Neural Networks 169 (2024) 623–636

Contents lists available at ScienceDirect

Neural Networks

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

Full Length Article

# AttentionMGT-DTA: A multi-modal drug-target affinity prediction using graph transformer and attention mechanism

Hongjie Wu [a], Junkai Liu [a,b], Tengsheng Jiang [c], Quan Zou [b], Shujie Qi [a], Zhiming Cui [a], Prayag Tiwari [d,*], Yijie Ding [b,*]

[a] School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, 215009, China
[b] Yangtze Delta Region Institute(Quzhou), University of Electronic Science and Technology of China, Quzhou, 324003, China
[c] Gusu School, Nanjing Medical University, Suzhou, 215009, China
[d] School of Information Technology, Halmstad University, Sweden

ARTICLE INFO

*Keywords:*
Drug–target affinity
Graph neural network
Graph transformer
Attention mechanism
Multi-modal learning

ABSTRACT

The accurate prediction of drug-target affinity (DTA) is a crucial step in drug discovery and design. Traditional experiments are very expensive and time-consuming. Recently, deep learning methods have achieved notable performance improvements in DTA prediction. However, one challenge for deep learning-based models is appropriate and accurate representations of drugs and targets, especially the lack of effective exploration of target representations. Another challenge is how to comprehensively capture the interaction information between different instances, which is also important for predicting DTA. In this study, we propose AttentionMGT-DTA, a multi-modal attention-based model for DTA prediction. AttentionMGT-DTA represents drugs and targets by a molecular graph and binding pocket graph, respectively. Two attention mechanisms are adopted to integrate and interact information between different protein modalities and drug-target pairs. The experimental results showed that our proposed model outperformed state-of-the-art baselines on two benchmark datasets. In addition, AttentionMGT-DTA also had high interpretability by modeling the interaction strength between drug atoms and protein residues. Our code is available at https://github.com/JK-Liu7/AttentionMGT-DTA.

## 1. Introduction

Drug discovery is a costly and time-consuming process. The typical process of a new drug's approval usually requires US $ 2.8 billion and takes 10 – 15 years (Wouters, McKee, & Luyten, 2020; Yang, Ding, Tang, & Guo, 2021). However, most drugs in clinical trial stages have not been approved and put into the market (Newman & Cragg, 2020). Recently, the identification of interactions between drugs and target proteins plays a vital role in drug discovery, which is a research hotspot and has been extensively studied (Ezzat, Wu, Li, & Kwoh, 2018; Qian, Ding, Zou, & Guo, 2022). Many traditional methods have been employed to predict the interaction of given drug-target pairs as a binary classification task (Bahi & Batouche, 2021; Ding, Tang, & Guo, 2021). However, binding affinity is a continuous value which reflects the binding strength between the drug and target (Ding, Tang, Guo, & Zou, 2022). Thus, the regression task of predicting drug-target affinity (DTA) has also become a key issue in the field of drug discovery and drug repositioning.

With the massive application of biomedical data (Ding, Tiwari, Guo, & Zou, 2022) and improvements in computational resources (Chao & Quan, 2021; Wang, Zhai, Ding, & Zou, 2023; Zhang, Tiwari, et al., 2021), deep learning-based methods (Cloninger & Klock, 2021; Mao, Shi, & Zhou, 2021) have been commonly used in bioinformatics, especially in DTA prediction fields (Wu, Ling, et al., 2021; Zhang, Song, et al., 2021). Generally, deep learning-based models include data preprocessing, a drug feature extraction module, a protein feature extraction module and a prediction module (Kimber, Chen, & Volkamer, 2021). The most widely used one-dimensional (1D) sequence representations of drugs and proteins are the Simplified Molecular Input Line Entry System (SMILES) and amino acid sequences, respectively. For deep learning-based models, the 1D representation vectors are fed into deep neural networks to predict the binding affinity. For instance, Öztürk, Özgür, and Ozkirimli (2018) used convolutional neural network (CNN) as encoders of drugs and targets. Lee, Keum, and Nam (2019) used molecular fingerprints as drug representations and linear layers as drug encoders. In addition, the recurrent neural network

* Corresponding authors.
*E-mail addresses:* hongjiewu@mail.usts.edu.cn (H. Wu), 1737969704@qq.com (J. Liu), 1911042002@post.usts.edu.cn (T. Jiang), zouquan@nclab.net (Q. Zou), 997196224@qq.com (S. Qi), zmcui@usts.edu.cn (Z. Cui), prayag.tiwari@ieee.org (P. Tiwari), wuxi_dyj@csj.uestc.edu.cn (Y. Ding).

https://doi.org/10.1016/j.neunet.2023.11.018
Received 26 June 2023; Received in revised form 29 September 2023; Accepted 7 November 2023
Available online 11 November 2023

(RNN) and long short-term memory (LSTM) are also used as feature extractors in DTA prediction (Karimi, Wu, Wang, & Shen, 2019; Zheng, Li, Chen, Xu, & Yang, 2020).

Although sequence-based methods have achieved remarkable performance in DTA prediction, it is not a natural way to represent drugs, as the molecular structure information may be lost (Wang, Tang, Ding, & Guo, 2021). Currently, molecular graphs have also been widely applied in DTA prediction (Jiang et al., 2020; Li et al., 2020; Nguyen et al., 2020; Torng & Altman, 2019; Tsubaki, Tomii, & Sese, 2018). For example, Nguyen et al. (2020) proposed GraphDTA to improve the prediction effect using various graph neural network (GNN) variant models. On the other hand, graph-based protein representations have also been used to predict DTA (Jiang et al., 2020; Li, Zhang, Guan, & Zhou, 2022a; Nguyen, Nguyen, Le, & Tran, 2022a; Zheng et al., 2020). Jiang et al. (2020) proposed the model named DGraphDTA, utilizing 2D contact maps obtained from protein sequences as the descriptor of proteins to obtain more structural information. Compared with 1D sequence-based methods, 2D graph-based models can leverage more topology information of drugs and proteins, which have shown great advantages in DTA prediction.

Despite recent progress in DTA prediction, there are still two major drawbacks of existing methods limiting performance. Most graph-based models use the 2D contact map or distance map to represent target proteins (Jiang et al., 2020; Li et al., 2020; Zheng et al., 2020). However, those graph representation methods are only an approximate abstraction of the structure of the protein, which are unable to accurately describe the complex three-dimensional (3D) structure of proteins (Ding, Tang, & Guo, 2020a). Thus, the lack of structure awareness limits the accuracy and generalization of the proposed models. To this end, we tend to use both 3D structure-based features and 1D sequence-based features in our method. On the one hand, the development of AlphaFold2 gives us the possibility of 3D structures of proteins for large-scale use. These accurate 3D structures contain a wealth of relevant information on the practicalities and configurations of protein binding pockets, which have certain positive impact on the binding process between drug molecules. On the other hand, the 1D sequence representations can provide contextual and biological information as a complement of 3D structures. We believe that the additional high-order sequence-based embeddings for proteins can provide clear and effective knowledge for our model to distinguish different proteins, by which the DTA prediction model can make full use of these features to measure the associations between both seen and unseen proteins in the dataset and thus improve the performance and generalization.

Furthermore, when integrating the 1D sequence-based features and graph-based features, existing methods often utilize global pooling and simple concatenation to form the final representation (Wang, Zheng, et al., 2022; Wu, Gao, Zeng, Zhang, & Li, 2022). Such operations are also very common in the incorporation and fusion of the drug encoder's and protein encoder's representation, which ignore complex multi-modal interactivity. In addition, the single concatenation process is not able to predict which protein binding sites contribute most to binding with the given drug. Thus, a cross-attention module was leveraged in our method to fuse and incorporate information from intra-modalities of protein information, and a joint attention mechanism was used to interact information from inter-modalities of protein and drug features.

To alleviate the abovementioned problems, we propose a novel multi-modal deep learning method named AttentionMGT-DTA for DTA prediction. Firstly, We construct drug graphs and protein pocket graphs with node and edge features, which are based on 2D molecular structure and 3D protein folding, respectively. Afterwards, the drug and protein graphs are fed into encoders to obtain the feature embeddings. The drug encoder contains a graph transformer architecture, while the protein encoder also includes a modality interaction module with 1D sequence-based embeddings. Ultimately, the extracted feature vectors are inputted into an attention-based prediction module for completing information fusion between diffierent instances and predicting DTAs. The main contributions of our work are summarized as follows:

- Our proposed model is the *first large-scale application of the predicted protein structures from AlphaFold2 in DTA prediction*. We constructed a residue-level protein pocket graph based on the AlphaFold Database to represent target proteins. Then two graph transformers were utilized for the feature extraction of protein pocket graphs and drug molecular graphs, respectively.
- To enrich the protein representations, we introduced self-supervised pretrained embeddings of protein amino acid sequences. To the best of our knowledge, our proposed model is the *first to use cross-attention to integrate information from two modalities of proteins, i.e., 1D sequences and 3D graphs*.
- We introduced a joint-attention mechanism to generate affinity results with the atom-residue interaction matrix. This allowed our model to obtain the interactive drug-target pair embedding and be *highly interpretable*, and learn which residues of proteins interact with the drug atoms.

The rest of this paper is organized as follows. In Section 2, related works associated to our study are presented, including the development in DTA prediction and the research in protein representation method. Section 3 presents our proposed AttentionMGT-DTA framework, which includes the description of drug molecule graph construction, protein pocket graph construction, graph transformer model, protein intra-modality interaction, and drug- protein inter-modality interaction. In Section 4, experimental results are represented and discussed. Ultimately, conclusions and future directions are given in Section 5.

## 2. Related work

In this section, we introduce related works from two aspects. First, we briefly review the existing researches and technologies for DTA prediction. Moreover, current development of protein representation methods in this field is also presented.

### 2.1. DTA prediction methods

#### 2.1.1. Deep learning-based DTA prediction methods

Recently, deep learning techniques have developed rapidly and achieved great success in computational methods for DTA prediction. Currently, most deep learning-based models for DTA prediction directly use drug and protein representations as features, encode them using various types of deep learning models, extract information from them, and combine the respective representations to predict binding affinities (Dhakal, McKay, Tanner, & Cheng, 2021; Ding, Tang, & Guo, 2020b). Öztürk et al. (2018) first applied CNN into DTA prediction. In their proposed DeepDTA model, CNNs were used to extract low-dimensional features of drugs and proteins, respectively. The obtained feature vectors were fed into a fully connected layer to calculate the binding affinity. Later, they added new features, including motif and domain information on proteins, and the proposed WideDTA (Öztürk, Ozkirimli, & Özgür, 2019) obtained better performance compared to DeepDTA on two benchmark datasets. Lee et al. (2019) proposed DeepConv-DTI, using CNN to convolve amino acid sequences of different lengths as an approach to obtain local residue patterns of proteins. Rifaioglu et al. (2020) first used a 2D image structure to represent the drug, reducing the information loss during data conversion, and the proposed DEEPScreen model also achieved satisfactory performance. Wang, Zhou, Li, and Li (2021) proposed DeepDTAF, which incorporated protein binding pockets as input features to integrate local and global contextual information. Karimi et al. (2019) proposed DeepAffinity, taking into account the possible dependencies between residues or atoms, based on a seq2seq autoencoder architecture combining CNN and RNN. Li, Zhao, and Li (2022) proposed a CO-VAE method, leveraging a novel co-regularized variational autoencoders to generate drug SMILES strings and target sequences, and a co-regularization part was employed to obtain the binding affinities. The above methods have shown promising prediction performances of the models.

