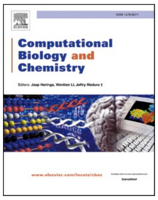
Contents lists available at ScienceDirect

Computational Biology and Chemistry

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



GTAMP-DTA: Graph transformer combined with attention mechanism for drug-target binding affinity prediction

Chuangchuang Tian [a], Luping Wang [a], Zhiming Cui [a], Hongjie Wu [a,b,*]

[a] *College of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China*
[b] *Suzhou Smart City Research Institute, Suzhou University of Science and Technology, Suzhou 215009, China*

ARTICLE INFO

Keywords:
Graph converter
Attention mechanism
Pre-trained
Deep learning
Drug-target affinity

ABSTRACT

Drug target affinity prediction (DTA) is critical to the success of drug development. While numerous machine learning methods have been developed for this task, there remains a necessity to further enhance the accuracy and reliability of predictions. Considerable bias in drug target binding prediction may result due to missing structural information or missing information. In addition, current methods focus only on simulating individual non-covalent interactions between drugs and proteins, thereby neglecting the intricate interplay among different drugs and their interactions with proteins. GTAMP-DTA combines special Attention mechanisms, assigning each atom or amino acid an attention vector. Interactions between drug forms and protein forms were considered to capture information about their interactions. And fusion transformer was used to learn protein characterization from raw amino acid sequences, which were then merged with molecular map features extracted from SMILES. A self-supervised pre-trained embedding that uses pre-trained transformers to encode drug and protein attributes is introduced in order to address the lack of labeled data. Experimental results demonstrate that our model outperforms state-of-the-art methods on both the Davis and KIBA datasets. Additionally, the model's performance undergoes evaluation using three distinct pooling layers (max-pooling, mean-pooling, sum-pooling) along with variations of the attention mechanism. GTAMP-DTA shows significant performance improvements compared to other methods.

1. Introduction

The process of drug discovery is a protracted, exorbitant and risky endeavor. Despite taking over a decade and incurring billions of dollars in costs, just 10% of drugs that reach clinical trials ultimately gain FDA approval and hit the market (Newman et al., 2020; Takebe et al., 2018). However, advancements in computer technology over recent decades have enhanced the efficacy of drug design and experimentation, and have accelerated the pace of drug development (Lin et al., 2020). In modern times, computer-aided drug design primarily focuses on the alignment of drug molecules with proteins, making the investigation of drug-target interactions a prevalent and crucial research subject (Wen et al., 2017).

In the past, vast compound databases have been mined for appropriate medicinal compounds via virtual screening. However, it takes an enormous amount of effort to experimentally validate the binding affinity between the medication and the target utilizing molecular docking technology (Kairys et al., 2019). Drug compounds can directly dock to determine binding affinity when proteins have well-established structure information. However, numerous proteins lack such structural details, and even with extensive time dedicated to homology modeling, obtaining comprehensive structural information remains uncertain (Yadav et al., 2020). To tackle this challenge, several artificial intelligence-driven approaches have been suggested to mitigate this problem. In recent years, traditional machine learning (ML) algorithms have gained popularity for constructing predictive models of compound-protein interactions as binary classification problems (Bahi et al., 2021). For predicting drug-target affinities (DTA), machine learning techniques have gradually replaced molecular docking.

In recent years, powerful deep learning methods have been applied to predict ligand binding sites, which have been widely used in molecular characterization and performance prediction (Zhao et al., 2021a; Zhen et al., 2022; Dhakal et al., 2022), and drug target affinity (DTA) prediction has been rapidly developed (Zhao et al., 2021b). A CNN-based model that examines the SMILES string and sequences of amino acids to predict DTA was introduced by Öztürk et al (Öztürk et al.,

[*] Corresponding author at: College of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China.
E-mail address: hongjie.wu@qq.com (H. Wu).

https://doi.org/10.1016/j.compbiolchem.2023.107982
Received 8 August 2023; Received in revised form 21 October 2023; Accepted 7 November 2023
Available online 15 November 2023
1476-9271/© 2023 Published by Elsevier Ltd.

2019). to record information about the topological structure of pharmaceuticals. Nguyen et al (Nguyen et al., 2020). introduced GraphDTA. This approach represents drugs as graphs and employs various graph neural networks (GNNs) to extract structural information, while protein features are learned using CNNs. Furthermore, Ding et al (Ding et al., 2020a). introduced a double Laplacian regularized least squares model (DLapRLS) and multi-core learning (HSIC-MKL) based on the Hilbert-Schmidt independence criterion to predict DTIs. Yang et al (Ding et al., 2019; Wang et al., 2021; Yang et al., 2021). applied multicore learning or multicore matrix decomposition to enhance prediction performance by integrating multiple information and adjusting the weight of the kernel matrix. A multicore learning method and graph-based semi-supervised learning techniques were also used by Ding et al (Ding et al., 2020b). to combine numerous sources of data and enhance prediction performance. Bi-ConvLSTM layers were employed in the heterogeneity graph attention model developed by Abdel-Basset et al (Abdel-Basset et al., 2020). to extract spatial-sequential data from SMILES string and learn topological information about medicinal molecules. In contrast, Ding et al (Ding et al., 2021). proposed a multi-view graph regularized link propagation model (MvGRLP) to tackle the challenge of integrating multiple sources of information for predicting new DTIs. The proposed model multi-perspective learning approach utilizes complementary and relevant information from different perspectives (features). Many studies have been performed using a multitude of CNNs and SE-Blocks (Hu et al., 2020) for protein sequencing. These methods, however, frequently ignore the crucial fact that only specific regions of the protein or a small number of atoms in the drug participate in interaction between molecules, rather than the entire structure, in favor of finding more potent modules for extracting drug or protein features.

Attention mechanisms were integrated in both DTI and DTA predictions in order to accurately depict the intermolecular interactions between amino acids and atoms (Bahdanau et al., 2015). Next-generation deep learning architectures based on the attention that has been successful in protein structure prediction and interpretation are yet to be developed (Dhakal et al., 2022). An attention-based DTI prediction model was developed by Tsubaki et al (Tsubaki et al., 2019). that encodes a drug as a fixed-length vector and uses a unilateral attention mechanism to assess the importance of subsequences in the protein to the molecule. In their studies, Chen et al (Chen et al., 2021). and Wang et al (Wang et al., 2020). have similarly used comparable attentional mechanisms. Additionally, Gao et al.'s (Gao et al., 2018) use of a bidirectional attention mechanism for DTI prediction enabled reciprocal attention between medicines and proteins. This process of bilateral attention has potential uses in DTA (Abbasi et al., 2020) and not only helps to identify binding sites on proteins but also makes it possible to explore key drug atoms. Zhu et al (Zhu et al., 2023). proposes a DTA prediction method (mutual transformer-drug target affinity [MT-DTA]) with interactive learning and an autoencoder mechanism, Increased informative interaction pathways between molecular sequence pairs complement the expression of correlations between molecular substructures. A novel multiple Laplacian regularization supporting vector machine (MLapSVM-LBS) with local behavioral similarity was also proposed by Sun et al (Sun et al., 2022). to predict DNA-binding proteins. Drawing from human behavior learning theory, MLapSVM-LBS can more effectively depict sample relationships through local behavioral similarity. Furthermore, Chen et al (Chen et al., 2020). separated medicines and proteins as independent sequences, creating a Transformer-based model called as TransformerCPI for forecasting drug-target interactions (DTIs), which was motivated by the exceptional feature-capturing capacity of Transformer (Vaswani et al., 2017) in dealing with two sequences.

