Bioinformatics, 38(3), 2022, 655–662
https://doi.org/10.1093/bioinformatics/btab715
Advance Access Publication Date: 19 October 2021
Original Paper

Sequence analysis

HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism

Qichang Zhao , Haochen Zhao, Kai Zheng and Jianxin Wang *

Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China

*To whom correspondence should be addressed.

Associate Editor: Jinbo Xu

Received on July 6, 2021; revised on September 24, 2021; editorial decision on October 10, 2021; accepted on October 13, 2021

Abstract

Motivation: Identifying drug–target interactions (DTIs) is a crucial step in drug repurposing and drug discovery. Accurately identifying DTIs *in silico* can significantly shorten development time and reduce costs. Recently, many sequence-based methods are proposed for DTI prediction and improve performance by introducing the attention mechanism. However, these methods only model single non-covalent inter-molecular interactions among drugs and proteins and ignore the complex interaction between atoms and amino acids.

Results: In this article, we propose an end-to-end bio-inspired model based on the convolutional neural network (CNN) and attention mechanism, named HyperAttentionDTI, for predicting DTIs. We use deep CNNs to learn the feature matrices of drugs and proteins. To model complex non-covalent inter-molecular interactions among atoms and amino acids, we utilize the attention mechanism on the feature matrices and assign an attention vector to each atom or amino acid. We evaluate HpyerAttentionDTI on three benchmark datasets and the results show that our model achieves significantly improved performance compared with the state-of-the-art baselines. Moreover, a case study on the human Gamma-aminobutyric acid receptors confirm that our model can be used as a powerful tool to predict DTIs.

Availability and implementation: The codes of our model are available at https://github.com/zhaoqichang/HpyerAttentionDTI and https://zenodo.org/record/5039589.

Contact: jxwang@mail.csu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Drug discovery and drug repurposing are highly valued in the current field of biomedicine (Agamah *et al.*, 2020; Ezzat *et al.*, 2019). Identifying drug–target interactions (DTIs) is a key step in drug discovery and drug repurposing (Zhao *et al.*, 2021). However, identifying DTIs in the wet lab is extremely costly and time-consuming due to the large-scale chemical space. To effectively shorten the time and reduce the cost, virtual screening (VS) has been developed to aid experimental drug discovery studies *in silico* (Rifaioglu *et al.*, 2019). Some classical VS methods, such as structure-based VS and ligand-based VS, have been studied for decades with great success in drug discovery (Himmat *et al.*, 2016; Maia *et al.*, 2020). However, the application of these VS methods is very limited. For example, structure-based VS methods cannot be performed when the 3D structure of a protein is unknown. Since accurately reconstructing the protein's structure is still a challenge, constructing structure-free approaches is gaining traction.

Recently, deep learning has achieved a superior performance compared with classical methods in many fields, such as computer vision and natural language processing (Hazra *et al.*, 2021; LeCun *et al.*, 2015). With the production of a large amount of biological activity data in recent years, predicting DTIs through deep learning technology becomes a research hotspot. In the early stages of deep learning applications in DTI prediction, researchers use manually crafted descriptors to represent drugs and proteins and design the fully connected neural network (FCNN) to make predictions (Tian *et al.*, 2016; Wen *et al.*, 2017). Because the descriptors are designed from a certain perspective and fixed during the training process, the descriptor-based methods could not extract task-related features. To this end, many end-to-end models are proposed. Lee *et al.* (2019)

proposed a model, called DeepConv-DTI, to predict DTIs. The model utilizes multi-scale one-dimensional convolutional neural network (1D-CNN) layers to obtain protein features and get Extended Connectivity Fingerprints (ECFP) (Rogers and Hahn, 2010) of drugs as input. To capture any relationship among atoms in a sequence, Shin *et al.* (2019) proposed a Transformer-based DTI model, which uses multi-layered bidirectional Transformer encoders (Vaswani *et al.*, 2017) to learn the high-dimensional structure of a molecule from the Simplified Molecular Input Line Entry System (SMILES) string. Zheng *et al.* (2020) proposed a framework to predict DTIs by representing proteins with 2D distance maps, and drugs with the SMILES string. From another perspective that processes the graph structure of compounds or proteins, GraphCPI (Quan *et al.*, 2019), Graph-CNN (Torng and Altman, 2019) and Lim *et al.* (2019) used graph neural networks (GNNs) to learn the representation of compounds and proteins.