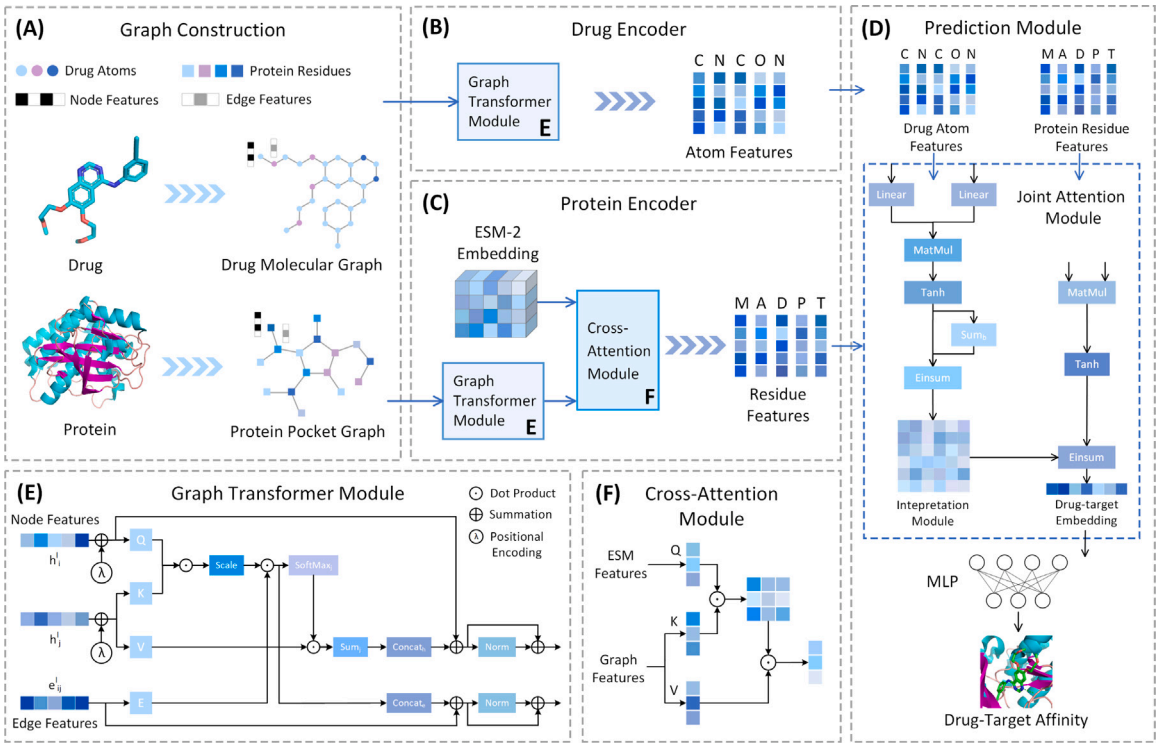


**Fig. 1.** Overall architecture of AttentionMGT-DTA. (A) Graph construction module, where we construct graph representations of drugs and proteins, including finding proteins' binding pockets. (B) Drug encoder module to extract drug features by the graph transformer. (C) Protein encoder module to extract protein features by using two different modalities of information. (D) Prediction module to predict the binding affinity of a drug-target pair, which includes an interpretation module to provide a deep explanation of which protein residues bind to the drug atoms. (E) The network structure of graph transformer in (B) and (C). (F) The detailed cross-attention module in (C).

### 2.1.2. GNN-based DTA prediction methods

Traditional CNN and RNN models represent drugs as strings of data format, the structural information of molecules may be lost as a result, which may reduce the predictive power of the model and the functional relevance of the learned potential space (Bagherian et al., 2020). Drugs and proteins can be naturally represented as graph structures with atom-level or residue-level nodes and edges between the nodes, and GNN updates these node features while considering the neighbors of each node to extract the global structural features (Zhang et al., 2022). For instance, Tsubaki et al. (2018) first applied GNN in DTA prediction, using GNN and CNN to extract features of compounds and proteins, respectively. Torng and Altman (2019) introduced graph convolutional networks into DTI prediction. In their approach, residues in protein binding pockets correspond to nodes and the calculated feature vectors represent their physicochemical properties, while drug molecules are also converted into graph structures. Feng, Dueva, Cherkasov, and Ester (2019) proposed PADME, which used ECFP encoding as well as graph representation, combined with protein feature vectors. MGraphDTA proposed by Yang, Zhong, Zhao, and Yu-Chian Chen (2022) introduced dense connectivity into GNNs and constructed an ultra-deep network structure to simultaneously capture the local and global structures of drugs. For target proteins, Jiang et al. (2020) first proposed a contact map-based protein graph method DGraphDTA, in which the researchers set a threshold of 0.5 to determine whether the residue nodes are connected to each other based on the contact probability obtained, and the results obtained correspond to the adjacency matrix of the protein. The research proposed by Zheng et al. (2020) used the 2D distance map of proteins as input and a Visual Question Answering system with a linear representation of drug molecules as query conditions to obtain the answer to whether the queried drug and protein interact with each other. Inspired by the above work, GEFA (Nguyen, Nguyen, Le, & Tran, 2022b), PSG-BAR (Pandey et al., 2022), and STAMP-DPI (Wang, Zheng, et al., 2022) also used contact maps to construct protein graphs and added pretrained language model embeddings as the node features. The above studies jointly show that the structure of drug molecules, and protein residues can be effectively expressed using graph data structures.

### 2.1.3. Attention mechanism-based DTA prediction methods

Attention mechanisms are gradually becoming more important in deep learning, including powerful NLP representation models such as transformer (Vaswani et al., 2017) and BERT (Devlin, Chang, Lee, & Toutanova, 2019), and are widely valued in bioinformatics. AttentionDTA (Zhao, Xiao, Yang, Li, & Wang, 2019) first correlated attention mechanisms with the binding affinity, using attention mechanisms to consider which subsequences in a protein are more significant for the drug and which subsequences in the drug are more significant for the protein, thus making the model more expressive. Chen et al. (2020) proposed TransformerCPI, a transformer-based model which managed some specific pitfalls in sequence-based models using transformer decoders to extract data features. Huang, Xiao, Glass, and Sun (2020) proposed MolTrans, which was an interpretable model extracting the interactions between substructures from unlabeled data by two transformer modules. CoaDTI proposed by Huang et al. (2022) also leveraged the transformer as the encoder for protein sequences to obtain the global representation at the amino acid level. Li, Zhang, Guan, and Zhou (2022b) improved the transformer by introducing the interformer, where two interacting transformer decoders were used to extract feature vectors of targets and drugs. ML-DTI (Yang, Zhong, Zhao, & Chen, 2021) devised a mutual learning mechanism based on multihead attention to facilitate the interaction between drug and protein encoders. MGPLI proposed by Wang, Hu, et al. (2022) utilized the transformer encoders to represent different-level features, completing a multi-granularity model with competitive performance. Zhao, Zhao, Zheng, and Wang (2021) proposed HyperAttentionDTI, adopting the attention mechanism to model complex inter-molecular interactions among drug atoms and protein amino acids.

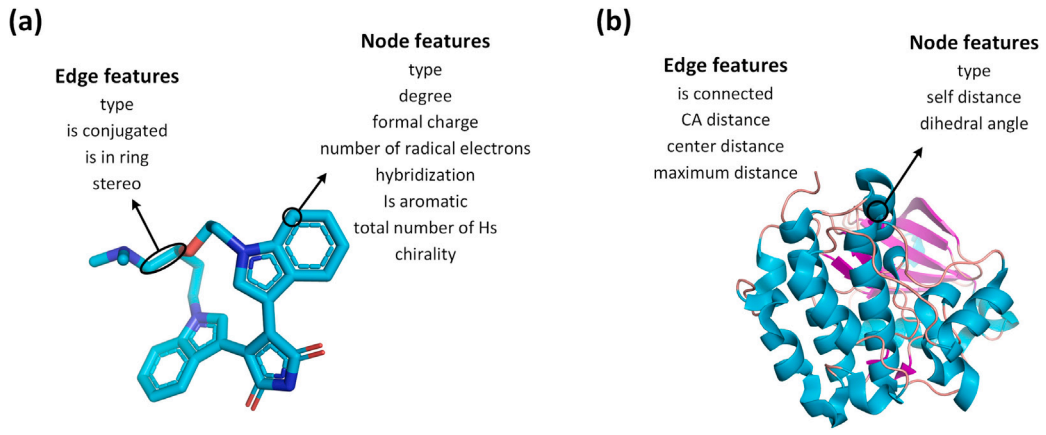


**Fig. 2.** The node and edge feature representations of drugs and proteins. (a) Atom-level drug molecular graph, which contains eight node features and four edge features. (b) Residue-level protein pocket graph, which contains three node features and four edge features.

The above study illustrates that methods based on GNN can effectively extract topological information from both drug molecules and protein residues with advanced and well-designed GNN models. The rich chemical and biological information contained in entities can be fully utilized to enhance the performance. Furthermore, the introduction of attention mechanism can promote the interpretability and generalization to some extents, which has also attracted considerable interest in this research field.

### 2.2. Protein representation methods

Proteins are organic macromolecules involved in various life activities and are composed of different amino acid sequences, which in turn form their unique 3D folding structures, resulting in orderly to disordered and conformational changes (Ding, Guo, Tiwari, & Zou, 2023). Understanding the sequence-structure-function relationship of proteins is a central issue in protein biology and is essential for investigating disease mechanisms, protein design and drug discovery (Bepler & Berger, 2021). The commonly used methods for protein data representation include 1D strings, 2D, and 3D graph structures.

The complete sequence of a protein, often referred to as the primary structure of a protein, is the order in which amino acids are arranged in a protein. All 20 types of amino acids in the protein primary structure can be encoded as a single letter, so one-hot encoding is usually used to represent the protein sequence (ElAbd et al., 2020). One-hot encoding converts the characters representing the amino acid sequence into a binary vector, which can represent the protein primary structure briefly and efficiently. Based on this technology mentioned above, some models have exploited DTA prediction methods which represent target proteins as 1D sequences (Chen et al., 2020; Karimi et al., 2019; Nguyen et al., 2020; Öztürk et al., 2018).

Although sequence is an effective way to store information about the primary structure of a protein, it cannot provide information on the 3D structure of a protein. For this reason, the protein structure can usually be converted into a spatial graph with chemical properties, which contains atomic or residue nodes and edges. Currently, the widely used protein graph representations include 2D and 3D graphs. Jiang et al. (2020) first proposed the utilization of the 2D contact map of proteins as its representation. Each protein includes hundreds of amino acid residues; however, the connection between residues is only a long chain, which does not contain any spatial information. The contact map is a representation of the protein structure, which is a 2D representation of the 3D structure of the protein and can also be used as an output for protein structure prediction. In this method, a threshold of 0.5 is set to obtain a contact map of $L \times L$, where $L$ is the number of nodes (residues). On the other hand, a major issue with the Protein Data Bank (PDB), the main source of protein data, is that the number of proteins with structural features is much smaller than the number of proteins with determined amino acid sequences (Li et al., 2020), by which the use of protein structural information for drug discovery is thus limited (Zhang et al., 2023). Hence, only a few methods have explored and developed protein graph representation methods based on 3D structures (Wang, Liu, et al., 2023; Wang, Zhang, et al., 2023; Yazdani-Jahromi et al., 2022). In this study, we introduce a protein representation method based on the 3D structure of proteins, with large-scale utilization and exploitation of protein structures predicted by AlphaFold2, which is the first time to the best of our knowledge to the DTA prediction problem with the ultimate motivation of finding an effective and accurate protein representation method and improve the performance.

## 3. Materials and methods

### 3.1. Graph construction

#### 3.1.1. Drug graph representation

The drug molecule was represented as a graph to obtain more chemical and topology information. The 2D undirected graphs of the drugs can be described as $G_D = (V_D, E_D)$, where nodes and edges represent drug atoms and covalent chemical bonds, respectively. In the drug graph, $V_D$ is the set of atom nodes with the feature vectors, and $E_D$ is the set of edges represented as the feature vectors. Fig. 2 illustrates the node and edge features for constructing drug graphs in our method, including eight types of node features and four types of edge features, which have been applied in previous studies to represent drug molecules (Jiang et al., 2021; Wu, Jiang, et al., 2021).