Inspired by earlier attention-based models (Tsubaki et al., 2019; Gao et al., 2018; Chen et al., 2020; Huang et al., 2021), our proposed method, GTAMP-DTA, utilizes the unique attention mechanism of graphical transformations to predict DTA. Graph Transformer introduces a powerful attention mechanism that outperforms both GCN and GAT in terms of adaptability. whereas GCN employs a fixed convolutional operation and GAT uses a fixed number of attention headers, Graph Transformer allows for the dynamic allocation of attention weights, enabling it to recognize and prioritize more relevant node connections. Specifically, GTAMP-DTA takes as inputs to the model graphical representations of the SMILES sequences of drugs and amino acid sequences of proteins. The Graph Transformer Module is utilized to encode the graph structures of drugs and proteins, obtaining their corresponding node features. Our method incorporates attention vectors for each amino acid-atom pair to capture the interactions and alter the feature representation across channels, in contrast to earlier attention-based models. Notably, the output feature vectors of drug-protein interactions are fused with pre-trained model feature vectors such as Mol2vec Embedding (Jaeger et al., 2018) and ESM-2 Embedding (Johnson et al., 2010) to extract complementary information about the interactions between proteins and drugs as well as between different drugs. Finally, after obtaining the feature vectors, they undergo processing through a fully connected neural network for DTA prediction. Our method consistently outperforms the benchmark model in comparison to other cutting-edge models in all evaluations. Furthermore, we conducted comparisons between different pooling layers (max-pooling, mean-pooling, and sum-pooling), as well as between models with and without attention. The results indicate that the attentional block effectively reduces the search space of binding points, highlighting its significance in enhancing the model's prediction accuracy.

The contributions of this study are summarized as follows:

- In order to enrich the spatial structure information of drugs and proteins(the molecular structure and bonding pattern of *a* drug, the secondary structure of *a* protein, and missing information refers to unavailable or incomplete information about the structure of *a* drug or *a* protein, as manifested by the lack of *a* protein structure at *a* speciffc binding site.), this study combined the drug and protein mapping conversion modules to obtain a richer feature vector. This approach effectively addresses the issue of information loss and mitigates the problem of irrational protein structure.
- To model the complex non-covalent intermolecular interactions between drug atoms and amino acids, we employed an attention mechanism on the feature vectors, associating a vector with each drug atom and amino acid.
- This study introduces self-supervised pre-trained embeddings to enhance protein/drug association signals. and integrates this advanced information into a unified framework to generate a richer drug-protein feature vector for predicting DTA.

2. Methods and materials

Fig. 1 shows the full architecture of our presented GTAMP-DTA network. The framework consists of four integral modules: the graph construction module, the drug and protein encoding module, the special attention module, and the output module. The graph construction module is responsible for generating molecular graphs with node and edge features for drugs and proteins, respectively. Subsequently, the encoder extracts feature embeddings from these molecular graphs. The obtained feature vectors are then directed to the special attention module, which assigns attention vectors to individual drug atoms and amino acids. Afterward, the feature vectors undergo processing through a mean-pooling layer and are merged with the feature vectors of a pre-trained model to obtain decision vectors. Ultimately, DTA prediction is carried out based on these decision vectors.

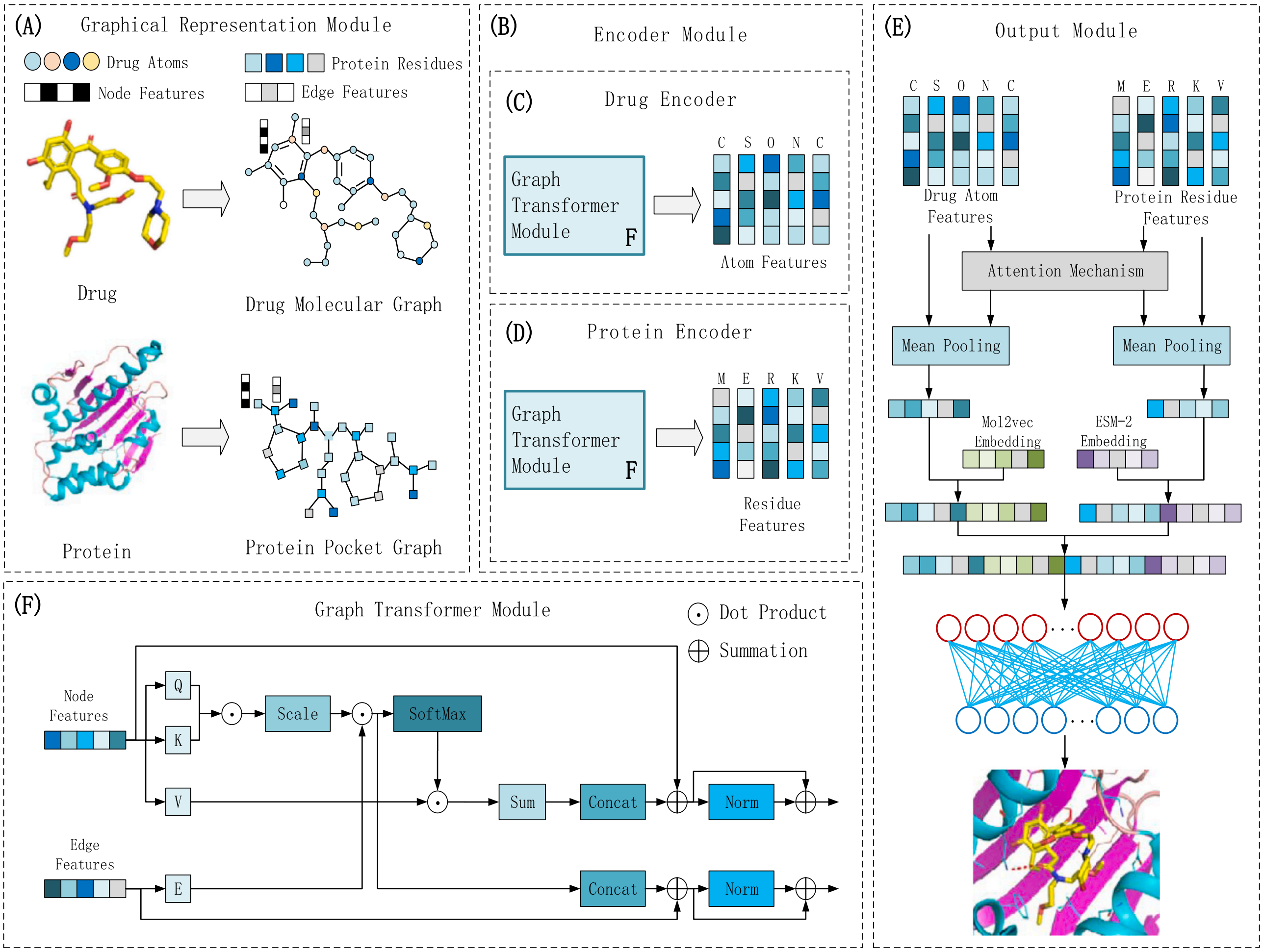


Fig. 1. The overall architecture of GTAMP-DTA. It consists of the following main modules: (A) graph representation of the module, which incorporates the SMILES of the drug and the graph representation of the protein sequence, including the identification of the protein's binding pocket. (B) includes the drug and protein encoder modules, denoted as (C) and (D) respectively, which employ graph converters to extract pertinent features from the drugs and proteins. The output module, denoted as (E), employs global mean-pooling operations alongside an attention mechanism to predict the binding affinity of the drug to the target. This prediction is made by leveraging both the extracted features and the pre-trained model's characteristics. (F) Network structure of the graph converters in (C) and (D).