Deep learning has driven the rapid development of drug–target affinity (DTA) prediction. Öztürk *et al.* (2018) proposed a model to predict DTAs, which processes the SMILES strings and amino acid sequences by CNN blocks. To extract the topological structure information of drugs, Nguyen *et al.* (2020) proposed GraphDTA for DTA prediction that represented drugs as graphs. GraphDTA uses various GNNs to capture the structural information of drugs and CNN to learn protein features. Furthermore, Abdel-Basset *et al.* (2020) proposed a heterogeneous graph attention model to learn topological information of drugs and Bi-ConvLSTM layers to model spatial-sequential information from SMILES strings. They used a dense CNN followed by SE-Block (Hu *et al.*, 2020) for protein sequences. The above methods tried to explore stronger modules to extract drug or protein features, but ignored the important fact that only certain parts of a protein or several atoms of a drug are involved in the intermolecular interactions, rather than the whole structure.

To model the inter-molecular interactions between amino acids and atoms, the attention mechanism (Bahdanau *et al.*, 2015) is introduced in DTI and DTA prediction. Tsubaki *et al.* (2019) proposed an attention-based model for DTI prediction. This model encodes a drug to a fixed-length vector and use the one-side attention mechanism to compute which subsequences in the protein are more important for the molecule. The models proposed by Chen *et al.* (2021) and Wang *et al.* (2020) also utilized this kind of attention mechanism. Furthermore, Gao *et al.* (2018) applied a two-side attention mechanism in DTI prediction to enable drugs and proteins to be aware of each other. The two-side attention mechanism can not only locate binding sites on proteins, but also explore the importance of atoms on drugs and is applied in DTA prediction (Abbasi *et al.*, 2020; Zhao *et al.*, 2019), recently. Inspired by the great ability of Transformer (Vaswani *et al.*, 2017) in capturing features between two sequences, Chen *et al.* (2020) regarded drugs and proteins as two kinds of sequences and proposed a Transformer-based model, named TransformerCPI, to predict DTIs. Huang *et al.* (2021) also proposed a Transformer-based model, MolTrans, which introduces the Transformer encoder in the feature extraction process to capture the semantic relations among substructures in drugs or proteins. These methods incorporate the attention mechanism to model single non-covalent inter-molecular interactions among drugs and proteins and get better performance than the models without attention mechanism. But, they ignore the fact that there are several non-covalent interaction types between drugs and proteins (e.g. hydrophobic interactions, hydrogen bonds and π-stackings).

Inspired by previous attention-based models (Chen *et al.*, 2020; Gao *et al.*, 2018; Huang *et al.*, 2021; Tsubaki *et al.*, 2019), we propose a bio-inspired end-to-end approach, named HpyerAttentionDTI, to predict DTIs. The input of our model is the SMILES string of drugs and amino acid sequence of proteins. We use stacked 1D-CNN layers to learn feature matrices from the input. Different from previous attention-based models, our model infers an attention vector for each amino acid–atom pair. These attention vectors not only present the interactions between amino acids and atoms, but also control the representation of features on the channel. After the attention block, the modified drug–protein feature vector is fed into fully connected neural networks to predict the DTIs. We compare our model with the state-of-the-art deep learning baselines on three wide-used datasets under four different drug discovery settings. The results show that HpyerAttentionDTI has competitive performance against the baselines under all settings. We perform a case study on the human Gamma-aminobutyric acid receptors to evaluate the ability of our model to predict existing DTIs. Moreover, we further visualize the attention score learned by HpyerAttentionDTI, and the results show that the attention block of our model is useful in reducing the search space for binding sites.

2 Materials and methods

2.1 Benchmark datasets

We extract the drug and target data from the DrugBank database (Wishart *et al.*, 2006) to establish the dataset of experiments. The data used in this study were released on January 3, 2020 (version 5.1.5). We manually discard the drugs which are inorganic compounds, very small molecule compounds [e.g. Iron (DB01592) and Zinc (DB01593)] or those of which the SMILES string cannot be recognized by RDKit python package (Landrum, 2006). Finally, 6655 drugs, 4294 proteins and 17 511 positive DTIs are obtained in total. Following common practice (Huang *et al.*, 2021; Wen *et al.*, 2017), we sample from the unlabeled drug–protein pairs to generate negative samples and obtain a balanced dataset with equal positive and negative samples. Moreover, we also construct two unbalanced benchmark datasets, Davis (Davis *et al.*, 2011) and KIBA (Tang *et al.*, 2014). Davis and KIBA record wet lab assay values measuring binding affinities among drugs and proteins. Following early works (Davis *et al.*, 2011; Öztürk *et al.*, 2018), the thresholds 5.0 and 12.1 are set for the Davis and KIBA datasets, respectively, to construct binary classification datasets. Because there are some proteins with the same amino acid sequence in the Davis dataset, we remove the duplicate drug-protein pairs to avoid label confusion. Table 1 summarizes DrugBank dataset, Davis dataset and KIBA dataset.