#### 3.1.2. Protein graph representation

The 3D structure of proteins plays crucial roles in the binding between drugs. Nevertheless, partial protein's crystallographic structure is unavailable in PDB, causing difficulties in utilizing this structure. More specifically, the number of available and unavailable protein PDB structure in the datasets can be found in Table 3. Considering the precision and accuracy of the 3D protein structure predicted by AlphaFold2 (Jumper, Evans, Pritzel, Green, & Hassabis, 2021), we employed the high-quality 3D structure of the protein based on AlphaFold2 to replace PDB structure. Each of the protein structure files in this study were downloaded by UniProt ID from the AlphaFold Database (Varadi et al., 2021), which were in the PDB file format. We demonstrated that the high-quality protein structure predicted by AlphaFold2 can also represent target proteins effectively and improve

**Table 1**
The node and edge features for drug graph representation.

| Type | Feature | Description | Dimension |
|---|---|---|---|
| Node features | Atom type | 'C', 'N', 'O', 'F', 'P', 'S', 'Cl', 'Br', 'I', 'B', 'Si', 'Fe', 'Zn', 'Cu', 'Mn', 'Mo', 'other' (one-hot) | 17 |
| | Atom degree | 0, 1, 2, 3, 4, 5, 6 (one-hot) | 7 |
| | Atom formal charge | 0 or 1 | 1 |
| | Atom num radical electrons | 0 or 1 | 1 |
| | Atom hybridization | 'sp', 'sp2', 'sp3', 'sp3d', 'sp3d2', 'other' (one-hot) | 6 |
| | Atom is aromatic | 0 or 1 | 1 |
| | Atom total num H | 0, 1, 2, 3, 4 (one-hot) | 5 |
| | Atom chirality | 'R', 'S', 'other' (one-hot) | 3 |
| Edge Features | Bond type | 'SINGLE', 'DOUBLE', 'TRIPLE', 'AROMATIC' (one-hot) | 4 |
| | Bond is conjugated | 0 or 1 | 1 |
| | Bond is in ring | 0 or 1 | 1 |
| | Bond stereo | 'STEREONONE', 'STEREOANY', 'STEREOZ', 'STEREOE' (one-hot) | 4 |

**Table 2**
The node and edge features for protein graph representation.

| Type | Feature | Description | Dimension |
|---|---|---|---|
| Node Features | Residue type | 'G', 'A', 'V', 'L', 'I', 'P', 'F', 'Y', 'W', 'S', 'T', 'C', 'M', 'N', 'Q', 'D', 'E', 'K', 'R', 'H', 'metal', 'other' (one-hot) | 22 |
| | Residue self distance | max and min values of the scaled distance of all atoms in a residue, the scaled distance between CA and O atoms, O and N atoms, C and N atoms | 5 |
| | Residue dihedral angle | 'phi', 'psi', 'omega', 'chi1' | 4 |
| Edge Features | Residue is connected | 0 or 1 | 1 |
| | Residue CA distance | scaled distance between the CA atoms | 1 |
| | Residue center distance | scaled distance between the center | 1 |
| | Residue max distance | max and min values of the scaled distance | 2 |

DTA prediction performance which enable our model to outperform several baselines.

Subsequently, the method proposed by Saberi Fathi and Tuszynski (2014) was introduced to identify the binding pockets of proteins. The bounding box coordinates calculated by the algorithm can be utilized to determine the binding pockets parts of the complete protein structure, which helped our model reduce the input data size and save computational resources. Moreover, although these determined pockets do not always correspond to the exact binding site, the protein graph representation using protein pockets can facilitate our method to be highly generalizable by directing our model to concentrate on generic topological features in pockets. This property is essential to be generalized reasonably to those scenarios where new proteins that are dissimilar to those have been seen in training data. The protein target graphs $G_P = (V_P, E_P)$ were then constructed, where nodes represented protein residues and edges represented interactions between two residues. Here, we set the threshold to 10.0 Å, which means that residue pairs whose minimum distances were less than 10.0 Åwere regarded as connected by edges. The threshold parameter was empirical and optimizable. Considering the model complexity and computation resource consumption, we constructed the protein pocket graph in residue-level instead of atom-level. For the graph features, we selected three node features and four edge features according to the study of Shen et al. (2022), as depicted in Fig. 2. Table 1 and 2 illustrate the list of drug and protein features in detail. We generated the protein features mainly through the MDAnalysis package (Michaud-Agrawal, Denning, Woolf, & Beckstein, 2011).

### 3.2. Drug encoder module

In our model, we used a graph transformer (Dwivedi & Bresson, 2020) framework as the drug encoder to extract the node representations of the drug graph, as shown in Fig. 1. The advanced graph transformer model has the following advantages compared with other GNN models: Firstly, the graph transformer with edge features can allow the introduction of more explicit chemical and biological property as edge features in DTA prediction, i.e., covalent chemical bonds of drugs and residue interactions of proteins. Secondly, the architecture of graph transformer model enables our approach to capture both local and global information, including universal and specific patterns of node and connection information. Thirdly, the innovative and improved attention mechanism in this model can facilitate our method to measure the importance of each atom and residue, which is beneficial to identify the exact binding sites. Furthermore, the positional encoding in this model is essential to encode node position information and represent graph topological structure, which is the attribute not found in other GNN models.

For the drug graph $G_D$ with node features $\alpha_i \in R^{d_n \times 1}$ for node $i$, and edge features $\beta_{ij} \in R^{d_e \times 1}$ for edges between node $i$ and node $j$, the input node and edge features were first fed into a linear projection layer to obtain hidden representations $h_i^0$ and $e_{ij}^0$ as follows:

$$h_i^0 = W_A^0 \alpha_i + b_A^0 \tag{1}$$

$$e_{ij}^0 = W_B^0 \alpha_i + b_B^0 \tag{2}$$

where $W_A^0 \in R^{d \times d_n}$, and $W_B^0 \in R^{d \times d_e}$ are the learnable weight parameters and $b_A^0, b_B^0 \in R^d$ are the biases of the linear layer. Then the positional encodings were added to the node features.

$$\lambda_i^0 = W_C^0 \lambda_i + b_C^0 \tag{3}$$

$$\hat{h}_i^l = h_i^0 + \lambda_i^0 \tag{4}$$

In our work, we exploited Laplacian eigenvectors as positional encoding in the graph transformer model (Dwivedi et al., 2020). In particular, the Laplacian eigenvectors of each graph were pre-calculated by the factorization of the graph Laplacian matrix as:

$$\Delta = \mathrm{I} - D^{-1/2} A D^{-1/2} = U^T \Lambda U \tag{5}$$

where $A$ is the adjacency matrix, $D$ is the corresponding degree matrix of the graph, and $\Lambda, U$ denote the eigenvalues and eigenvectors, respectively. The $k$ -smallest non-trivial eigenvectors of the node $i$ are employed as its corresponding positional encoding, which are represented as $\lambda_i$.

The graph transformer update node and edge features were mainly based on the multi-head attention mechanism. The following equations define the detailed update process of the $l$ th layer:

$$Q_{ij}^{k,l} = Q^{k,l} Norm\left(h_i^l\right),\ K_{ij}^{k,l} = K^{k,l} Norm\left(h_j^l\right),\ V_{ij}^{k,l} = V^{k,l} Norm\left(h_j^l\right) \quad (6)$$

$$E_{ij}^{k,l} = E^{k,l} Norm\left(e_{ij}^l\right) \quad (7)$$

$$w_{ij}^{k,l} = softmax_j\left(\left(\frac{Q_{ij}^{k,l} h_i^l \cdot K_{ij}^{k,l} h_j^l}{\sqrt{d_k}}\right) \cdot E_{ij}^{k,l} e_{ij}^l\right) \quad (8)$$

$$\hat{h}_i^{l+1} = h_i^l + O_h^l\left(Concat_{k=1}^{h_{gt}}\left(\sum_{j \in \mathcal{N}_i} w_{ij}^{k,l} V_{ij}^{k,l} h_j^l\right)\right) \quad (9)$$

$$\hat{e}_{ij}^{l+1} = e_{ij}^l + O_e^l\left(Concat_{k=1}^{h_{gt}}\left(w_{ij}^{k,l}\right)\right) \quad (10)$$

where $Q^{k,l}, K^{k,l}, V^{k,l}, E^{k,l} \in R^{d_k \times d}, O_h^l, O_e^l \in R^{d \times d}$ are parameters of the linear layers. $k = 1$ to $h_{gt}$ means the number of attention heads; $d_k$ denotes the dimension of each head. Finally, $\hat{h}_i^{l+1}$ and $\hat{e}_{ij}^{l+1}$ were fed into feed forward networks with residue connections and batch normalization layers as follows:

$$h_i^{l+1} = \hat{h}_i^{l+1} + W_{h2}^l\left(ReLU\left(W_{h1}^l Norm\left(\hat{h}_i^{l+1}\right)\right)\right) \quad (11)$$

$$e_{ij}^{l+1} = \hat{e}_{ij}^{l+1} + W_{e2}^l\left(ReLU\left(W_{e1}^l Norm\left(\hat{e}_{ij}^{l+1}\right)\right)\right) \quad (12)$$

where $W_{h1}^l, W_{e1}^l \in R^{2d \times d}$ and $W_{h2}^l, W_{e2}^l \in R^{d \times 2d}$. Through the aforementioned graph transformer module, the node features, i.e., atom features of drug $X_d$ were obtained for further prediction.

### 3.3. Protein encoder module

Similarly, the graph transformer model is used to encode the protein graphs to learn the residue features. In addition to the 3D biological structure information of the protein pocket graph, the pretrained embeddings from protein sequences were also used in our model to enrich multi-modal protein information representations. More specifically, we introduced ESM-2 (Lin et al., 2023), which is a protein language model for protein structure, function and other properties prediction from individual sequences. The pretrained ESM-2 model was employed to encode amino acid sequences into embedding vectors. Each embedding vector contains a wealth of contextual information from the 1D protein sequence. Furthermore, we obtained the protein graph embeddings which only represent its binding pocket part, while the ESM-2 embeddings represent the full protein sequence. To address this inconsistency, the ESM-2 embeddings were further processed. Specifically, we cut these embeddings and took only the part of them that corresponds to the binding pocket, denoted as $X_{ps} \in R^{n_p \times d_p}$ ($n_p$ is the length of the protein pocket and $d_p$ is the dimension of protein embedding). We believe that using only incomplete pretrained embeddings does not affect their utility, as the embedded fragments also contain local feature information due to the contextual nature.

The concatenation strategy was a simple integration method for preserving the information from individual modalities by encoding and mixing two modalities of embeddings, which was widely employed in previous studies (Wang, Zheng, et al., 2022; Xu, Hu, Leskovec, & Jegelka, 2019). However, different modalities of protein embeddings are correlated with each other and simple concatenation is not sufficient to capture the interaction information and reveal the relationship between them, some essential structural or sequence information may be lost as a result. Hence, we introduced a cross-attention module to perform the modality interaction between 1D sequences and 3D graphs of proteins, as shown in Fig. 1. The introduction of cross-attention mechanism has following benefits: On the one hand, comprehensive and reasonable feature embeddings can be obtained by the attention scores to measure the importance of each residue in the process of binding. On the other hand, this module can fuse and incorporate abundant biological properties from dual modalities, making our model to learn more informative feature representations. In this module, given the embeddings from graph transformer module $X_{pg} \in R^{n_p \times d_p}$ and embeddings from pretrained models $X_{ps} \in R^{n_p \times d_p}$, we calculated the multi-modal output $X_p$ as follows:

$$Q = W_Q X_{ps},\ K = W_K X_{pg},\ V = W_V X_{pg} \quad (13)$$

$$X_p = attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (14)$$

The cross-attention module facilitated our model to learn the relationship between independent protein modalities (1D sequences and 3D structure) by the modality interactions of two encoded protein embeddings. Thus, our model can capture more comprehensive and effective information from the correlated protein modalities, which promote the DTA prediction.