2.1. Graphical representation module

2.1.1. Schematic representation of drugs

This study uses graphs indicated as $G_c = (V_c\ , A_c)$ to represent drug molecules in order to obtain more detailed molecular structure and chemical information. $V_c \in \mathbb{R}^{n\times f}$ is a collection of n atomic nodes, with each node being represented by an f-dimensional feature vector. An adjacency matrix identifies the set of edges in the molecular graph known as $A_c \in \mathbb{R}^{n\times n}$. The presence of an edge in the adjacency matrix is determined by the covalent bond between the corresponding atoms. For drug characterization, nodes usually correspond to atoms, while edges indicate chemical bonds. Enhanced node characterization may include atom type and atom connectivity. These features provide valuable information for understanding the chemical and structural characteristics of each atom, allowing the model to differentiate between different drug molecules based on their composition.

2.1.2. Protein diagram representation

The spatial relationships between each residue in the graph were described using structural properties of proteins to help us better comprehend the 3-dimensional spatial structure of proteins. The model used protein residues as nodes, and the edges represented the interactions between two residues. To forecast the node properties of the protein graph, Hidden Markov Model (HMM), position-specific scoring matrix (PSSM), and SPIDER3.29 were used. The evolutionary data represented by PSSM and HMM is derived from protein-related gene sequences in protein sequences. After three iterations, PSSM curves were obtained from PSI-BLASTv2.7.1 and the UniRef90 sequence database. HMM curves were generated in HHBLITSv3.0.3 with the default parameter alignment of the HMM database for UniClust30 curves. The structural features consist of 14 attributes that reflect the protein's secondary structure as predicted by SPIDER3. For protein contact maps, the SPOT-Contact method was used to generate protein contact maps. This method predicts the contact probability of all residue pairs in a protein employing information on the protein sequence and evolutionary coupling as input. Finally, we obtain a protein contact map $G_p = (V_p, E_p)$, where the nodes represent protein residues, and the edges represent the interactions between two residues such as non-covalent bonds and spatial proximity. Nodal features of proteins may include amino acid types, physicochemical properties, and structural information such as solvent accessibility and secondary structure. Edge features may capture information on pairwise amino acid interactions, such as hydrogen bonding, hydrophobic interactions, and electrostatic forces. To reduce model complexity and computational resource consumption, constructing protein pocket maps at the residue level instead of the atomic level is done, as non-covalent interactions are more significant than covalent bonds in the 3D structure of proteins (Lin et al., 2022).

2.2. Encoder module

In our model, we utilize the graph transformer framework proposed by Dwivedi et al (Dwivedi et al., 2020). as the drug encoder to extract node representations from the drug graph, as depicted in Fig. 1F. For the drug graph G_D, which consists of node features $\alpha_i \in R^{d_n \times 1}$ for each node i and edge features $\beta_{ij} \in R^{d_e \times 1}$ for edges connecting node i and node j, the initial node and edge features are processed through a linear projection layer to obtain hidden representations h_i^0 and e_{ij}^0, respectively, as follows:

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

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

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

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

$$\check{h}_i^{\varrho} = h_i^0 + \lambda_i^0 \tag{4}$$

where $W_C^0 \in R^{d \times k}$, $b_C^0 \in R^d$ are weights and biases, respectively. λ_i represents the initial feature of each node. Here, we employ the Laplacian positional encodings according to the initial graph transformer architecture.

The multi-head attention mechanism serves as the basic foundation for the graph transformer's update node and edge functionalities. The following equations define the detailed update process of the lth layer:

$$Q_{ij}^{k,l} = Q^{k,l} Norm\left(h_i^l\right) \tag{5}$$

$$K_{ij}^{k,l} = K^{k,l} Norm\left(h_j^l\right) \tag{6}$$

$$V_{ij}^{k,l} = V^{k,l} Norm\left(h_i^l\right) \tag{7}$$

$$E_{ij}^{k,l} = E^{k,l} Norm\left(e_{ij}^l\right) \tag{8}$$

$$w_{ij}^{k,l} = softmax_j\left(\left(\frac{Q_{ij}^{k,l} h_i^l \bullet K_{ij}^{k,l} h_j^l}{\sqrt{d_k}}\right) \bullet E_{ij}^{k,l} e_{ij}^l\right) \tag{9}$$

$$\check{h}_i^{l+1} = h_i^l + O_h^l\left(Concat_{k=1}^H\left(\sum_{j \in N_i} w_{ij}^{k,l} V_{ij}^{k,l} h_j^l\right)\right) \tag{10}$$

$$\check{e}_{ij}^{l+1} = e_{ij}^l + O_e^l\left(Concat_{k=1}^H\left(w_{ij}^{k,l}\right)\right) \tag{11}$$

Where $Q^{k,l},\ K^{k,l},\ V^{k,l},\ E^{k,l} \in R^{d_k \times d}$,$O_h^{\varrho},\ O_e^{\varrho} \in R^{d \times d}$ are parameters of the linear layers. $k = 1$ to H means the number of attention heads; d_k denotes the dimension of each head; *Norm* represents batch normalization operation and *Concat* represents concatenation operation. To ensure numerical stability, the outputs of softmax operation are clamped to a value between −5 and +5. Finally, $\check{h}_i^{l+1}$ and $\check{e}_{ij}^{l+1}$ are fed into feed forward networks with residue connections and batch normalization layers as:

$$h_i^{l+1} = \check{h}_i^{l+1} + W_{h2}^l\left(ReLU\left(W_{h1}^l Norm\left(\check{h}_i^{l+1}\right)\right)\right) \tag{12}$$

$$e_{ij}^{l+1} = \check{e}_{ij}^{l+1} + W_{e2}^l\left(ReLU\left(W_{e1}^l Norm\left(\check{e}_{ij}^{l+1}\right)\right)\right) \tag{13}$$

where W_{h1}^l,$W_{e1}^l \in R^{2d \times d}$ and $W_{h2}^l,\ W_{e2}^l \in R^{2d \times d}$. *Norm* serves as a normalization function, while *ReLU* functions as an activation function, enhancing the network's nonlinear expression capability. The atomic features X_d、X_p of drugs and proteins are obtained for further prediction by the above mentioned graph transformer module.

2.3. Attention module

We developed a dedicated attention module, known as GDS-Attention, to recognize and prioritize local dependencies between atoms and amino acids. It recognizes that certain atoms in a drug molecule may have stronger interactions with specific amino acids in a protein and vice versa. Different atom-amino acid pairs can be assigned different attention weights based on their importance in specific interactions, and the model can focus attention on these key interactions, thereby improving the ability to accurately predict drug-target binding. Unlike previous approaches that solely addressed the spatial dimension, GDS-Attention considers both the spatial and channel dimensions to model this interdependence. By utilizing the known drug potential characteristic matrix $D_{GraphTransformer}$ and protein potential characteristic matrix $P_{GraphTransformer} = \{p_1, p_2, \ldots, p_M\}$ from the Graph Transformer Module, we generate an attention matrix $A \in \mathbb{R}^{N \times M \times f}$. This attention matrix effectively captures drug-protein interactions across both spatial and channel dimensions.

More precisely, when given d_i and p_j, we initially convert them into attention vectors da_i and pa_j using a multilayer perceptron (MLP) to distinguish feature extraction from attention modeling.

$$da_i = F(W_d \bullet d_i + b) \tag{14}$$

$$pa_j = F\left(W_p \bullet p_j + b\right) \tag{15}$$

where F is a non-linear activation function, $W_d \in \mathbb{R}^{f \times f}$ and $W_p \in \mathbb{R}^{f \times f}$ represents the weight matrix, with b representing the bias vector, the attention vector $A_{i,j} \in \mathbb{R}^f$ is computed as follows:

$$A_{i,j} = F\left(W_a \bullet \left(da_i + pa_j\right) + b\right) \tag{16}$$

where $W_a \in R^{2f \times f}$ represents the weight matrix.