2.2 Proposed model

HyperAttentionDTI consists of three parts: CNN block, attention block and output block. Given the drug's SMILES strings and protein's amino acid sequences, CNN block extracts feature matrices from the sequences of drugs and proteins. Then the feature matrices are feed into the attention block to get a decision vector. Finally, the output block performs prediction according to the decision vector. An overview of the proposed HyperAttentionDTI is depicted in Figure 1.

2.2.1 Embedding layer

The amino acid sequences of proteins and SMILES strings of drugs are the input of HyperAttentionDTI. The SMILES strings of drugs are made up of 64 different characters, and there are 20 different amino acids in proteins. HyperAttentionDTI starts with two embedding layers to transform each amino acid and SMILES character to the corresponding embedding vector. After embedding layers, we get the embedding matrix $P_e \in \mathbb{R}^{M\times ep}$ for protein and $D_e \in \mathbb{R}^{N\times ed}$ for drug, where M and N are the lengths of protein and drug strings, respectively, and ep and ed are the sizes of embedding for protein and drug strings, respectively.

2.2.2 CNN block

There are two independent CNN blocks in our model, one for drugs and one for proteins. The CNN block contains three consecutive 1D-CNN layers, which can efficiently extract sequence semantic

Table 1. Summary of the benchmark datasets

Datasets	Protein	Drug	Interaction	Positive	Negative
DrugBank	4294	6655	35 022	17 511	17 511
Davis	379	68	25 772	7320	18 452
KIBA	225	2068	116 350	22 154	94 196

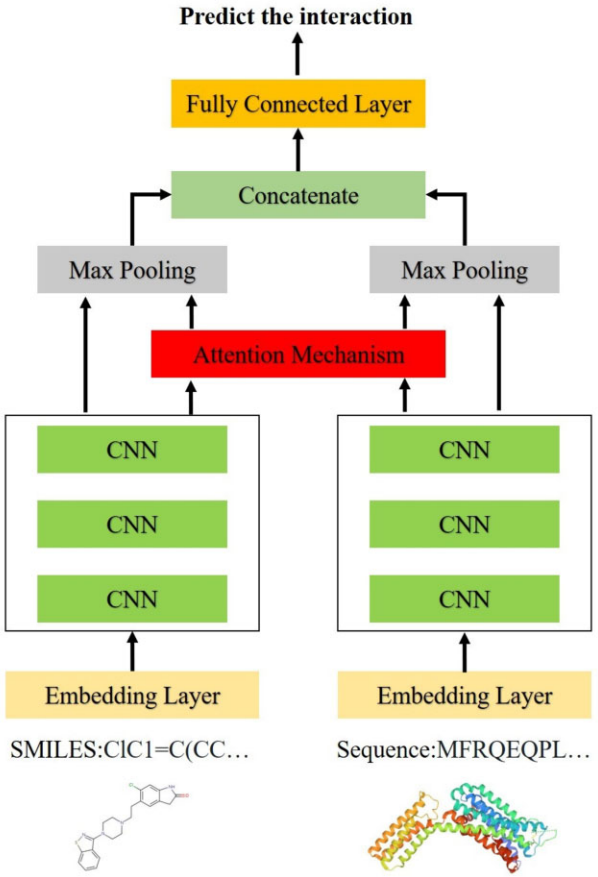

Fig. 1. The network architecture of HyperAttentionDTI

information (Kim, 2014). 1D-CNN is capable to capture important local patterns from the whole space. When the convolution filter slides on proteins or SMILES strings, the different combinations of amino acids or substructure of the drug are captured to get the latent feature vectors which contain chemical relationships among themselves. Given the embedding matrices P_e and D_e from embedding layers, CNN blocks generate the latent feature matrices $P_{\text{cnn}} \in \mathbb{R}^{M \times f}$ for proteins and $D_{\text{cnn}} \in \mathbb{R}^{N \times f}$ for drugs, where f is the number of filters of the last 1D-CNN layer.