### 3.4. Prediction module

Following the drug and protein encoder, the learned feature embedding vectors of drug $X_d$ and protein $X_p$ were inputted into the prediction network to output the interactive drug-target embedding $X_{dp}$. Instead of concatenating the two representations followed by the multi-layer perceptron (MLP) as in previous studies, we adopted a joint-attention mechanism (Karimi et al., 2019; Karimi, Wu, Wang, & Shen, 2021) to integrate and interact information, as shown in Fig. 1. The drug embedding $X_d \in R^{n_d \times d_d}$ and protein embedding $X_p \in R^{n_p \times d_p}$ were processed as follows to obtain the attention score:

$$N_{ij} = \tanh\left(\left(x_i^d\right)^T W_{A1} x_j^p\right)\ \forall i = 1, \ldots, n_d,\ \forall j = 1, \ldots, n_p \quad (15)$$

$$\alpha_{ij} = \frac{\exp\left(N_{ij}\right)}{\sum_{i',j'} \exp\left(N_{i'j'}\right)} \forall i = 1, \ldots, n_d,\ \forall j = 1, \ldots, n_p \quad (16)$$

where $i$ is the index of the drug atom and $j$ is the index of the protein residue, $W_{A1}$ denotes the weight matrix. The attention scores $\alpha_{ij}$ constitute the attention matrix $A \in R^{d_d \times d_p}$, where each position represents the interaction strength between the $i$ th drug atom and $j$ th protein residue. Then the drug-target representation is embedded by $X_{dp}$ as:

$$f_{ij} = \tanh\left(W_{A2} x_i^d + W_{A3} x_j^p + b_A\right) \quad (17)$$

$$X_{dp} = \sum_{i,j} f_{ij} \alpha_{ij} \quad (18)$$

where $W_{A2}, W_{A3}$ and $b_A$ are learnable parameters.

The calculated interactive drug-target feature was then passed to the MLP to get the final affinity value. In this study, the MLP was composed of three fully connected layers with a ReLU activation function and dropout layer. As DTA prediction is a regression task, the mean square error (MSE) was utilized as the loss function:

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(P_i - Y_i\right)^2 \quad (19)$$

where $n$ means the number of drug-target pairs, $P_i$ and $Y_i$ are the predictive binding affinity value and the ground truth value.

**Table 3**
Summary of the refined benchmark datasets.

| Datasets | Drugs | Proteins | Available PDBs | Unavailable PDBs | Interactions | Active | Inactive |
|---|---|---|---|---|---|---|---|
| Davis | 68 | 361 | 275 | 86 | 24 548 | 1649 | 22 899 |
| KIBA | 2052 | 229 | 194 | 35 | 117 148 | 24 543 | 92 641 |

**Table 4**
Hyperparameter settings of AttentionMGT-DTA.

| Hyperparameter | Setting |
|---|---|
| Threshold of protein residue graph | [5, 8, **10**, 12, 15] |
| Number of graph transformer layers | [2, 3, **5**, 10] |
| Number of attention heads | [1, 2, 4, **8**] |
| Dropout rate of graph transformer | 0.2 |
| Dimension of drug embedding | [32, 64, **128**, 256] |
| Dimension of protein embedding | [32, 64, **128**, 256] |
| Learning rate | 1e − 4 |
| Batchsize | 60 |
| Epoch | 1000 |

**Table 5**
Performance comparison between our model and baseline methods on the Davis dataset.

| Dataset | Model | $r_m^2$ | CI | MSE |
|---|---|---|---|---|
| Davis | DeepDTA | 0.690 (0.035) | 0.882 (0.016) | **0.191** |
| | AttentionDTA | 0.697 (0.005) | 0.888 (0.007) | 0.195 |
| | GraphDTA | 0.682 (0.028) | 0.876 (0.019) | 0.194 |
| | TransformerCPI | 0.658 (0.033) | 0.872 (0.015) | 0.201 |
| | ML-DTI | 0.627 (0.022) | 0.869 (0.009) | 0.196 |
| | MGPLI | 0.620 (0.017) | 0.884 (0.004) | 0.218 |
| | AttentionMGT-DTA | **0.699 (0.027)** | **0.891 (0.005)** | 0.193 |

**Table 6**
Performance comparison between our model and baseline methods on the KIBA dataset.

| Dataset | Model | $r_m^2$ | CI | MSE |
|---|---|---|---|---|
| KIBA | DeepDTA | 0.766 (0.085) | 0.892 (0.026) | 0.152 |
| | AttentionDTA | 0.742 (0.015) | 0.880 (0.001) | 0.158 |
| | GraphDTA | 0.760 (0.049) | 0.888 (0.023) | 0.203 |
| | TransformerCPI | 0.721 (0.022) | 0.875 (0.009) | 0.205 |
| | ML-DTI | 0.764 (0.025) | 0.890 (0.005) | 0.177 |
| | MGPLI | 0.753 (0.016) | 0.891 (0.004) | 0.159 |
| | AttentionMGT-DTA | **0.786 (0.018)** | **0.893 (0.001)** | **0.140** |

## 4. Experiments and results

### 4.1. Datasets

We compared our AttentionMGT-DTA model with other baseline models on two benchmark datasets, the Davis dataset and KIBA dataset (Davis et al., 2011; Tang et al., 2014). The binding affinity value is usually expressed by indicators such as dissociation constant ($K_d$), inhibition constant ($K_i$), and the half maximal inhibitory concentration ($IC_{50}$). The affinity in Davis dataset was evaluated by the $K_d$ value, which reflects the selective measurements with the constant values of dissociation of the kinase protein family and associated inhibitor. For KIBA dataset, the affinity values were measured by a method named KIBA, which uses the statistical information contained in $K_d$, $K_i$ and $IC_{50}$ to optimize the consistency between them. The affinity values in the KIBA dataset are mainly in the range of 10 to 13, with most around 11. Furthermore, the duplicate samples in the original datasets were removed to reduce their impact on model training. Table 3 shows the summary statistics of the refined benchmark datasets.

### 4.2. Experimental settings

In our experiment, AttentionMGT-DTA was implemented by Pytorch. The Adam optimizer (Kingma & Ba, 2017) was used for model training with the learning rate of 0.0001. The learning rate decay strategy was also employed where learning rate reduced by 20% when there was no improvement of the MSE index in test datasets in 50 epochs. We utilized Nvidia RTX 3090 GPU for the experiments. The best settings of hyperparameter optimization are presented in Table 4.

### 4.3. Evaluation metrics

For the regression task of DTA prediction, we used MSE, concordance index (CI) and regression toward the mean ($r_m^2$) to evaluate the performance of our model. CI (Gönen & Heller, 2005) is an evaluation metric which reflects the correctness of the result by calculating the difference between the prediction value and the actual value, as follows:

$$CI = \frac{1}{Z} \sum_{\delta_j > \delta_i} h\left(b_i - b_j\right) \tag{20}$$

$$h(x) = \begin{cases} 0 & x < 0 \\ 0.5 & x = 0 \\ 1 & x > 0 \end{cases} \tag{21}$$

MSE is a commonly used index to measure the error. Given $N$ samples with corresponding prediction affinity value $y_i$ and ground truth affinity value $\hat{y}_i$, MSE is defined as follows:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2 \tag{22}$$

$r_m^2$ is a metric evaluating the external predictive performance. A model was regarded acceptable if and only if $r_m^2 \geq 0.5$. $r_m^2$ is defined as follows:

$$r_m^2 = r^2 * \left(1 - \sqrt{r^2 - r_0^2}\right) \tag{23}$$

where $r$ denotes the squared correlation coefficients between the observed and predicted values with intercepts and $r_0$ is the coefficient without intercepts.

### 4.4. Results

#### 4.4.1. The performance on benchmark datasets

Here, we compared our model with the following baselines: DeepDTA (Öztürk et al., 2018), AttentionDTA (Zhao et al., 2019), GraphDTA (Nguyen et al., 2020), TransformerCPI (Chen et al., 2020), ML-DTI (Yang, Zhong, et al., 2021) and MGPLI (Wang, Hu, et al., 2022), which are state-of-the-art approaches.

- DeepDTA (Öztürk et al., 2018) adopted two CNN modules to extract embeddings from drug SMILES and protein sequence, respectively. Then a MLP was employed for prediction from the concatenated features.
- AttentionDTA (Zhao et al., 2019) introduced attention mechanism to measure the importance of different subsequences in drugs and proteins, which enhanced the representational capability.
- GraphDTA (Nguyen et al., 2020) was a GNN-based method which utilized various types of GNN models to encode drug molecule graphs and a CNN to encode protein sequence.
- TransformerCPI (Chen et al., 2020) proposed a modified transformer architecture to complete information interaction between drug and protein sequential representations.

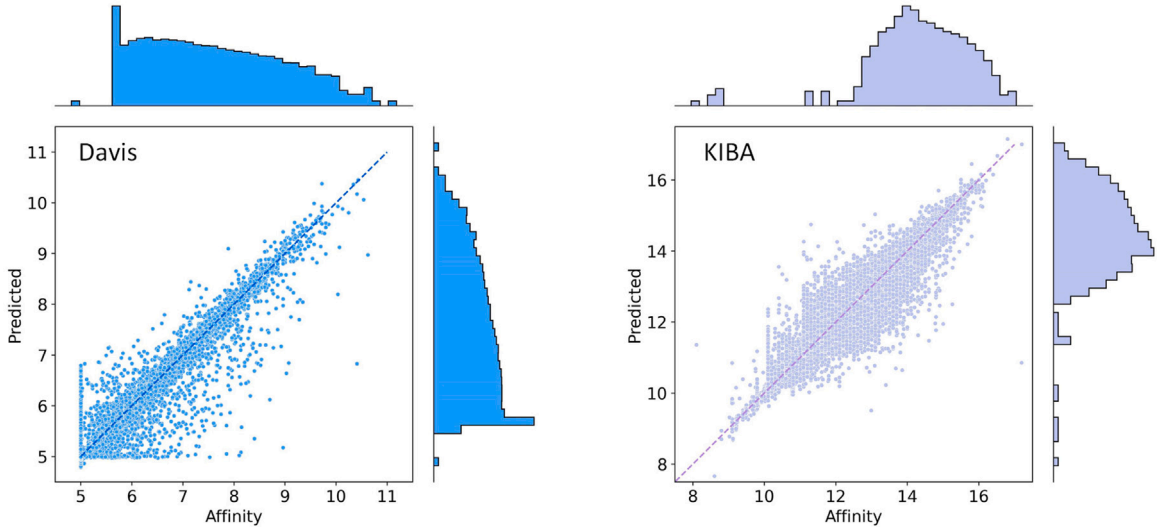


**Fig. 3.** Ground truth affinities (x-axis) vs predicted affinities (y-axis) for drug-target pairs in Davis and KIBA datasets.