Following these operations, the attention matrix $A \in R^{N \times M \times f}$ is obtained. The attention matrix $A_d \in \mathbb{R}^{N \times f}$ for drugs and the attention matrix $A_p \in \mathbb{R}^{M \times f}$ for proteins are obtained by mean operations over different dimensions.

$$A_d = Sigmoid(MEAN(A, 2)) \tag{17}$$

$$A_p = Sigmoid(MEAN(A, 1)) \tag{18}$$

Where $MEAN(Input, \mathrm{dim})$ represents the mean operation, and $\mathrm{Sigmoid}$ is the activation function that maps any attention score to the range (0, 1). Subsequently, the latent feature matrices D_a and P_a are updated as follows:

$$D_a = D_{GraphTransformer} \bullet 0.5 + D_{GraphTransformer} \odot A_d \tag{19}$$

$$P_a = D_{GraphTransformer} \bullet 0.5 + D_{GraphTransformer} \odot A_p \tag{20}$$

where $\odot$ denotes the operation of element multiplication. Subsequently, the global mean-pooling layer operation is performed on D_a and P_a to derive the feature vectors υ_{drug} and $\upsilon_{protein}$, respectively.

2.4. Output module

In order to improve the ability to capture drug-drug associations (DDAs), we incorporate graphic-level Mol2vec (Johnson et al., 2010) features as a high-level representation of molecules. The use of pre-trained embeddings of proteins and molecules provides additional implicit information, thereby aiding in distinguishing between various proteins and proteins. Ultimately, the resulting feature vector of drugs, denoted as X_d, is obtained. On the basis of protein structure, this study also uses protein sequences for pre-training to enrich the multimodal information of proteins. In particular, we will introduce ESM-2 (Lin et al., 2022), a protein language model that uses a single gene sequence to predict protein structure, function, and other characteristics. Amino acid sequence was encoded into embedding vector using ESM-2 gene engineering model. Each vector encompasses a substantial amount of contextual information about the 1D protein sequence. To harness this, we embed the protein sequence into ESM-2 and utilize the pre-trained language model for ESM-2 embedding. Finally, the feature vector of the output protein is X_p.

We first take the vectors of output X_p and X_d and average them along the length of the sequence dimension to get enriched vectors of features for medicines and proteins. The obtained feature vectors $X_p^{'}$ and $X_d^{'}$ are connected to form the drug-target feature vector. In the classification module, the interaction vector is through a linear layer and activation function Tanh.

$$f = mean\left(X_p^{'}\right) \oplus mean\left(X_d^{'}\right) \tag{21}$$

$$s = Tanh(W_s f + b_s) \tag{22}$$

where the symbol $\oplus$ denotes the splicing operation, i.e., splicing two vectors in a specific dimension. *mean* denotes the mean operation; $W_s \in \mathbb{R}^{2d}$ signifies the weight matrix, while $b_s \in R$ represents the bias.

GTAMP-DTA utilizes the cross-entropy loss as its objective function to be minimized.

$$L = -\frac{1}{N}\sum_{i}^{N} y_i \log\sigma(s_i) + (1 - y_i)\log(1 - \sigma(s_i)) \tag{23}$$

where N is the number of samples, and $y_i \in \{0, 1\}$ is the label, identifying the indicator variable if the relation is the same as the relation of samples. We employ a lookahead optimizer (Zhang et al., 2019) along with the inner optimizer SGD to reduce the loss.

2.5. Datasets

GTAMP-DTA was evaluated on two publicly available datasets, namely the Davis dataset (Davis et al., 2011) and the KIBA dataset (Tang et al., 2014). These datasets are well-regarded benchmarks for drug-target affinity prediction in previous studies. The Davis dataset contains 30,056 interactions involving 442 proteins and 68 ligands. It offers selective measurements of the kinase protein family and its associated inhibitors, along with their corresponding dissociation constant values. The KIBA dataset comprises bioactivities of kinase inhibitors measured using the KIBA approach, which considers various efficacy indices of the inhibitors, such as K_i, K_d, and $IC_5 0$. The dataset includes binding affinities measured for interactions between 467 proteins and 52,498 ligands.

Öztürk et al (Öztürk et al., 2019). observed that in the KIBA dataset, 99% of the protein pairs had at most 60% Smith-Waterman (S-W) similarity, while for the Davis dataset, 92% of the protein pairs had at most 60% target similarity. These statistics indicate that both datasets are non-redundant. To ensure a fair assessment, we employed a 5-fold cross-validation method in our experiments. All data were divided equally into five subsets, where four subsets were used for training, and one subset was reserved for testing. Therefore, the dataset may be broken down into five scenarios, and each scenario was used to evaluate the suggested model. The final performance metric was then the average score.

3. Experimental results

3.1. Evaluation metrics

Consistency index (CI), proposed by GÖnen and Heller (GÖnen et al., 2005), is a model evaluation metric. It quantifies the disparity between the predicted output of a model and the actual results. The formula for the CI is as follows:

$$CI = \frac{1}{Z}\sum_{\delta_j > \delta_i} h(b_i - b_j) \tag{24}$$

In the equation, b_i represents the predicted value of δ_i, b_j represents the predicted value of δ_j, $h(x)$ denotes the step function, and Z stands for the normalized hyperparameters. Typically, the step function $h(x)$ is defined as follows:

$$h(x) = \begin{cases} 0, x < 0 \\ 0.5, x = 0 \\ 1, else \end{cases} \tag{25}$$

The mean squared error (MSE) is a statistical measure that directly quantifies the difference between estimated and true values. It is calculated as the expectation of the squared loss given a sample of n estimates and their corresponding true values:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{26}$$

where $\hat{y}_i$ represents the estimated value of the i_{th} sample and y_i represents the true value of the i_{th} sample.

In the WidedDTA model, the Pearson correlation coefficient to assess the deviation between the actual and predicted values of binding affinity (Solomon and Richard., 1951). Pearson's correlation coefficient was defined as:

$$Pearson = \frac{\mathrm{cov}(p, y)}{\sigma(p)\sigma(y)} \tag{27}$$

where cov represents the covariance between the predicted value p and the original value y, and σ denotes the standard deviation.

A metric used to assess a model's capability for external prediction is regression toward the mean (r_m2 index). It measures how much a variable, if it is initially very large, generally approaches the average the next time. It is determined how to calculate the r_m2 index:

$$r_m^2 = r^2 * \left(1 - \sqrt{(r^2 - r_0^2)}\right) \tag{28}$$

where r represents the squared correlation coefficient with an intercept, and r_0 denotes the coefficient without an intercept.

3.2. Result and discussion

The purpose of this section is to assess the performance of the GTAMP-DTA model using the Davis and KIBA datasets. We examined the max-pooling, mean-pooling, and sum-pooling approaches on both datasets to determine the model's optimal performance. Additionally, we performed a number of comparative tests utilizing the GTAMP-DTA model with or without attentional processes, or with various attentional mechanisms, to assess the efficacy of our suggested model. The baseline models included in this study were: SimBoost (He et al., 2017), DeepDTA (Öztürk et al., 2019), GraphDTA (Nguyen et al., 2020), WideDTA (Öztürk et al., 2019), DeepCDA (Abbasi et al., 2020), MT-DTI (Shin et al., 2019), MATT_DTI (Zeng et al., 2021) and FusionDTA (Yuan et al., 2022).

