug_seq	0.163	0.889	0.756
MultiDTA-protein_seq	0.168	0.886	0.760
MultiDTA-drug_gra	0.178	0.863	0.751
MultiDTA-protein_gra	0.189	0.872	0.746
MultiDTA	**0.156**	**0.890**	**0.761**

The bold values represent the highest value in each column

4.1.2 Analysis of experimental results

Through comparative experiments, we have consistently observed that graph-based methods, such as GANsDTA [11] and DeepGS [26], tend to outperform sequence-based methods, including KronRLS [10], SimBoost [7], DeepDTA [8] and MT-DTI [9]. These findings underscore the critical importance of three-dimensional modeling for both drug molecules and proteins. This modeling approach allows

Table 6 The experimental results of the attention module ablation on the Davis dataset

	MSE	*CI*	R_m^2
MultiDTA-noattention	0.2404	0.8851	0.6835
MultiDTA	**0.2312**	**0.8932**	**0.6946**

The bold values represent the highest value in each column

Table 7 The experimental results of the attention module ablation on the KIBA dataset

	MSE	*CI*	R_m^2
MultiDTA-noattention	0.1591	0.8837	0.7527
MultiDTA	**0.1565**	**0.8903**	**0.7611**

The bold values represent the highest value in each column

the model to capture topological structural information, enriching their representations and leading to more precise predictions of Drug-Target Interactions (DTA). Subsequent ablation experiments and attention visualizations further corroborate this conclusion.

Simultaneously, we have also discovered that our proposed model consistently achieves more competitive performance across almost all metrics on both datasets when compared to using sequence-based or graph-based methods in isolation. This highlights the effectiveness of our approach, which combines information from four distinct channels with an attention mechanism. Through this method, we successfully integrate the advantages of sequence-based and graph-based approaches, allowing us to extract advanced representations embedded within the sequences of drug-target pairs and their three-dimensional topological structural information. As a result, we obtain a more comprehensive representation. We firmly believe that this discovery will provide a fresh perspective for researchers in the field of DTI.

4.2 Ablation study

To explore the impact of each part of the input in the model on the results, we performed ablation experiments on each part of the input on the Davis and KIBA dataset. Tables 4 and 5 shows the results of three evaluation metrics of ablation experimental models and the original model on the Davis and KIBA datasets.

The MultiDTA-drug_seq indicates the input without the drug sequence, MultiDTA-protein_seq indicates the input without the protein sequence, MultiDTA-drug_graph indicates the input without the drug graph, and MultiDTA-protein_graph indicates the input without the protein graph. It can be seen from the experimental results of Table 4 that the best result in the ablation model is MultiDTA-drug_seq, CI, MSE, and R_m^2 reached 0.233, 0.893, and 0.694, respectively, which is only close to the complete model we proposed. Nevertheless, other ablation models are significantly inferior to MultiDTA. From the experimental results, the drug sequence has the least impact on the model. This is because SMILES strings only provide simplified chemical representations of drug molecules, And the same drug molecule usually has multiple SMILES representations. Compared with the topological structure of the drug, the chemical information carried by the drug sequence is limited, so the drug sequence has less impact on the model performance. Also from the experimental results in Table 5, it can be seen that our full model completely outperforms the ablation experimental model. The results show of ablation experiments that each part of the model input is effective for the prediction results.

We conducted ablation experiments on the attention module in two datasets to prove the effectiveness of our model's attention mechanism fusion. MultiDTA-noattention is a variant of our model with attention removed. Tables 6 and 7 shows the performance comparison of MultiDTA and MultiDTA-noattention on two datasets.

4.3 Interpretability analysis

To investigate the interpretability of the model, we conduct an analysis of the attention weights and visualized their values. We extract the value x_i obtained from Formula 4 in the experiment and then visualize it. As shown in Fig. 2, the experiment was conducted on two different datasets, where the darker the color is, the higher the weight gets.

From Fig. 2, it can be seen that for different datasets, the four different representations contribute differently to the DTA prediction. In general, the input represented by

Fig. 2 Visualization of Attention Weights Learned from the Davis and KIBA Datasets. x1 and x2 respectively represent the drug and protein representations used for extracting graph structural information, while x3 and x4 represent the drug and protein representations used for extracting sequence information. Darker colors indicate higher weights assigned to the respective channels

the graph structure contributes more to the final result. On the KIBA dataset, x2 has the highest contribution to the results, while on the contrary, x1 has the highest contribution on the Davis dataset.

The different visualization results of the two datasets are caused by the different structures of the datasets. The Davis dataset contains a considerably larger number of proteins than drugs, with each drug interacting with multiple proteins. This characteristic is advantageous in studying the diversity and specificity of drugs. Analyzing the interactions of drugs with various proteins enables the efficient acquisition of valuable information regarding their structures, properties, and mechanisms of action. Conversely, the KIBA dataset has a much larger number of drugs than proteins, with each protein containing multiple interaction data information. This characteristic is better suited for extracting protein information. Therefore, in different datasets, different visualization results are presented.

5 Limitations

Although our model has shown good performance by utilizing the graphical structure of protein data as input, it's worth noting that the highly precise 3D structures of proteins account for only a small fraction of known protein sequences. Protein graph structures are more complex and subject to more significant changes than protein sequences. Typically, a protein's graph structure consists of multiple amino acid residues, each with several degrees of freedom. Consequently, the computational complexity of calculating the graph structure is high and requires substantial computing resources. This can pose challenges to significantly improving prediction accuracy, and in the future, more effective methods will be required to obtain protein graph structures.