**Table 7**
Performance comparison between our model and baseline methods on Davis dataset under cold-start settings.

| Setting | Model | CI | MSE | $r_m^2$ |
|---|---|---|---|---|
| Drug cold-start | DeepDTA | 0.633 (0.030) | 0.675 (0.262) | 0.062 (0.029) |
| | AttentionDTA | 0.649 (0.031) | **0.630 (0.274)** | 0.091 (0.031) |
| | GraphDTA | 0.660 (0.035) | 0.722 (0.271) | 0.108 (0.070) |
| | TransformerCPI | 0.608 (0.059) | 0.788 (0.255) | 0.057 (0.044) |
| | ML-DTI | 0.640 (0.030) | 0.735 (0.195) | 0.113 (0.068) |
| | MGPLI | 0.644 (0.051) | 0.711 (0.189) | 0.088 (0.045) |
| | ELECTRA-DTA | 0.659 (0.055) | 0.667 (0.129) | 0.094 (0.069) |
| | AttentionMGT-DTA | **0.696 (0.034)** | 0.729 (0.213) | **0.162 (0.081)** |
| Target cold-start | DeepDTA | 0.780 (0.008) | 0.424 (0.039) | **0.351 (0.036)** |
| | AttentionDTA | 0.748 (0.008) | 0.490 (0.056) | 0.246 (0.022) |
| | GraphDTA | 0.755 (0.006) | 0.494 (0.061) | 0.277 (0.020) |
| | TransformerCPI | 0.725 (0.009) | 0.503 (0.051) | 0.244 (0.026) |
| | ML-DTI | 0.732 (0.008) | 0.459 (0.032) | 0.281 (0.025) |
| | MGPLI | 0.766 (0.005) | 0.485 (0.048) | 0.308 (0.019) |
| | ELECTRA-DTA | 0.804 (0.006) | 0.435 (0.042) | 0.318 (0.018) |
| | AttentionMGT-DTA | **0.829 (0.005)** | **0.422 (0.031)** | 0.284 (0.017) |
| Drug-target cold-start | DeepDTA | 0.597 (0.034) | 0.679 (0.101) | 0.037 (0.029) |
| | AttentionDTA | 0.560 (0.036) | 0.676 (0.147) | 0.017 (0.015) |
| | GraphDTA | 0.603 (0.032) | 0.779 (0.084) | 0.058 (0.045) |
| | TransformerCPI | 0.564 (0.029) | 0.739 (0.155) | 0.047 (0.042) |
| | ML-DTI | 0.601 (0.033) | 0.728 (0.143) | 0.060 (0.033) |
| | MGPLI | 0.587 (0.027) | 0.761 (0.148) | 0.050 (0.052) |
| | ELECTRA-DTA | 0.605 (0.026) | 0.617 (0.105) | 0.054 (0.045) |
| | AttentionMGT-DTA | **0.613 (0.031)** | **0.612 (0.082)** | **0.065 (0.046)** |

- ML-DTI (Yang, Zhong, et al., 2021) established a mutual learning mechanism to bridge the gap between feature extraction modules from a global perspective, improving the generalization and interpretation ability.
- MGPLI (Wang, Hu, et al., 2022) employed the transformer encoders to capture potential interactions between atoms and residues, and CNN layers were used to fuse and intergrate multi-granular information.

It is noteworthy that 5-fold cross-validation was applied in the experiments on benchmarks to prevent overfitting and ensure the fairness of comparison. The results are presented in Tables 5 and 6.

As shown, AttentionMGT-DTA attained the best performance in the $r_m^2$ metric on the Davis dataset, achieving a 0.2% improvement compared to baseline models. For the large-scale dataset, on the KIBA dataset, AttentionMGT-DTA was superior to other baseline models in terms of all metrics. In detail, our model improved the CI metric by 1.0% and reduced the MSE metric by 1.2% compared to state-of-the-art methods. Furthermore, our model also achieved a 2.0% increase in the $r_m^2$ index over previous methods. For GraphDTA, which used GNN as drug encoder, AttentionMGT-DTA achieved a 1.0% 5.1% and 2.2% improvement on average in terms of CI, MSE and $r_m^2$, respectively. When compared with those transformer-based methods, i.e., TransformerCPI and MGPLI, our method also demonstrated its advantages in DTA prediction. These improvement indicated the superiority of 3D protein graph, which enabled our model to represent target proteins precisely and comprehensively. These results suggested that AttentionMGT-DTA achieved a very competitive DTA prediction performance on the large dataset by attention-based multi-modal information fusion. When extracting features of target proteins, the models and interactions of sequence and graph structure information significantly improve protein representations. When extracting interactive features of drugs and targets, the attention module can preserve integrated information between drug-target pairs. The possible reason for performance differences of our model on the Davis and KIBA datasets is the dependence on the protein 3D structures from the AlphaFold Database. The variable quality of protein structures obtained from AlphaFold2 led to the relatively mediocre performance of our model on Davis dataset.

**Table 8**
The parameter setting analysis of the threshold of protein residue graphs on the Davis dataset.

| Threshold Setting (Å) | Average number of edges | Average degree | $r_m^2$ | CI | MSE |
|---|---|---|---|---|---|
| 5.0 | 6688 | 8.98 | 0.661 (0.023) | 0.872 (0.007) | 0.209 (0.006) |
| 8.0 | 14 475 | 19.56 | 0.692 (0.029) | 0.885 (0.006) | 0.194 (0.009) |
| 10.0 | 21 564 | 29.19 | **0.699 (0.027)** | **0.891 (0.005)** | **0.193 (0.010)** |
| 12.0 | 30 756 | 41.66 | 0.694 (0.025) | 0.887 (0.005) | 0.195 (0.012) |
| 15.0 | 47 137 | 63.80 | 0.670 (0.031) | 0.871 (0.006) | 0.208 (0.011) |

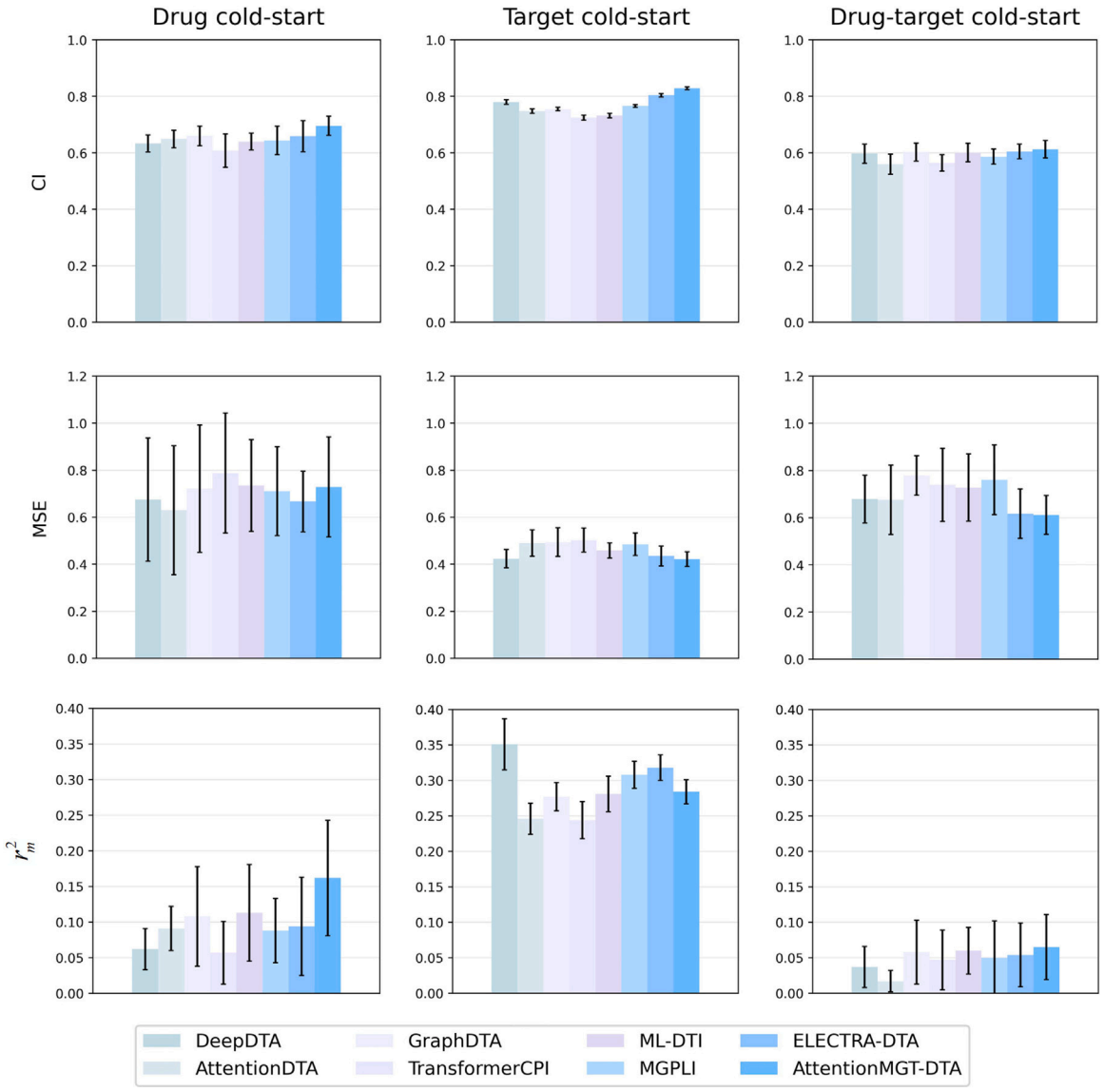


**Fig. 4.** Performance comparison of AttentionMGT-DTA and baseline methods under cold-start settings on Davis dataset.

To further analyze the experimental results, we also plotted the predicted affinity and ground truth for the Davis and KIBA dataset. Fig. 3 illustrates the scatter plot of the predicted affinities against the actual affinities of the two datasets. Setting the $x$-axis to the ground truth and the $y$-axis to the predicted value, an ideal model can generate a straight line $y = x$. As shown in Fig. 3, the samples are on or close to the straight line, which are symmetrically distributed. In addition, the result suggests our model performed better on the KIBA dataset, as the points were highly densely distributed around the straight line $y = x$.

### 4.4.2. The performance of cold-start

The generalization and robustness are critical issues in DTA prediction methods, especially in the case of unseen drugs and unseen proteins. Under the setting of DTA prediction task, the data redundancy caused by similar or identical drugs or proteins may result in simpler prediction task, which may confuse the performance evaluation of methods. From a practical application perspective, most drugs and proteins in the training set would not appear in the test set. Whether the model continues to perform well when encountering unseen data is an important challenge. Therefore, in the experiments, we performed three cold-start schemes following the setting of previous work (Wang, Wen, et al., 2022). Specifically, three different splitting settings were implemented as:

- Drug cold-start: Each drug that appears in the training set does not appear in the test set.

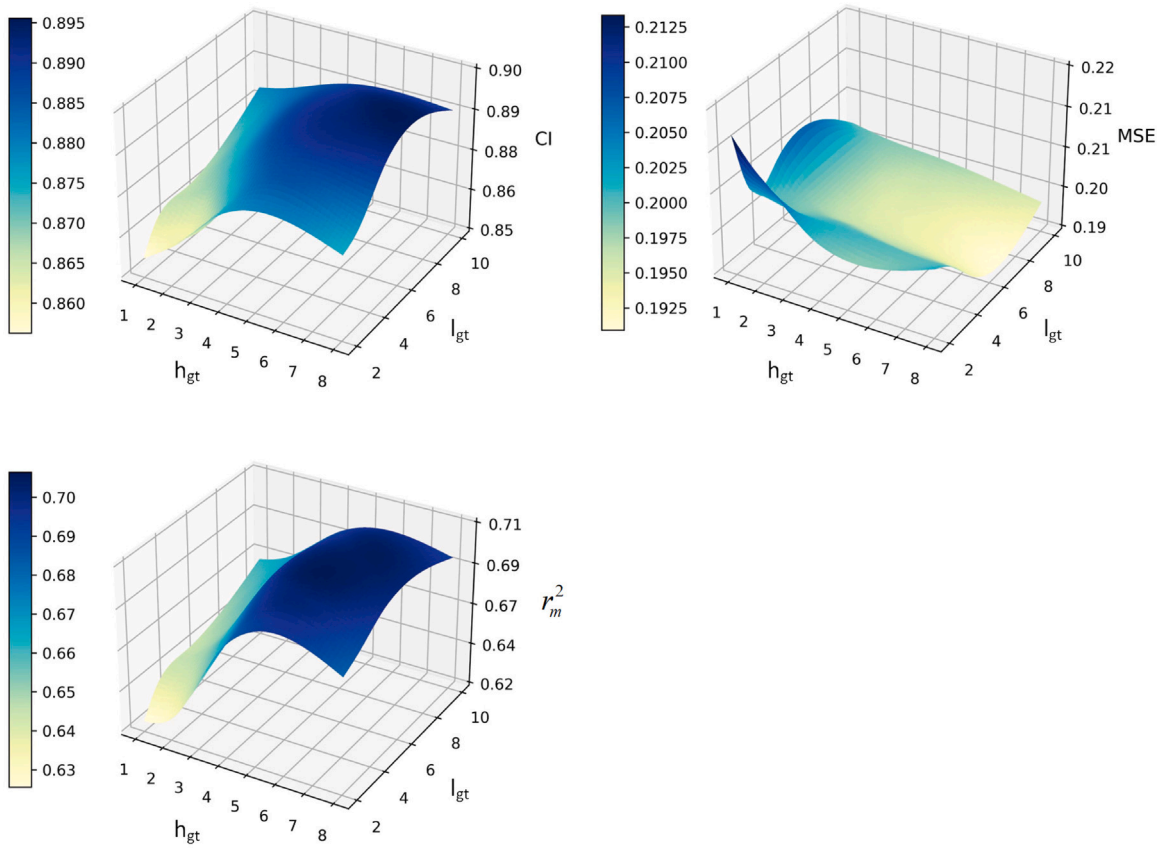


**Fig. 5.** The parameter setting analysis of heads and layers of the graph transformer models on Davis dataset.

- Target cold-start: Each target that appears in the training set does not appear in the test set.
- Drug-target cold-start: Both drug and target which appear in the training set do not appear in the test set.