Table 1
Performance of the GTAMP-DTA and baseline models on the Davis dataset.

Model	CI	MSE	Pearson	r_m^2
SimBoost	0.872	0.282	-	0.644
DeepDTA	0.878	0.261	-	0.630
GraphDTA	0.886	0.229	-	-
DeepCDA	0.891	0.248	-	0.649
MT-DTI	0.887	0.245	-	0.665
MATT_DTI	0.891	0.227	-	0.683
WideDTA	0.886	0.262	0.820	-
FusionDTA	0.913	0.208	-	0.743
GTAMP-DTA	0.923	0.177	0.861	0.755

3.2.1. *The performance of GTAMP-DTA*

Table 1 compares the GTAMP-DTA model's performance with benchmark models while evaluating its performance on the Davis dataset. The results indicate that GTAMP-DTA outperforms existing models across all metrics. Specifically, the CI and Pearson of GTAMP-DTA improved by 0.020 and 0.041, respectively, and the MSE decreased by 0.021 compared to the baseline model WideDTA. Moreover, when compared to the base model FusionDTA, GTAMP-DTA shows a notable improvement in the r_m^2 index of 0.012.

The GTAMP-DTA model's and baseline models' findings on the KIBA dataset are shown in Table 2. The findings demonstrate that GTAMP-DTA model outperforms the baseline models significantly across all evaluation measures. In particular, the CI, MSE, and R2m of the GTAMP-DTA model demonstrated improvements of 0.011, 0.007, and 0.005, respectively, in comparison to the state-of-the-art model FusionDTA. Additionally, the GTAMP-DTA model achieves a 0.039 increase in Pearson index over the WideDTA baseline model.

Fig. 2 illustrates the comparison between true affinity and predicted values for both the Davis and KIBA datasets. The x-axis represents the ground truth, and the y-axis represents the predictions. The variance between the expected affinity and the actual value is shown by the vertical distance $|\Delta y|$ between each point and $y = x$. The distribution of the expected and actual affinities is shown by the histograms at the edges. The results show that the data points in both datasets tend to be symmetric around $y = x$, with a denser distribution around $y = x$ in the KIBA dataset.

The y-axis displays the predicted value for each data point, and the x-axis displays the actual value for each data point. Each sample's vertical distance $|\Delta y|$ from $y = x$ shows the difference between the predicted and actual values of its affinity.

3.2.2. *Performance comparison of different pooling layer methods*

Different pooling layer approaches in the model structure allow the model to concentrate on various portions of the intermediate sequence and update the layer parameters in accordance with various gradients. The three most popular pooling techniques are max-pooling, mean-pooling, and sum-pooling, which use maximizing, averaging, and sum functions, respectively, to group the feature maps of a sequence into tokens.

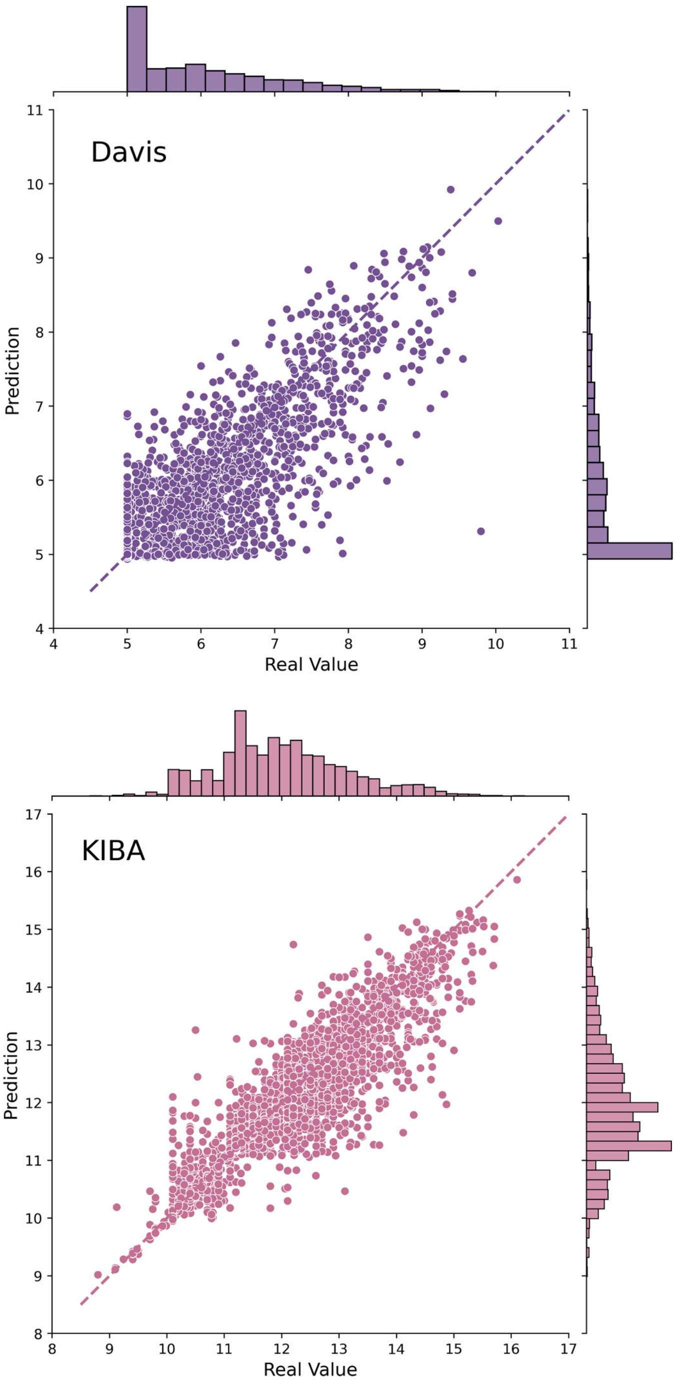


Fig. 2. The real affinity against the predicted value on Davis and KIBA datasets.

To examine the impact of different pooling methods on the performance of the proposed model, we conducted three controlled experiments during the validation stage. Notably, the model parameters for each experimental group were held constant, with the only variation being the utilization of distinct pooling methods. Fig. 3 illustrates the results of the experiments using max-pooling, mean-pooling, and sum-pooling on the Davis dataset. The results showed that the CI, MSE and r_m^2 of the mean-pooling layer reached 0.923, 0.177 and 0.755, compared to the CI, MSE and r_m^2 of the max-pooling layer were 0.879, 0.201 and 0.677, and the CI, MSE and r_m^2 of the sum-pooling layer reached 0.876, 0.225 and 0.664. For the Davis dataset, it is obvious that the mean-pooling layer outperforms any other two pooling strategies.

Fig. 4 depicts the performance results of max-pooling, mean-pooling, and sum-pooling on the KIBA dataset. The results demonstrate that the

Table 2
Performance of the GTAMP-DTA and baseline models on the KIBA dataset.