2.2.3 Attention block

We design a special attention module, called HyperAttention. Different with the attention mechanisms in early works (Huang *et al.*, 2021; Tsubaki *et al.*, 2019), HyperAttention models the semantic interdependencies not only in spatial dimensions but also in channel dimensions between drug subsequences and protein subsequences. Given the latent feature matrices D_{cnn} for drugs and $P_{\text{cnn}} = \{p_1, p_2, ..., p_M\}$ for proteins from CNN blocks, we generate an attention matrix $A \in \mathbb{R}^{N \times M \times f}$, which contains the interactions between a drug and a protein in spatial and channel dimensions.

More precisely, given d_i and p_j, we first transform them into attention vectors, da_i and pa_j, by multi-layer perceptron (MLP) to separate the feature extractor and attention modeling.

$$da_i = F(W_d \cdot d_i + b) \tag{1}$$

$$pa_j = F(W_p \cdot p_j + b) \tag{2}$$

where F is a non-linear activation function (e.g. ReLU), $W_d \in \mathbb{R}^{f \times f}$ and $W_p \in \mathbb{R}^{f \times f}$ are the weight matrices, and b is the bias vector. Then, the attention vector $A_{i,j} \in \mathbb{R}^f$ is calculated as:

$$A_{i,j} = F(W_a \cdot (da_i + pa_j) + b), \tag{3}$$

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

After these operations, we get the attention matrix $A \in \mathbb{R}^{N \times M \times f}$. By performing mean operations on different dimensions, the attention matrix $A_d \in \mathbb{R}^{N \times f}$ for drugs and $A_p \in \mathbb{R}^{M \times f}$ for proteins are generated.

$$A_d = \text{Sigmoid}(MEAN(A, 2)), \tag{4}$$

$$A_p = \text{Sigmoid}(MEAN(A, 1)), \tag{5}$$

where $MEAN(Input, dim)$ is the mean operation which returns the mean value of each row of *Input* in the given dimension *dim*, and *Sigmoid* is the activate function to map any attention score to the range (0, 1). The latent feature matrices D_a and P_a are updated:

$$D_a = D_{\text{cnn}} \cdot 0.5 + D_{\text{cnn}} \odot A_d, \tag{6}$$

$$P_a = P_{\text{cnn}} \cdot 0.5 + D_{\text{cnn}} \odot A_p, \tag{7}$$

where $\odot$ denotes element-wise product. We then apply a global-max pooling operation over D_a and P_a to obtain the feature vectors, v_{drug} and v_{protein}, which are concatenated and feed into the output block.

2.2.4 Output block

The output block consists of multilayer fully connected neural networks (FCNNs). The activation function of FCNN is Leaky Rectified Linear Unit (Leaky ReLU) (He *et al.*, 2015) with the negative slope of 0.01. Each FCNN is followed by a Dropout layer to prevent overfitting. The last layer output the probability $\hat{y}$ indicating the likelihood of interaction. As a binary classification task, we use binary cross entropy loss to train our model:

$$\text{loss} = -[y\log(\hat{y}) + (1 - y)\log(1 - (\hat{y}))], \tag{8}$$

where y is the ground truth label.

2.3 Implementation

HyperAttentionDTI is implemented in PyTorch (Paszke *et al.*, 2019). For optimization parameters, we use the AdamW optimizer (Loshchilov and Hutter, 2019) with the default learning rate of 1e-4 and the weight decay coefficient of 1e-4. The input embedding is of size 64, which means that we represent each character in SMILES or amino acid sequence with a 64-dimensional dense vector. Each CNN block consists of three stacked 1D-CNN layers with 32, 64 and 96 filters, respectively. And the window sizes of the CNN block are 4, 6, 8 for drugs and 4, 6, 12 for proteins. The output block consists of four fully connected layers, in which the numbers of neurons are 1024, 1024, 512 and 2, respectively. The dropout rate is 0.1. We set the batch size to be 32. We perform early stopping to solve the overfitting problem. If the loss of models on the validation set does not decrease within 20 epochs, the training will stop.

3 Experiments and results

3.1 Experimental setup

Metrics: As DTI prediction is a classification task, we use the accuracy, precision, recall, AUC (Area under the receiver operating characteristic curve) and AUPR (Area under the precision-recall curve) as metrics to measure the performance of models. The best results are highlighted in bold for each metric.

Evaluation strategies: Suppose P_{train} and D_{train} are the sets of proteins and drugs in the training set. When predicting the interaction between a drug d and a protein p in the testing set, there are four different experimental settings to make a comprehensive comparison:

- E_1. Both d and p appear in the training set: $d \in D_{\text{train}}$ and $p \in P_{\text{train}}$.