These cold-start schemes represent a more realistic and complicated environment which is closer to real-world application scenarios. The results of cold-start settings on the Davis dataset are shown in Table 7 and Fig. 4. We compared AttentionMGT-DTA with seven baseline models. As demonstrated, all models showed significant performance degradation, indicating the complexity and distress of this more realistic condition. Our model obtained the best overall performance in the three cold-start settings. In the drug cold-start setting, our model demonstrated advantages in the CI and $r_m^2$ metric, which increased by 3.6% and 4.9% respectively. Under the target cold-start condition, the advantages of our approach were equally obvious, improving 2.5% and 0.4% on CI and MSE values, respectively. Moreover, it can be seen that in the drug-target cold-start schemes, where drugs and targets during training are both absent in the test set, our model showed impressive stability compared with other baselines, which outperformed baselines in all three metrics.

We conjecture the robust performance of our model under the cold-start settings can be explained from two aspects: First is that our graph-based representation method with profuse biological and chemical property feature information for drug molecules and protein pockets, helping our model to learn universal and general topological knowledge near binding sites, which can enhance the generalization capability when encountering unfamiliar data. The second is due to the introduction of pretrained language embeddings, which providing generic protein property information and facilitating our model to learn prominent unseen protein representations by the interaction module. This merit allowed AttentionMGT-DTA to show better advantages in the target cold-start and drug-target cold-start schemes. In general, AttentionMGT-DTA demonstrated a relatively stable and robust performance in cold-start settings, which proved that our multi-modal attention-based network design was effective in undiscovered DTA application scenarios.

### 4.5. Parameter optimization and analysis

We explored the effect of the hyperparameters in AttentionMGT-DTA: (1) The threshold of protein graph construction; (2) The number of heads and layers of the graph transformer module; (3) The dimension of embeddings of drugs and proteins. We implemented the parameter analysis experiment on the Davis dataset.

In our construction of protein graph, the threshold determined the number of neighbor residues to which each residue can be connected, which was an important hyperparameter affecting the complexity and accuracy of protein residue graphs. Empirically speaking, increasing the threshold value can aggregate more structural and topological information of proteins. However, excessive threshold value may lead to noisy connections and over-smooth issues in GNN. In addition, increasing the threshold will increase the size of the graph data, thus enhancing the resource consumption for model training. Hence, we evaluated AttentionMGT-DTA with the threshold value changed from {5, 8, 10, 12, 15}. Table 8 illustrated the detail result of the parameter setting analysis of the threshold of protein residue graphs. Among them, average number of edges denoted the average number of edges per protein graph in Davis dataset and average degree represented average node degrees under corresponding settings. As shown, our model achieved the highest performance in terms of all three metrics when the threshold was set to 10.0. Afterwards, we estimated the impact

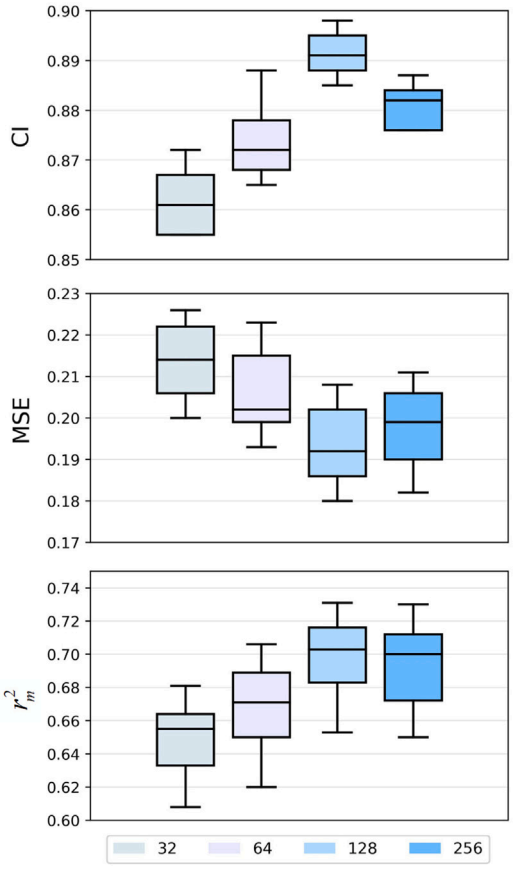


**Fig. 6.** The parameter setting analysis of the embedding sizes on Davis dataset.

of two hyperparameters in the graph transformer module with grid search: the number of attention heads with search range {1, 2, 4, 8} and the number of layers with search range {2, 3, 5, 10}. As illustrated in Fig. 5, the two axes denoted the number of attention heads $h_{gt}$ and the number of graph transformer layers $l_{gt}$ respectively, the metric values for our model reached the maximum at the same time when heads were set to 8 and layers equal to 5. Ultimately, considering that the dimension of feature embeddings affected the capability of the model to learn features, we also performed search experiment to explore the best parameters setting of embedding sizes of drugs and proteins as Fig. 6, from which we can find that there were significant differences in evaluation metrics between different dimension sizes. Overall, based on the above experimental results and analysis we have determined the corresponding optimal parameter settings in AttentionMGT-DTA.

### 4.6. Ablation study

#### 4.6.1. The effect of model architecture modules

To validate the contribution and effectiveness of each module in AttentionMGT-DTA, we conducted ablation studies on the Davis dataset. We performed ablation experiments by removing pretrained protein embeddings, cross-attention and joint-attention modules. The performances of the models with different modules are listed in Table 9. Specifically, the first model AttentionMGT-DTA$_{concat}$ concatenated the protein graph embeddings and pretrained sequence embeddings instead of using the cross-attention mechanism. AttentionMGT-DTA$_{single}$ is the model without pretrained protein embedding features. For the third model AttentionMGT-DTA$_{max}$, the joint-attention mechanism was removed by adding a max pooling layer after graph transformers, and the pretrained embedding was also aggregated into a vector with the same

**Table 9**
Results of ablation study on Davis dataset.

| Model | $r_m^2$ | CI | MSE |
|---|---|---|---|
| AttentionMGT-DTA$_{concat}$ | 0.663 (0.026) | 0.876 (0.006) | 0.206 (0.006) |
| AttentionMGT-DTA$_{single}$ | 0.678 (0.021) | 0.889 (0.005) | 0.198 (0.006) |
| AttentionMGT-DTA$_{max}$ | 0.667 (0.021) | 0.886 (0.007) | 0.198 (0.006) |
| AttentionMGT-DTA$_{mean}$ | 0.673 (0.030) | 0.886 (0.009) | 0.199 (0.007) |
| AttentionMGT-DTA$_{PDB}$ | 0.675 (0.025) | 0.887 (0.006) | 0.196 (0.012) |
| AttentionMGT-DTA | **0.699 (0.027)** | **0.891 (0.005)** | **0.193 (0.010)** |

dimension by the maximizing function. Analogously, AttentionMGT-DTA$_{mean}$ was the model utilizing the mean pooling method to replace the joint-attention module.

As shown in Table 9, AttentionMGT-DTA maintained the best performance compared with the variant models. Removal of the cross-attention module significantly decreased the prediction performance of the model, which clearly indicated the importance of multi-modal interaction. For the ablation study of pretrained embeddings, the results suggested that the 1D sequence embedding introduced by our model provided efficient high-level protein sequence information, which resulted in improved performance. Furthermore, AttentionMGT-DTA also outperformed the third and fourth variant model, indicating that the joint-attention mechanism enhanced the performance of our model. We speculate that this was mainly due to the information interaction capability between drug and target data of our model, which outperformed the traditional pooling and concatenation strategy and was beneficial for DTA prediction. Overall, the ablation experiments indicated that the modality and data interaction modules in our model were effective for improving the prediction performance.

#### 4.6.2. The effect of AlphaFold structures

As described above, the availability of the protein's crystallographic structure is also a major factor affecting final performance. Considering that partial protein structure information is missing and unavailable in the PDB database, we consequently selected protein structures from AlphaFold Database instead of PDB. We further investigated the effect of AlphaFold structures on the overall performance by replacing AlphaFold structures with available PDB structures. In particular, the number of available PDB protein structures in Davis dataset is 275, while the number of missing structures supplemented by AlphaFold is 86. As reported in Table 9, the model performance rose by 0.4% 0.3% and 2.4% in terms of CI, MSE and $r_m^2$ index. This improvement may be due to the presence of more structurally similar proteins in dataset which enable our model to learn complex binding properties and features between drug-target pairs, while the mixed existence of two database structures can affect the learning ability of the model. To this end, we demonstrated that the utilization of high-quality AlphaFold-predicted structures enhanced the performance of the proposed model and facilitated our method outperform existing baselines.