Model	CI	MSE	Pearson	r_m^2
SimBoost	0.836	0.222	-	0.629
DeepDTA	0.863	0.194	-	0.673
GraphDTA	0.891	0.139	-	-
DeepCDA	0.889	0.176	-	0.682
MT-DTI	0.882	0.152	-	0.738
MATT_DTI	0.889	0.150	-	0.756
WideDTA	0.875	0.179	0.856	-
FusionDTA	0.906	0.130	-	0.793
GTAMP-DTA	0.917	0.123	0.895	0.798

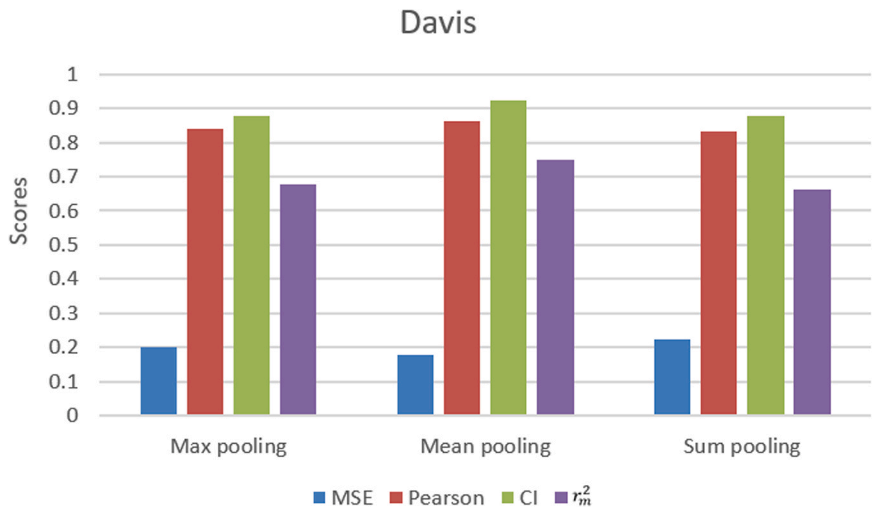


Fig. 3. Performance of different pooling methods on the Davis dataset.

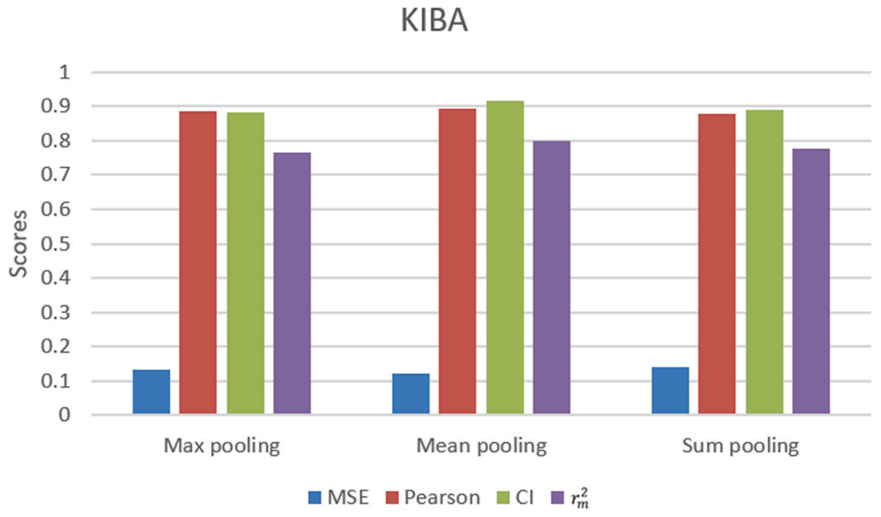


Fig. 4. Performance of different pooling methods on the KIBA dataset.

CI and r_m^2 of the mean-pooling layer reach 0.917 and 0.798, surpassing that of max-pooling and sum-pooling which have CI indices of 0.881 and 0.890, respectively, and r_m^2 of 0.767 and 0.775, respectively. Furthermore, the MSE of mean-pooling layer amounts to 0.123, which is lower than the MSE values of max-pooling (0.133) and sum-pooling (0.139). Thus, for the KIBA dataset, the mean-pooling layer employed as a feature aggregator exhibits superior performance to that of sum-pooling and max-pooling.

To summarize, the GTAMP-DTA model presented in this study demonstrates competitive or superior performance compared to other deep learning baseline models across all experimental settings. This is due to the utilization of the attention mechanism, which enables dynamic feature adjustments for drugs and proteins under various combinations, thus enhancing the model's predictive capacity.

3.3. *Importance of attention mechanism*

In order to evaluate the importance of attention and the performance of the model, we proposed two sub-models and conducted comparison tests with and without pre-trained vectors Its hyper-parameters were chosen as vector dimension of 256, number of attention heads of 4, learning rate of 0.0001, batch size of 50, and optimizer of SGD. The attention block is not included in the first sub-model, which is named No-Attention-DTA. In order to obtain drug and protein feature vectors, it directly applies global average pooling operation to the output of the transformer block. The output block is then provided with these vectors after they have been concatenated for prediction. The second sub-model is called Attention-DTA, which uses a bilateral attention mechanism. The attention weights are generated based on the drug feature matrix $D_{transformer}$ and the protein feature matrix $P_{transformer}$ as follows:

$$A_i = Sigmoid\left(D_{transformer} \bullet P_{transformer}^T\right) \tag{29}$$

The effectiveness of various models on the Davis dataset is shown in Table 3. Models with and without the attention mechanism are contrasted, and it becomes clear that the attention mechanism did actually lead to improvements. This emphasizes the significance of establishing connections between drug features and protein features in DTA

Table 3
Model Comparisons: With and Without Attention Block.

Methods	CI	MSE	Pearson	r_m^2
No-Attention-DTI	0.801	0.235	0.781	0.694
Attention-DTI	0.823	0.197	0.808	0.727
No-pre-trained-DTA	0.882	0.217	0.831	0.629
GTAMP-DTA	0.923	0.177	0.861	0.755

prediction. It is noteworthy that the GTAMP-DTA model performed the best, demonstrating that our suggested attention mechanism is more appropriate for use with transformer-based models than conventional attention methods. Additionally, we evaluated a number of activation functions in the block of attention and discovered that the ReLU function produced the best outcomes. This observation is most likely related to the protein and drug feature matrices that were extracted.

3.4. Visualization with attention weights

Our model may examine the protein-molecule interaction mechanism by utilizing the GTAMP-DTA architecture module. The positions highlighted by the self-attention mechanism provide a logical justification for the binding activity prediction and make it easier to quickly identify crucial protein-molecule interaction sites during further activity analysis.

We selected two different complexes from the RCSB Protein Data Bank (PDB) (Burley et al., 2019) to serve as examples. We used the attention weight that was determined using the molecular feature as the Query and the protein feature as the Key for this purpose. In order to gather attention data on proteins, we then computed the mean attention weight at the molecular level. To be more precise, we contrasted the violet-colored top-weighted residues of the model proteins and the ligand's atoms with the actual protein-ligand interaction sites retrieved from the PDB. The chemical atoms and amino acids with the greatest weights showed significant overlap with the actual contact locations. The attention bar in Fig. 5a for protein CA13 (UniProt ID: Q8N1Q1) indicates the residues HIS-121 and THR-201, which have a lot in common with the important pocket residues seen in the cocrystal complex (PDB: 4KNM). The highlighted residues (ARG-63, SER-69) and ligand functional groups in the significance maps for protein GCK (UniProt ID: P35557) in Fig. 5b show considerable resemblance to observed interactions in the cocrystal complex (PDB: 4DHY). The results suggest that the model can effectively analyze the interaction mechanism between molecules and target proteins, providing researchers with new ideas.