6 Conclusion

The paper introduces a novel end-to-end deep learning framework, MultiDTA, for Drug-Target Affinity (DTA) prediction. The model takes drug and protein sequences as well as graph structures as inputs and leverages the two-dimensional and three-dimensional spatial structures of drugs and proteins for DTA prediction. Furthermore, we have designed an attention-based fusion mechanism, enabling the model to learn the contributions from four channels. Experimental results demonstrate that our model exhibits strong competitiveness in terms of performance compared to other state-of-the-art DTA prediction models. Despite employing relatively simple networks such as CNN and GCN, our results surpass those of other models. Looking ahead, we can further enhance our model's performance by utilizing more advanced networks like Graph Neural Networks and Transformers.

Acknowledgements This work is supported by grant from the Natural Science Foundation of China (No. 62072070).

Data availability The code of MultiDTA and the data are available at: https://github.com/dengjiejin/MultiDTA.

References

1. Ezzat A, Wu M, Li X-L, Kwoh C-K (2019) Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. Brief Bioinform 20(4):1337–1357
2. Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J, Zhang Y (2016) Drug-target interaction prediction: databases, web servers and computational models. Brief Bioinform 17(4):696–712
3. Gao KY, Fokoue A, Luo H, Iyengar A, Ping Z (2018) Interpretable drug target prediction using deep neural representation. In: Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18
4. Öztürk H, Ozkirimli E, Özgür A (2016) A comparative study of smiles-based compound similarity functions for drug-target interaction prediction. BMC Bioinformatics 17(1):1–11
5. Wang L, You Z-H, Chen X, Xia S-X, Liu F, Yan X, Zhou Y, Song K-J (2018) A computational-based method for predicting drug-target interactions by using stacked autoencoder deep neural network. J Comput Biol 25(3):361–373
6. Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, Lu H (2017) Deep-learning-based drug-target interaction prediction. J Proteome Res 16(4):1401–1409
7. He T, Heidemeyer M, Ban F, Cherkasov A, Ester M (2017) Simboost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. Journal of cheminformatics 9(1):1–14
8. Öztürk H, Özgür A, Ozkirimli E (2018) Deepdta: deep drug-target binding affinity prediction. Bioinformatics 34(17):821–829
9. Shin B, Park S, Kang K, Ho JC (2019) Self-attention based molecule representation for predicting drug-target interaction. In: Machine Learning for Healthcare Conference, pp. 230–248. PMLR
10. Pahikkala T, Airola A, Pietilä S, Shakyawar S, Szwajda A, Tang J, Aittokallio T (2015) Toward more realistic drug-target interaction predictions. Brief Bioinform 16(2):325–337
11. Zhao L, Wang J, Pang L, Liu Y, Zhang J (2020) Gansdta: Predicting drug-target binding affinity using gans. Front Genet 10:1243
12. Zhao Q, Xiao F, Yang M, Li Y, Wang J (2019) Attentiondta: prediction of drug–target binding affinity using attention model. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 64–69. IEEE
13. Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S (2021) Graphdta: predicting drug-target binding affinity with graph neural networks. Bioinformatics 37(8):1140–1147

14. Cheng Z, Yan C, Wu F-X, Wang J (2021) Drug-target interaction prediction using multi-head self-attention and graph attention network. IEEE/ACM Trans Comput Biol Bioinf 19(4):2208–2218
15. Li M, Lu Z, Wu Y, Li Y (2022) Bacpi: a bi-directional attention neural network for compound-protein interaction and binding affinity prediction. Bioinformatics 38(7):1995–2002
16. Weininger D (1988) Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. J Chem Inf Comput Sci 28(1):31–36
17. Liu T, Bao J, Wang J, Zhang Y (2018) A hybrid cnn-lstm algorithm for online defect recognition of co2 welding. Sensors 18(12):4369
18. Wigington C, Stewart S, Davis B, Barrett B, Price B, Cohen S (2017) Data augmentation for recognition of handwritten words and lines using a cnn-lstm network. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 1:639–645 . IEEE
19. Wu C, Wu F, Chen Y, Wu S, Yuan Z, Huang Y (2018) Neural metaphor detecting with cnn-lstm model. In: Proceedings of the Workshop on Figurative Language Processing, pp. 110–114
20. Landrum G (2010) Rdkit: open-source cheminformatics. release 2014.03. 1. arXiv preprint, 1908
21. Zhao L, Akoglu L (2019) Pairnorm: Tackling oversmoothing in gnns. arXiv:1909.12223
22. Anastassiadis T, Deacon SW, Devarajan K, Ma H, Peterson JR (2011) Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. Nat Biotechnol 29(11):1039–1045
23. Tang J, Szwajda A, Shakyawar S, Xu T, Hintsanen P, Wennerberg K, Aittokallio T (2014) Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J Chem Inf Model 54(3):735–743
24. Roy K, Chakraborty P, Mitra I, Ojha PK, Kar S, Das RN (2013) Some case studies on application of rm2 metrics for judging quality of quantitative structure-activity relationship predictions: emphasis on scaling of response data. J Comput Chem 34(12):1071–1082
25. Pratim Roy P, Paul S, Mitra I, Roy K (2009) On two novel parameters for validation of predictive qsar models. Molecules 14(5):1660–1701
26. Lin X (2020) Deepgs: Deep representation learning of graphs and sequences for drug-target binding affinity prediction. arXiv:2003.13902