### 4.7. Interpretation and visualization

The joint-attention module in AttentionMGT-DTA can be utilized to analyze which protein binding sites are more likely to bind with the drug molecule. By inputting the drug graph and protein graph representations, our model can generate attention matrices, which demonstrated the importance of each protein residue and drug atom in the drug-target binding. The attention matrix can provide a biological and reasonable explanation for the probabilities, which is one advantage of our model. Fig. 7 shows examples of the attention visualization of the proposed model. To exemplify the interpretability of the model, we chose two complexes from PDB (Burley et al., 2018) for binding visual analysis. We colored the top-weighted positions of protein residues and drug atoms. In particular, the colored residues highlight the positions of the protein binding pockets with high attention scores,

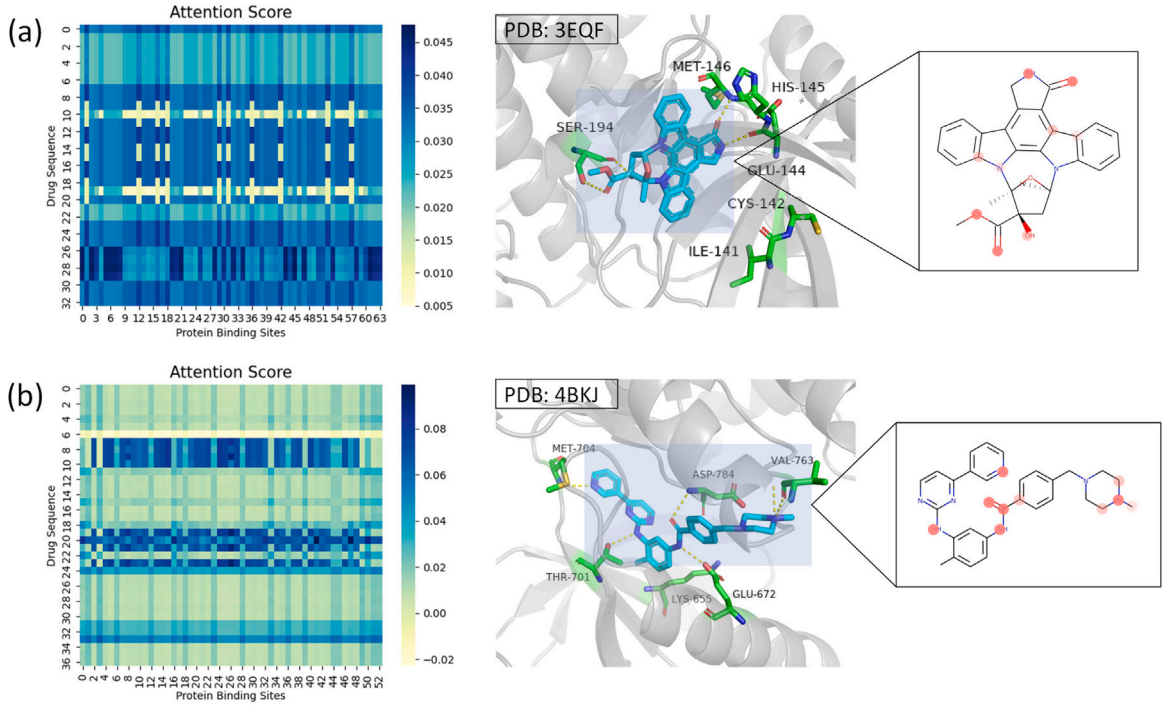


**Fig. 7.** Attention visualization of DTAs. Left: Heatmaps of attention matrices. Middle: Drugs and highlighted residues are represented in blue and green, respectively. Right: Drug structure with highlighted atoms in red. (a) The attention visualization of 3EQF. (b) The attention visualization of 4BKJ.

and the red color highlights the focused drug atoms. The attention scores were obtained by averaging the attention matrix on the protein dimension and drug dimension, respectively.

For protein MAP2K1 (Uniport ID: Q02750) in Fig. 7, the highlighted residues identified by the model include ILE141, CYS142, GLU144, HIS145, MET146 and SER194, which partially overlapped with the binding site residues observed in the complex structure (PDB ID: 3EQF). Analogously, it was observed that the drug atoms with high attention values were within or surrounding the protein pockets. For protein DDR1 (UniProt ID: Q08345) in Fig. 7, the key residues (LYS655, GLU672, THR701, MET704, VAL763, ASP784) and molecule atoms were also similar to the observed binding in the co-crystal complex (PDB ID: 4BKJ). Overall, most residues captured by our model were located in the binding sites, but there were still some incorrect predictions. The results indicated that the attention mechanism can extract meaningful binding information by learning the important protein residues and drug atoms. Moreover, the results suggested that the proposed model can help us understand DTA accurately and comprehensively, which is beneficial for researchers when studying the binding and interaction mechanism between target proteins and drugs.

## 5. Conclusions

In this article, we propose a novel model named AttentionMGT-DTA for DTA prediction, based on the attention mechanism to capture relationships between various independent modalities. AttentionMGT-DTA employs multi-modal protein residue-level features and drug atom-level features through graph and attention-based encoders. Simultaneously, our model can learn the attention matrix using the information fusion module between drug and protein, which can provide significant interpretation of biological meaning. Experimental results on public datasets demonstrated that our model performed better than existing models. When encountering unknown drugs and proteins, AttentionMGT-DTA is also robust and effective. Moreover, the superiority of our model also indicated that the protein structure predicted from AlphaFold Database was of high quality, which can provide accurate and effective structural information for downstream tasks.

Although our model has been proven to have excellent performance, there is still room for further improvement. We only used one type of modality, 2D molecular graphs to represent drugs, and no consideration is given to other representations such as sequence and fingerprints. In future work, we will focus on the integration and fusion of both drug and target in multi-modal deep learning models. Additionally, the algorithm used in our work to determine protein binding pocket is sometimes inaccurate, leading to the mismatched pockets of proteins in the datasets, which can affect performance to some extent. Therefore, introducing and applying more approaches to find cryptic pockets of proteins, which enable the presence of more sites in the protein that can bind to drugs, is another future direction to improve the performance and strengthen the generalizable capability.

## Code availability

The code is freely available at GitHub: https://github.com/JK-Liu7/AttentionMGT-DTA.

## Declaration of competing interest

The authors declare no conflict of interests.

## Data availability

Github link is given in the paper.

## Acknowledgments

This work has been supported by the National Natural Science Foundation of China (62073231, 62176175, 62172076), National Research Project (2020YFC2006602), Provincial Key Laboratory for Computer Information Processing Technology, Soochow University (KJS2166), Opening Topic Fund of Big Data Intelligent Engineering Laboratory of Jiangsu Province (SDGC2157), Postgraduate Research and Practice Innovation Program of Jiangsu Province, Zhejiang Provincial Natural Science Foundation of China (Grant No. LY23F020003), and the Municipal Government of Quzhou, China (Grant No. 2023D038).

## References


Bagherian, M., Sabeti, E., Wang, K., Sartor, M. A., Nikolovska-Coleska, Z., & Najarian, K. (2020). Machine learning approaches and databases for prediction of drug–target interaction: A survey paper. *Briefings in Bioinformatics*, *22*(1), 247–269. http://dx.doi.org/10.1093/bib/bbz157.

Bahi, M., & Batouche, M. (2021). Convolutional neural network with stacked autoencoders for predicting drug-target interaction and binding affinity. *International Journal of Data Mining, Modelling and Management*, *13*(1–2), 81–113. http://dx.doi.org/10.1504/IJDMMM.2021.112914.

Bepler, T., & Berger, B. (2021). Learning the protein language: Evolution, structure, and function. *Cell Systems*, *12*(6), 654–669. http://dx.doi.org/10.1016/j.cels.2021.05.017.

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., et al. (2018). RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research*, *47*(D1), D464–D474. http://dx.doi.org/10.1093/nar/gky1004.

Chao, W., & Quan, Z. (2021). A machine learning method for differentiating and predicting human-infective coronavirus based on physicochemical features and composition of the spike protein. *Chinese Journal of Electronics*, *30*(5), 815–823. http://dx.doi.org/10.1049/cje.2021.06.003.

Chen, L., Tan, X., Wang, D., Zhong, F., Liu, X., Yang, T., et al. (2020). TransformerCPI: Improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, *36*(16), 4406–4414. http://dx.doi.org/10.1093/bioinformatics/btaa524.

Cloninger, A., & Klock, T. (2021). A deep network construction that adapts to intrinsic dimensionality beyond the domain. *Neural Networks*, *141*, 404–419. http://dx.doi.org/10.1016/j.neunet.2021.06.004.

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., et al. (2011). Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, *29*(11), 1046–1051. http://dx.doi.org/10.1038/nbt.1990.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Dhakal, A., McKay, C., Tanner, J. J., & Cheng, J. (2021). Artificial intelligence in the prediction of protein–ligand interactions: Recent advances and future directions. *Briefings in Bioinformatics*, *23*(1), bbab476. http://dx.doi.org/10.1093/bib/bbab476.

Ding, Y., Guo, F., Tiwari, P., & Zou, Q. (2023). Identification of drug-side effect association via multi-view semi-supervised sparse model. *IEEE Transactions on Artificial Intelligence*, 1–12. http://dx.doi.org/10.1109/TAI.2023.3314405.

Ding, Y., Tang, J., & Guo, F. (2020a). Identification of drug–target interactions via dual Laplacian regularized least squares with multiple kernel fusion. *Knowledge-Based Systems*, *204*, Article 106254. http://dx.doi.org/10.1016/j.knosys.2020.106254.

Ding, Y., Tang, J., & Guo, F. (2020b). Identification of drug-target interactions via fuzzy bipartite local model. *Neural Computing and Applications*, *32*, 10303–10319. http://dx.doi.org/10.1007/s00521-019-04569-z.

Ding, Y., Tang, J., & Guo, F. (2021). Identification of drug-target interactions via multi-view graph regularized link propagation model. *Neurocomputing*, *461*, 618–631. http://dx.doi.org/10.1016/j.neucom.2021.05.100.

Ding, Y., Tang, J., Guo, F., & Zou, Q. (2022). Identification of drug–target interactions via multiple kernel-based triple collaborative matrix factorization. *Briefings in Bioinformatics*, *23*(2), http://dx.doi.org/10.1093/bib/bbab582, bbab582.

Ding, Y., Tiwari, P., Guo, F., & Zou, Q. (2022). Shared subspace-based radial basis function neural network for identifying ncRNAs subcellular localization. *Neural Networks*, *156*, 170–178. http://dx.doi.org/10.1016/j.neunet.2022.09.026.

Dwivedi, V. P., & Bresson, X. (2020). A generalization of transformer networks to graphs. arXiv:2012.09699.

Dwivedi, V. P., Joshi, C. K., Luu, A. T., Laurent, T., Bengio, Y., & Bresson, X. (2020). Benchmarking graph neural networks. http://dx.doi.org/10.48550/arXiv.2003.00982, arXiv preprint arXiv:2003.00982.

ElAbd, H., Bromberg, Y., Hoarfrost, A., Lenz, T., Franke, A., & Wendorff, M. (2020). Amino acid encoding for deep learning applications. *BMC Bioinformatics*, *21*, 1–14. http://dx.doi.org/10.1186/s12859-020-03546-x.

Ezzat, A., Wu, M., Li, X.-L., & Kwoh, C.-K. (2018). Computational prediction of drug–target interactions using chemogenomic approaches: An empirical survey. *Briefings in Bioinformatics*, *20*(4), 1337–1357. http://dx.doi.org/10.1093/bib/bby002.

Feng, Q., Dueva, E., Cherkasov, A., & Ester, M. (2019). PADME: A deep learning-based framework for drug-target interaction prediction. arXiv:1807.09741.

Gönen, M., & Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, *92*(4), 965–970.

Huang, L., Lin, J., Liu, R., Zheng, Z., Meng, L., Chen, X., et al. (2022). CoaDTI: Multi-modal co-attention based framework for drug–target interaction annotation. *Briefings in Bioinformatics*, *23*(6), bbac446. http://dx.doi.org/10.1093/bib/bbac446.

Huang, K., Xiao, C., Glass, L. M., & Sun, J. (2020). MolTrans: Molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, *37*(6), 830–836. http://dx.doi.org/10.1093/bioinformatics/btaa880.

Jiang, D., Hsieh, C.-Y., Wu, Z., Kang, Y., Wang, J., Wang, E., et al. (2021). InteractionGraphNet: A novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions. *Journal of Medicinal Chemistry*, *64*(24), 18209–18232. http://dx.doi.org/10.1021/acs.jmedchem.1c01830, PMID: 34878785.

Jiang, M., Li, Z., Zhang, S., Wang, S., Wang, X., Yuan, Q., et al. (2020). Drug–target affinity prediction using graph neural network and contact maps. *RSC Advances*, *10*, 20701–20712. http://dx.doi.org/10.1039/D0RA02297G.

Jumper, J., Evans, R., Pritzel, A., Green, T., & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. http://dx.doi.org/10.1038/s41586-021-03819-2.

Karimi, M., Wu, D., Wang, Z., & Shen, Y. (2019). DeepAffinity: Interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, *35*(18), 3329–3338. http://dx.doi.org/10.1093/bioinformatics/btz111.