4. Conclusions

This study introduces the novel GTAMP-DTA model for predicting drug-target binding affinity by utilizing a graph converter and an attention mechanism. The model incorporates drug and protein map transducers to capture complex interactions and employs the attention mechanism to dynamically adjust the features of drugs and proteins. In order to improve the accuracy of DTA prediction, the model also makes use of protein and molecule pre-training data. We did a comparison analysis on two major datasets (Davis and KIBA) to evaluate the effectiveness of our model. We looked at the effects of various pooling layers (maximum pooling, average pooling, and total pooling), as well as attention mechanisms. The experimental results demonstrate a significant improvement in the performance of all metrics, including CI, MSE, Pearson, and r_m^2. Finally, we map weights of attention to protein sequences, which permits further investigation of how future medications will bind to target proteins and helps to reduce the scope of searches for

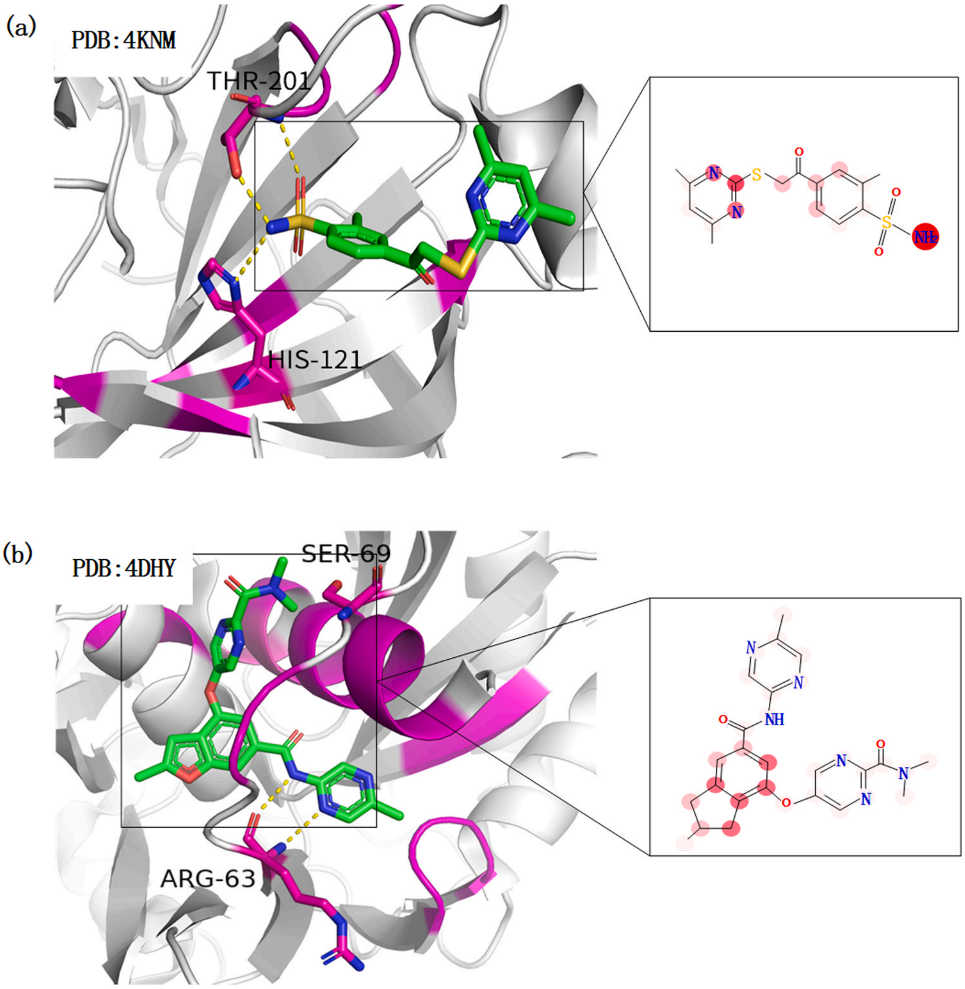


Fig. 5. Visualizing Attention Weights for Pocket and Ligand Pairs.

binding sites.

Even though GTAMP-DTA has shown to be effective in predicting DTA, here is still room for development. Firstly, identifying and extracting meaningful protein features remains a challenging but crucial task, as not all protein characteristics have a significant impact on prediction accuracy. Additionally, addressing information transmission between proteins and between proteins and drugs will be a key focus of future research efforts.

Author statement

I have made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; and I have drafted the work or revised it critically for important intellectual content; and I have approved the final version to be published; and I agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All persons who have made substantial contributions to the work reported in the manuscript, including those who provided editing and writing assistance but who are not authors, are named in the Acknowledgments section of the manuscript and have given their written permission to be named. If the manuscript does not include acknowledgments, it is because the authors have not received substantia contributions from nonauthors.

CRediT authorship contribution statement

Chuangchuang Tian: Conceptualization, Methodology, Data curation, Formal analysis, Validation, Visualization, Writing – original draft, Writing – review & editing. **Luping Wang**: Supervision, Writing - review & editing. **Zhiming Cui**: Supervision, Writing – review & editing. **Hongjie Wu**: Conceptualization, Resources, Supervision, Project administration, Funding acquisition, Reviewing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (62073231, 62176175), National Research Project (2020YFC2006602), Opening Topic Fund of Big Data Intelligent Engineering Laboratory of Jiangsu Province(SDGC2157), Provincial Key Laboratory for Computer Information Processing Technology , Soochow University(KJS2166).

References

Abbasi, K., Razzaghi, P., Poso, A., et al., 2020. Deepcda: deep cross-domain compound–protein affinity prediction through lstm and convolutional neural networks. Bioinformatics. https://doi.org/10.1093/bioinformatics/btaa544.

Abdel-Basset, M., Hawash, H., Elhoseny, M., et al., 2020. Deeph-dta: deep learning for predicting drug-target interactions: a case study of covid-19 drug repurposing. IEEE Access 8, 170433–170451. https://doi.org/10.1109/ACCESS.2020.3024238.

Bahdanau, D., Cho, K., Bengio, Y., , 2015. Neural machine translation by jointly learning to align and translate. Int. Conf. Learn. Represent. ICLR 10.48550/arXiv.1409.0473..

Bahi, M., Batouche, M., et al., 2021. Convolutional neural network with stacked autoencoders for predicting drug-target interaction and binding affinity. Int. J. Data Min. Model. Manag. 13 (1-2), 81–113. https://doi.org/10.1504/IJDMMM.2021.112914.

Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., et al., 2019. RCSB protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. 47, 464–474. https://doi.org/10.1093/nar/gky1004.

Chen, L., Tan, X., Wang, D., 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. https://doi.org/10.1093/bioinformatics/btaa524.

Chen, W., Chen, G., Zhao, L., et al., 2021. Predicting drug–target interactions with deep-embedding learning of graphs and sequences. J. Phys. Chem. A 125 (25), 5633–5642. https://doi.org/10.1021/acs.jpca.1c02419.

Davis, M.I., Hunt, J.P., Herrgard, S., et al., 2011. Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol. 29 (11), 1046–1051. https://doi.org/10.1038/nbt.1990.

Dhakal, Ashwin, McKay, Cole, Tanner, John J., et al., 2022. Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions. Brief. Bioinforma. 23 (1), bbab476 https://doi.org/10.1093/bib/bbab476.

Ding, J., Tang, J., Guo, F., 2019. Identification of drug-side effect association via semisupervised model and multiple kernel learning. IEEE J. Biomed. Health Inform. 23 (6), 2619–2632. https://doi.org/10.1109/JBHI.2018.2883834.

Ding, Y., Tang, J., Guo, F., et al., 2020a. Identification of drug–target interactions via dual laplacian regularized least squares with multiple kernel fusion. Knowl. Based Syst. 204, 106254 https://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 Comput. Appl. 32, 10303–10319. https://doi.org/10.1007/s00521-019-04569-z.

Ding, Y., Tang, J., Guo, F., et al., 2021. Identification of drug-target interactions via multi-view graph regularized link propagation model. Neurocomputing 461, 618–631. https://doi.org/10.1016/J.NEUCOM.2021.05.100.