- E_2. There are no interactions of drug d in the training set, while there are interactions of protein p in the training set: $d \notin D_{\text{train}}$ and $p \in P_{\text{train}}$.
- E_3. There are no interactions of protein p in the training set, while there are interactions of drug d in the training set: $d \in D_{\text{train}}$ and $p \notin P_{\text{train}}$.
- E_4. Neither d nor p appear in the training dataset: $d \notin D_{\text{train}}$ and $p \notin P_{\text{train}}$.

We perform 10 times repeated 5-fold cross-validation to assess the predictive ability of models. For each time, we conduct different random split of datasets under different random seeds.

The search of hyper-parameters: There are four important hyper-parameters in our model, namely the learning rate, the weight decay coefficient, the batch size and the dropout rate. These hyper-parameters are determined by grid-search on the DrugBank dataset. In grid-search, the learning rate is in [1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7], the batch size is in [8, 16,32, 64, 128, 256, 512], the weight decay coefficient is in [1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7] and the dropout rate is in [0.1, 0.2, 0.3, 0.4, 0.5]. In general, the learning rate directly determines the performance and the batch size are correlated with the learning rate. So, we first determine the learning rate and the batch size in a grid-search. After the learning rate and batch size are fixed, we select the weight decay coefficient and the dropout rate to improve the robustness of our model. The optimized learning rate, weight decay coefficient, batch size and dropout rate are 1e-4, 1e-4, 32 and 0.1, respectively.

3.2 Baselines

GNN-CPI (Tsubaki *et al.*, 2019)**:** GNN-CPI encodes drugs and proteins by graph neural network and 1D-CNNs, respectively, and use the one-side attention mechanism to consider which subsequences in a protein are important for a drug. And the feature vectors of drugs and proteins are concatenated and feed into FCNN to predict DTIs.

GNN-PT (Wang *et al.*, 2020)**:** GNN-PT utilizes GNNs to extract drug vectors and Transformer and CNN for protein representation. A one-side attention mechanism is used to get the protein vector. It then is concatenated with the drug vector and fed into the FCNN for final prediction.

DeepEmbedding-DTI (Chen *et al.*, 2021)**:** DeepEmbedding-DTI encodes drugs and proteins by GNNs and BiLSTM, respectively, and use the transformer-based model to learn embedding vectors of protein sequences.

GraphDTA (Nguyen *et al.*, 2020)**:** GraphDTA applies GNNs and CNNs for drug and protein representation, respectively. According to the reports provided in Nguyen *et al.* (2020), we choose GAT_GCN as the extractor in the later experiment comparison. We add a Sigmoid activation function after the last layer to change it for DTI prediction.

DeepConv-DTI (Lee *et al.*, 2019)**:** DeepConv-DTI utilizes FCNNs to process ECFP fingerprints of drugs and applies multi-scale 1D-CNN and global max-pooling layer to extract various length local pattern in protein sequences. Then the abstract feature vectors of drugs and proteins are concatenated and feed into FCNNs to predict DTIs.

TransformerCPI (Chen *et al.*, 2020)**:** TransformerCPI is based on the Transformer architecture, which regards drugs and proteins as two kinds of sequences. After generating protein sequence representation and atom representation from CNN and GCN, respectively, TransformerCPI gets interaction features through the decoder of Transformer and uses linear layers to output the interaction probability.

MolTrans (Huang *et al.*, 2021)**:** MolTrans is also based on the Transformer architecture. MolTrans uses frequent consecutive subsequence mining module to decompose drugs and proteins into a set of explicit sequences of substructures. Then it utilizes Transformer embedding modules to obtain the augmented contextual embedding for drugs and proteins. Next, MolTrans models the interaction map by dot product and applies CNN and FCNN on the interaction map to predict DTIs.

We follow the same hyper-parameter setting described in the papers of baselines.

3.3 Performance evaluation under the setting E_1

We first compare our model with baselines on the DrugBank dataset under the setting E_1. We divide DrugBank dataset into training, validation and testing sets in a 16:4:5 ratio. Each model is evaluated on the testing set when its performance no longer improves on the validation set. Table 2 shows the accuracy, precision, recall, AUC and AUPR results of various baselines and the proposed method on the DrugBank dataset. From Table 2, our approach gets 0.023, 0.007, 0.012, 0.028 and 0.018 improvements in accuracy, precision, recall, AUC and AUPR over the best performance of baselines, respectively.