Karimi, M., Wu, D., Wang, Z., & Shen, Y. (2021). Explainable deep relational networks for predicting compound–protein affinities and contacts. *Journal of Chemical Information and Modeling*, *61*(1), 46–66. http://dx.doi.org/10.1021/acs.jcim.0c00866, PMID: 33347301.

Kimber, T. B., Chen, Y., & Volkamer, A. (2021). Deep learning in virtual screening: Recent applications and developments. *International Journal of Molecular Sciences*, *22*(9), http://dx.doi.org/10.3390/ijms22094435.

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. arXiv:1412.6980.

Lee, I., Keum, J., & Nam, H. (2019). DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Computational Biology*, *15*(6), 1–21. http://dx.doi.org/10.1371/journal.pcbi.1007129.

Li, S., Wan, F., Shu, H., Jiang, T., Zhao, D., & Zeng, J. (2020). MONN: A multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Systems*, *10*(4), 308–322.e11. http://dx.doi.org/10.1016/j.cels.2020.03.002.

Li, F., Zhang, Z., Guan, J., & Zhou, S. (2022a). Effective drug–target interaction prediction with mutual interaction neural network. *Bioinformatics*, *38*(14), 3582–3589. http://dx.doi.org/10.1093/bioinformatics/btac377.

Li, F., Zhang, Z., Guan, J., & Zhou, S. (2022b). Effective drug–target interaction prediction with mutual interaction neural network. *Bioinformatics*, *38*(14), 3582–3589. http://dx.doi.org/10.1093/bioinformatics/btac377.

Li, T., Zhao, X.-M., & Li, L. (2022). Co-VAE: Drug-target binding affinity prediction by co-regularized variational autoencoders. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 8861–8873. http://dx.doi.org/10.1109/TPAMI.2021.3120428.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123–1130. http://dx.doi.org/10.1126/science.ade2574.

Mao, T., Shi, Z., & Zhou, D. (2021). Theory of deep convolutional neural networks III: Approximating radial functions. *Neural Networks*, *144*, 778–790. http://dx.doi.org/10.1016/j.neunet.2021.09.027.

Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., & Beckstein, O. (2011). MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, *32*(10), 2319–2327. http://dx.doi.org/10.1002/jcc.21787.

Newman, D. J., & Cragg, G. M. (2020). Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of Natural Products*, *83*(3), 770–803. http://dx.doi.org/10.1021/acs.jnatprod.9b01285, PMID: 32162523.

Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., & Venkatesh, S. (2020). GraphDTA: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, *37*(8), 1140–1147. http://dx.doi.org/10.1093/bioinformatics/btaa921.

Nguyen, T. M., Nguyen, T., Le, T. M., & Tran, T. (2022a). GEFA: Early fusion approach in drug-target affinity prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(2), 718–728. http://dx.doi.org/10.1109/TCBB.2021.3094217.

Nguyen, T. M., Nguyen, T., Le, T. M., & Tran, T. (2022b). GEFA: Early fusion approach in drug-target affinity prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(2), 718–728. http://dx.doi.org/10.1109/TCBB.2021.3094217.

Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: Deep drug–target binding affinity prediction. *Bioinformatics*, *34*(17), i821–i829. http://dx.doi.org/10.1093/bioinformatics/bty593.

Öztürk, H., Ozkirimli, E., & Özgür, A. (2019). WideDTA: Prediction of drug-target binding affinity. arXiv:1902.04166.

Pandey, M., Radaeva, M., Mslati, H., Garland, O., Fernandez, M., Ester, M., et al. (2022). Ligand binding prediction using protein structure graphs and residual graph attention networks. *Molecules*, *27*(16), http://dx.doi.org/10.3390/molecules27165114, URL https://www.mdpi.com/1420-3049/27/16/5114.

Qian, Y., Ding, Y., Zou, Q., & Guo, F. (2022). Identification of drug-side effect association via restricted Boltzmann machines with penalized term. *Briefings in Bioinformatics*, *23*(6), http://dx.doi.org/10.1093/bib/bbac458, bbac458.

Rifaioglu, A. S., Nalbat, E., Atalay, V., Martin, M. J., Cetin-Atalay, R., & Doğan, T. (2020). DEEPScreen: High performance drug–target interaction prediction with convolutional neural networks using 2-D structural compound representations. *Chemical Science*, *11*, 2531–2557. http://dx.doi.org/10.1039/C9SC03414E.

Saberi Fathi, S. M., & Tuszynski, J. A. (2014). A simple method for finding a protein's ligand-binding pockets. *BMC Structural Biology*, *14*(1), 1–9. http://dx.doi.org/10.1186/1472-6807-14-18.

Shen, C., Zhang, X., Deng, Y., Gao, J., Wang, D., Xu, L., et al. (2022). Boosting protein–ligand binding pose prediction and virtual screening based on residue–atom distance likelihood potential and graph transformer. *Journal of Medicinal Chemistry*, *65*(15), 10691–10706. http://dx.doi.org/10.1021/acs.jmedchem.2c00991, PMID: 35917397.

Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., et al. (2014). Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis. *Journal of Chemical Information and Modeling*, *54*(3), 735–743. http://dx.doi.org/10.1021/ci400709d, PMID: 24521231.

Torng, W., & Altman, R. B. (2019). Graph convolutional neural networks for predicting drug-target interactions. *Journal of Chemical Information and Modeling*, *59*(10), 4131–4149. http://dx.doi.org/10.1021/acs.jcim.9b00628, PMID: 31580672.

Tsubaki, M., Tomii, K., & Sese, J. (2018). Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, *35*(2), 309–318. http://dx.doi.org/10.1093/bioinformatics/bty535.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., et al. (2021). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, *50*(D1), D439–D444. http://dx.doi.org/10.1093/nar/gkab1061.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. arXiv:1706.03762.

Wang, J., Hu, J., Sun, H., Xu, M., Yu, Y., Liu, Y., et al. (2022). MGPLI: Exploring multi-granular representations for protein–ligand interaction prediction. *Bioinformatics*, *38*(21), 4859–4867. http://dx.doi.org/10.1093/bioinformatics/btac597.

Wang, L., Liu, H., Liu, Y., Kurtin, J., & Ji, S. (2023). Learning hierarchical protein representations via complete 3D graph networks. In *The eleventh international conference on learning representations*. URL https://openreview.net/forum?id=9X-hgLDLYkQ.

Wang, H., Tang, J., Ding, Y., & Guo, F. (2021). Exploring associations of non-coding RNAs in human diseases via three-matrix factorization with hypergraph-regular terms on center kernel alignment. *Briefings in Bioinformatics*, *22*(5), http://dx.doi.org/10.1093/bib/bbaa409, bbaa409.

Wang, J., Wen, N., Wang, C., Zhao, L., & Cheng, L. (2022). ELECTRA-DTA: A new compound-protein binding affinity prediction model based on the contextualized sequence encoding. *Journal of cheminformatics*, *14*(1), 1–14. http://dx.doi.org/10.1186/s13321-022-00591-x.

Wang, Y., Zhai, Y., Ding, Y., & Zou, Q. (2023). SBSM-Pro: Support bio-sequence machine for proteins. http://dx.doi.org/10.48550/arXiv.2308.10275, arXiv preprint arXiv:2308.10275.

Wang, Z., Zhang, Q., HU, S.-W., Yu, H., Jin, X., Gong, Z., et al. (2023). Multi-level protein structure pre-training via prompt learning. In *The eleventh international conference on learning representations*. URL https://openreview.net/forum?id=XGagtiJ8XC.

Wang, P., Zheng, S., Jiang, Y., Li, C., Liu, J., Wen, C., et al. (2022). Structure-aware multimodal deep learning for drug–protein interaction prediction. *Journal of Chemical Information and Modeling*, *62*(5), 1308–1317. http://dx.doi.org/10.1021/acs.jcim.2c00060, PMID: 35200015.

Wang, K., Zhou, R., Li, Y., & Li, M. (2021). DeepDTAF: A deep learning method to predict protein–ligand binding affinity. *Briefings in Bioinformatics*, *22*(5), bbab072. http://dx.doi.org/10.1093/bib/bbab072.

Wouters, O. J., McKee, M., & Luyten, J. (2020). Research and development costs of new drugs—Reply. *JAMA*, *324*(5), 518. http://dx.doi.org/10.1001/jama.2020.8651.

Wu, Y., Gao, M., Zeng, M., Zhang, J., & Li, M. (2022). BridgeDPI: A novel Graph Neural Network for predicting drug–protein interactions. *Bioinformatics*, *38*(9), 2571–2578. http://dx.doi.org/10.1093/bioinformatics/btac155.

Wu, Z., Jiang, D., Wang, J., Hsieh, C.-Y., Cao, D., & Hou, T. (2021). Mining toxicity information from large amounts of toxicity data. *Journal of Medicinal Chemistry*, *64*(10), 6924–6936. http://dx.doi.org/10.1021/acs.jmedchem.1c00421, PMID: 33961429.

Wu, H., Ling, H., Gao, L., Fu, Q., Lu, W., Ding, Y., et al. (2021). Empirical potential energy function toward ab initio folding g protein-coupled receptors. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(5), 1752–1762. http://dx.doi.org/10.1109/TCBB.2020.3008014.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? arXiv:1810.00826.

Yang, H., Ding, Y., Tang, J., & Guo, F. (2021). Drug–disease associations prediction via multiple kernel-based dual graph regularized least squares. *Applied Soft Computing*, *112*, Article 107811. http://dx.doi.org/10.1016/j.asoc.2021.107811.

Yang, Z., Zhong, W., Zhao, L., & Chen, C. Y.-C. (2021). ML-DTI: Mutual learning mechanism for interpretable drug–target interaction prediction. *The Journal of Physical Chemistry Letters*, *12*(17), 4247–4261. http://dx.doi.org/10.1021/acs.jpclett.1c00867, PMID: 33904745.

Yang, Z., Zhong, W., Zhao, L., & Yu-Chian Chen, C. (2022). MGraphDTA: Deep multi-scale graph neural network for explainable drug–target binding affinity prediction. *Chemical Science*, *13*, 816–833. http://dx.doi.org/10.1039/D1SC05180F.

Yazdani-Jahromi, M., Yousefi, N., Tayebi, A., Kolanthai, E., Neal, C. J., Seal, S., et al. (2022). AttentionSiteDTI: An interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification. *Briefings in Bioinformatics*, *23*(4), bbac272. http://dx.doi.org/10.1093/bib/bbac272.

Zhang, Z., Chen, L., Zhong, F., Wang, D., Jiang, J., Zhang, S., et al. (2022). Graph neural network approaches for drug-target interactions. *Current Opinion in Structural Biolog*, *73*, Article 102327. http://dx.doi.org/10.1016/j.sbi.2021.102327.

Zhang, F., Song, H., Zeng, M., Wu, F.-X., Li, Y., Pan, Y., et al. (2021). A deep learning framework for gene ontology annotations with sequence- and network-based information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(6), 2208–2217. http://dx.doi.org/10.1109/TCBB.2020.2968882.

Zhang, Y., Tiwari, P., Song, D., Mao, X., Wang, P., Li, X., et al. (2021). Learning interaction dynamics with an interactive LSTM for conversational sentiment analysis. *Neural Networks*, *133*, 40–56.

Zhang, Z., Xu, M., Jamasb, A., Chenthamarakshan, V., Lozano, A., Das, P., et al. (2023). Protein representation learning by geometric structure pretraining. arXiv:2203.06125.

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* (pp. 64–69). http://dx.doi.org/10.1109/BIBM47256.2019.8983125.

Zhao, Q., Zhao, H., Zheng, K., & Wang, J. (2021). HyperAttentionDTI: Improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics*, *38*(3), 655–662. http://dx.doi.org/10.1093/bioinformatics/btab715.

Zheng, S., Li, Y., Chen, S., Xu, J., & Yang, Y. (2020). Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence*, *2*(2), 134–140. http://dx.doi.org/10.1038/s42256-020-0152-y.