V.P. Dwivedi, et al. A Generalization of transformer networks to graphs. arXiv preprint arXiv: 2012.09699, 2020. https://doi.org/10.48550/arXiv.2012.09699.

K.Y. Gao, A. Fokoue, H. Luo, et al. Interpretable Drug Target Prediction Using Deep Neural Representation. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence 2018; IJCAI-18, pages 3371-3377. https://doi.org/10.24963/ijcai.2018/468.

GÖnen, M., Heller, G., et al., 2005. Concordance probability and discriminatory power in proportional hazards regression. Biometrika 92 (4), 965–970. https://doi.org/10.1093/biomet/92.4.965.

He, T., Heidemeyer, M., Ban, F., et al., 2017. Simboost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. J. Chem. 9 (1), 1–14. https://doi.org/10.1186/s13321-017-0209-z.

Hu, J., Shen, L., Albanie, S., et al., 2020. Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. 42 (8), 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372.

Huang, K., Xiao, C., Glass, L.M., et al., 2021. MolTrans: molecular interaction transformer for drug–target interaction prediction. Bioinformatics 37 (6), 830–836. https://doi.org/10.1093/BIOINFORMATICS/BTAA880.

Jaeger, S., Fulle, S., Turk, S., et al., 2018. Mol2vec: unsupervised machine learning approach with chemical intuition. J. Chem. Imf. Model. 58, 27–35. https://doi.org/10.1021/acs.jcim.7b00616.

Johnson, E.R., Keinan, S., Mori-Sanchez, P., et al., 2010. Revealing noncovalent interactions. J. Am. Chem. SOC 132 (18), 6498–6506. https://doi.org/10.1021/ja100936w.

Kairys, V., Baranauskiene, L., Kazlauskiene, M., et al., 2019. Binding affinity in drug design: experimental and computational techniques. Expert Opin. Drug Discov. 14 (8), 755–768. https://doi.org/10.1080/17460441.2019.1623202.

Lin, X., Li, X., Lin, X., et al., 2020. A review on applications of computational methods in drug screening and design. Molecules 25 (6), 1375. https://doi.org/10.3390/molecules25061375.

Lin, Z.M., Akin, H., Rao, R.S., et al., 2022. Evolutionary-scale prediction of atomic level protein structure with a language model, 07.20.500902 bioRXiv Prepr. bioRXiv. https://doi.org/10.1101/2022.07.20.500902.

Newman, D.J., Cragg, G.M., et al., 2020. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. J. Nat. Prod. 83 (3), 770–803. https://doi.org/10.1021/acs.jnatprod.9b01285.

Nguyen, T., Le, H., Quinn, T.P., et al., 2020. Graphdta: predicting drug–target binding affinity with graph neural networks. Bioinformatics. https://doi.org/10.1093/bioinformatics/btaa921.

Öztürk, H., Özgür, A., Ozkirimli, E., et al., 2019. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics 34 (17), i821–i829. https://doi.org/10.1093/bioinformatics/bty593.

H. Öztürk, E. Ozkirimli, A. Özgür Widedta: prediction of drug-target binding affinity. arXiv preprint arXiv:1902.04166. 2019. https://doi.org/10.48550/arXiv.1902.04166.

B. Shin, S. Park, K. Kang, et al. Self-attention based molecule representation for predicting drug-target interaction. arXiv preprint arXiv:1908.06760, 2019.

Solomon, K., Richard, A.L., 1951. On information and sufficiency. Ann. Math. Stat. 22 (1), 79–86. https://doi.org/10.1016/S0022-5193(05)80753-2.

Sun, M., Prayag, T., Qian, Y., et al., 2022. MLapSVM-LBS: predicting DNA-binding proteins via a multiple Laplacian regularized support vector machine with local behavior similarity. Knowl. Based Syst. 250, 109174 https://doi.org/10.1016/J.KNOSYS.2022.109174.

Takebe, T., Imai, R., Ono, S., et al., 2018. The current status of drug discovery and development as originated in United States academia: the influence of industrial and academic collaboration on drug discovery and development. Clin. Transl. Sci. 11 (6), 597–606. https://doi.org/10.1111/cts.12577.

Tang, J., Szwajda, A., Shakyawar, S., et al., 2014. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J. Chem. Inf. Model. 54 (3), 735–743. https://doi.org/10.1021/ci400709d.

Tsubaki, M., Tomii, K., Sese, J., et al., 2019. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics 35 (2), 309–318. https://doi.org/10.1093/bioinformatics/bty535.

Vaswani, A., Shazeer, N., Parmar, N., et al., 2017. Attention is all you need. Proc. 31st Int. Conf. Neural Inf. Process. Syst. 6000–6010. https://doi.org/10.48550/arXiv.1706.03762.

Wang, H., Tang, J., Ding, Y., et al., 2021. Exploring associations of non-coding RNAs in human diseases via three-matrix factorization with hypergraph-regular terms on center kernel alignment. Brief. Bioinforma. 22 (5) https://doi.org/10.1093/BIB/BBAA409.

Wang, J., Li, X., Zhang, H., et al., 2020. GNN-PT: enhanced prediction of compound protein interactions by integrating protein transformer. arXiv. https://doi.org/10.48550/arXiv.2009.00805.

Wen, M., Zhang, Z., Niu, S., et al., 2017. Deep-learning- based drug-target interaction prediction. J. Proteome Res. 16 (4), 1401–1409.

Yadav, A.R., Mohite, S.K., et al., 2020. Homology modeling and generation of 3d-structure of protein. Res. J. Pharm. Dos. Forms Technol. 12 (4), 313–320. https://doi.org/10.5958/0975-4377.2020.00052.X.

Yang, H., Ding, Y., Tang, J., et al., 2021. Drug–disease associations prediction via multiple kernel-based dual graph regularized least squares. Appl. Softw. Comput. 112, 107811 https://doi.org/10.1016/J.ASOC.2021.107811.

Yuan, W., Chen, G., Chen, Y.C., 2022. FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity predictin. Brief. Bioinforma. 23, 1–13. https://doi.org/10.1093/BIB/BBAB506.

Zeng, Y., Chen, X., Luo, Y., et al., 2021. Deep drug-target binding affinity prediction with multiple attention blocks. Brief. Bioinform 22 (5), 1–10. https://doi.org/10.1093/BIB/BBAB117.

Zhang, M., Lucas, J., Ba, J., et al., 2019. Lookahead optimizer: k steps forward, 1 step back. In: Wallach, H., Arochelle, H.L., Eygelzimer, A.B. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc,, Vancouver, USA. https://doi.org/10.48550/arXiv.1907.08610.

Zhao, Q., Yang, M., Cheng, Z., Li, Y., et al., 2021b. Biomedical data and deep learning computational models for predicting compound-protein relations. IEEE/ACM Trans. Comput. Biol. Bioinforma. 19 (4), 2092–2110. https://doi.org/10.1109/TCBB.2021.3069040.

Zhao, Q., Yang, M., Cheng, Z., et al., 2021a. Biomedical data and deep learning computational models for predicting compound-protein relations. IEEE/ACM Trans. Comput. Biol. Bioinforma. https://doi.org/10.1109/TCBB.2021.3069040.

Zhen, L., Jiang, M., Wang, S., Zhang, S., 2022. Deep learning methods for molecular representation and property prediction. Drug Discov. Today, 103373. https://doi.org/10.1016/j.drudis.2022.103373.

Zhu, Z., Yao, Z., Qi, G., et al., 2023. Associative learning mechanism for drug-target interaction prediction. CAAI Trans. Intell. Technol. https://doi.org/10.1049/cit2.12194.