To reduce bias and noises from the random generation of negative DTIs, we compare our model with baselines on the Davis and KIBA datasets under the setting E_1. These two datasets are unbalanced. Tables 3 and 4 show the results of these models on the Davis and KIBA datasets, respectively. On Davis and KIBA, MolTrans and GNN-CPI get the highest precision, respectively, but our model far outperforms baselines on other metrics. On Davis dataset, our model gets 0.024, 0.074, 0.02 and 0.055 improvements in accuracy, recall, AUC and AUPR over the best performance of baselines, respectively. On KIBA dataset, our model gets 0.004, 0.149, 0.021 and, 0.041 improvements in accuracy, recall, AUC and AUPR over the best performance of baselines, respectively.

3.4 Performance evaluation under *de novo* setting

To test the robustness of our model, we evaluate our model and baselines under the setting E_2, E_3 and E_4 on the DrugBank dataset. The settings E_2 and E_3 are more realistic situation than the setting E_1 on drug discovery. To test these models under the setting E_2/E_3, we randomly select 20% drugs/proteins and treat all DTIs associated with these drugs/proteins as the testing set. The other DTIs are used as training set and validation set, with a ratio of 4:1. The results under the setting E_2 and E_3 are in Tables 5 and 6, respectively. These results suggested that HpyerAttentionDTI achieve superior performance over the state-of-the-art baselines, with capability to handle realistic situation on drug discovery.

In the most challenging setting E_4, we randomly select 20% drugs and 20% proteins and treat all DTIs composed of these drugs and proteins as the testing set. And the DTIs, which are not associated with these drugs and proteins, are used as training set and validation set, with a ratio of 4:1. The results under the setting E_4 are described in Table 7. As we can see, there is a significant performance decline in all models, but our model still performs the best performance.

We also compare our model with the baselines under these *de novo* settings on the Davis and KIBA datasets and get the similar results (the results are available in Supplementary Tables S2–S7). Overall, our model achieves competitive or greater performance against the state-of-the-art deep learning baselines in all settings. The reason is that our model uses attentional mechanism to dynamically adjust the features of drugs and proteins in different combinations of them, compared with the baselines in which the extracted features of drugs and proteins are fixed.

3.5 The effectiveness of attention block

To evaluate the importance of the attention block, we propose three sub-models. The first one is No_Attention_DTI, in which there is no attention module. We directly apply global-max pooling operations over the output of the CNN blocks to get drug and protein feature vectors. We concatenate vectors and feed them into Output block to make predictions. The second sub-model is based on the two-side attention mechanism, named Attention_DTI. Given the drug feature matrix D_{cnn} and protein feature matrix P_{cnn}, the attention weight is generated such that

Table 2. Comparison results of the proposed model and baselines on the DrugBank dataset under the setting E_1

Methods	Accuracy (Std)	Precision (Std)	Recall (Std)	AUC (Std)	AUPR (Std)
GNN-CPI	0.731 (0.005)	0.737 (0.005)	0.716 (0.004)	0.802 (0.005)	0.811 (0.004)
GNN-PT	0.754 (0.001)	0.726 (0.007)	0.817 (0.010)	0.839 (0.006)	0.839 (0.012)
DeepEmbedding-DTI	0.758 (0.009)	0.768 (0.015)	0.738 (0.015)	0.841 (0.009)	0.848 (0.005)
GraphDTA	0.757 (0.001)	0.751 (0.010)	0.769 (0.011)	0.821 (0.002)	0.797 (0.001)
DeepConv-DTI	0.770 (0.003)	0.792 (0.003)	0.736 (0.004)	0.845 (0.003)	0.844 (0.002)
TransformerCPI	0.764 (0.002)	0.750 (0.003)	0.792 (0.003)	0.837 (0.003)	0.836 (0.002)
MolTrans	0.787 (0.002)	0.786 (0.002)	0.792 (0.002)	0.861 (0.002)	0.856 (0.002)
HyperAttentionDTI	**0.810** (0.002)	**0.799** (0.001)	**0.829** (0.001)	**0.889** (0.001)	**0.897** (0.001)

Table 3. Comparison results of the proposed model and baselines on the Davis dataset under the setting E_1

Methods	Accuracy (Std)	Precision (Std)	Recall (Std)	AUC (Std)	AUPR (Std)
GNN-CPI	0.819 (0.001)	0.731 (0.002)	0.570 (0.002)	0.863 (0.001)	0.745 (0.002)
GNN-PT	0.827 (0.001)	0.693 (0.020)	0.706 (0.021)	0.882 (0.007)	0.774 (0.010)
DeepEmbedding-DTI	0.836 (0.008)	0.760 (0.017)	0.618 (0.024)	0.878 (0.011)	0.775 (0.020)
GraphDTA	0.817 (0.001)	0.743 (0.014)	0.530 (0.017)	0.859 (0.004)	0.743 (0.007)
DeepConv-DTI	0.830 (0.001)	0.750 (0.002)	0.698 (0.001)	0.8669 (0.001)	0.777 (0.001)
TransformerCPI	0.822 (0.001)	0.688 (0.003)	0.688 (0.003)	0.877 (0.001)	0.767 (0.001)
MolTrans	0.842 (0.00)	**0.782** (0.003)	0.617 (0.004)	0.900 (0.001)	0.784 (0.002)
HyperAttentionDTI	**0.866** (0.001)	0.754 (0.002)	**0.780** (0.001)	**0.920** (0.001)	**0.839** (0.001)

Table 4. Comparison results of the proposed model and baselines on the KIBA dataset under the setting E_1

Methods	Accuracy (Std)	Precision (Std)	Recall (Std)	AUC (Std)	AUPR (Std)
GNN-CPI	0.867 (0.002)	0.727 (0.002)	0.477 (0.007)	0.864 (0.005)	0.673 (0.005)
GNN-PT	0.876 (0.005)	0.691 (0.006)	0.647 (0.007)	0.901 (0.002)	0.741 (0.005)
DeepEmbedding-DTI	0.878 (0.002)	0.741 (0.005)	0.556 (0.016)	0.889 (0.003)	0.727 (0.006)
GraphDTA	0.889 (0.001)	**0.775** (0.020)	0.594 (0.032)	0.914 (0.001)	0.776 (0.007)
DeepConv-DTI	0.878 (0.001)	0.708 (0.002)	0.636 (0.003)	0.898 (0.001)	0.703 (0.001)
TransformerCPI	0.870 (0.001)	0.669 (0.003)	0.631 (0.003)	0.888 (0.001)	0.708 (0.001)
MolTrans	0.881 (0.001)	0.710 (0.003)	0.645 (0.003)	0.905 (0.001)	0.708 (0.003)
HyperAttentionDTI	**0.893** (0.001)	0.689 (0.001)	**0.796** (0.002)	**0.935** (0.001)	**0.814** (0.002)

Table 5. Comparison results of the proposed model and baselines on the DrugBank dataset under the setting E_2

Methods	Accuracy (Std)	Precision (Std)	Recall (Std)	AUC (Std)	AUPR (Std)
GNN-CPI	0.618 (0.035)	0.649 (0.047)	0.522 (0.064)	0.662 (0.040)	0.692 (0.035)
GNN-PT	0.615 (0.008)	0.675 (0.013)	0.436 (0.036)	0.655 (0.011)	0.684 (0.019)
DeepEmbedding-DTI	0.655 (0.019)	0.699 (0.046)	0.538 (0.064)	0.713 (0.027)	0.729 (0.045)
GraphDTA	0.577 (0.003)	0.635 (0.023)	0.361 (0.023)	0.616 (0.021)	0.621 (0.0.19)
DeepConv-DTI	0.658 (0.022)	0.710 (0.025)	0.535 (0.016)	0.704 (0.023)	0.728 (0.011)
TransformerCPI	0.648 (0.011)	0.685 (0.021)	0.541 (0.022)	0.702 (0.023)	0.670 (0.022)
MolTrans	0.662 (0.011)	0.732 (0.012)	0.584 (0.023)	0.726 (0.022)	0.745 (0.023)
HyperAttentionDTI	**0.718** (0.011)	**0.774** (0.023)	**0.612** (0.011)	**0.785** (0.011)	**0.809** (0.013)

Table 6. Comparison results of the proposed model and baselines on the DrugBank dataset under the setting E_3

Methods	Accuracy (Std)	Precision (Std)	Recall (Std)	AUC (Std)	AUPR (Std)
GNN-CPI	0.655 (0.059)	0.673 (0.057)	0.613 (0.065)	0.716 (0.065)	0.725 (0.076)
GNN-PT	0.713 (0.021)	0.751 (0.026)	0.641 (0.063)	0.777 (0.025)	0.783 (0.016)
DeepEmbedding-DTI	0.711 (0.013)	0.743 (0.035)	0.644 (0.056)	0.791 (0.011)	0.795 (0.012)
GraphDTA	0.731 (0.004)	0.737 (0.012)	0.716 (0.047)	0.801 (0.022)	0.799 (0.011)
DeepConv-DTI	0.719 (0.031)	0.731 (0.031)	0.699 (0.024)	0.788 (0.021)	0.797 (0.025)
TransformerCPI	0.735 (0.032)	0.730 (0.031)	**0.742** (0.035)	0.799 (0.028)	0.796 (0.023)
MolTrans	0.728 (0.024)	0.755 (0.021)	0.673 (0.023)	0.794 (0.021)	0.804 (0.019)
HyperAttentionDTI	**0.743** (0.015)	**0.765** (0.017)	0.698 (0.021)	**0.818** (0.014)	**0.834** (0.011)

Table 7. Comparison results of the proposed model and baselines on the DrugBank dataset under the setting E_4

Methods	Accuracy (Std)	Precision (Std)	Recall (Std)	AUC (Std)	AUPR (Std)
GNN-CPI	0.557 (0.023)	0.586 (0.027)	0.418 (0.078)	0.589 (0.030)	0.607 (0.015)
GNN-PT	0.544 (0.016)	0.589 (0.047)	0.313 (0.037)	0.580 (0.021)	0.590 (0.037)
DeepEmbedding-DTI	0.596 (0.023)	0.666 (0.050)	0.442 (0.079)	0.665 (0.026)	0.669 (0.046)
GraphDTA	0.544 (0.003)	0.586 (0.027)	0.307 (0.018)	0.579 (0.025)	0.579 (0.023)
DeepConv-DTI	0.585 (0.033)	0.631 (0.022)	0.415 (0.025)	0.638 (0.022)	0.622 (0.021)
TransformerCPI	0.611 (0.022)	0.669 (0.023)	**0.467** (0.024)	0.657 (0.027)	0.664 (0.025)
MolTrans	0.623 (0.021)	0.705 (0.024)	0.450 (0.026)	0.685 (0.022)	0.706 (0.025)
HyperAttentionDTI	**0.647** (0.021)	**0.752** (0.019)	0.455 (0.023)	**0.723** (0.018)	**0.741** (0.023)

$$A_i = \text{Sigmoid}(D_{\text{cnn}} \cdot P_{\text{cnn}}^{\text{T}}), \tag{9}$$

The third one is based on the multi-head attention mechanism, named Multi-Head_DTI. For each head i, the attention weight is generated such that

$$Da_i = F(Wd_i \cdot D_{\text{cnn}} + b), \tag{10}$$

$$Pa_i = F(Wp_i \cdot P_{\text{cnn}} + b), \tag{11}$$

$$A_i = \text{Sigmoid}(Da_i \cdot Pa_i^{\text{T}}), \tag{12}$$

where $Wd_i \in \mathbb{R}^{f \times d}$ and $Wp_i \in \mathbb{R}^{f \times d}$ are the weight matrix and T is the transpose operation. And the final attention weight $A \in \mathbb{N}^{M \times f}$ is generated by

$$A = \frac{1}{K} \cdot \sum_{i=0}^{K} A_i, \tag{13}$$

where K is the number of heads. By definition, the multi-head attention mechanism also has the ability to simulate the complex interaction between atoms and amino acids. However, this mechanism introduces a large number of model parameters, the number of which depends on hyperparameter K. Empirically, we set K to be 8.

Table 8 shows the prediction performance among these models on the DrugBank dataset. As it is shown, by comparing models with and without attention mechanism, it is concluded that utilizing attention mechanism achieves improvements indeed. This indicates that it is necessary to establish the correlation between drug features and protein features in DTI inference process. Moreover, HyperAttentionDTI gets the best performance, which indicates that our proposed attention mechanism is more suitable to be applied to the CNN-based model than the conventional attention mechanisms. It is worth mentioning that we tried different activation functions in the attention block and found that the ReLU function achieved the best result. We speculate that it is related to the extracted drug and protein feature matrices.

3.6 Case studies

To evaluate the reliability of our model, we perform a case study using actual FDA-approved drugs targeting specific proteins, the human Gamma-aminobutyric acid receptors (GABARs). GABARs are selected to perform the case study, since they are the most important inhibitory chloride ion channels in the central nervous system and are major targets for a wide variety of drugs (Sigel and Steinmann, 2012; Zhu *et al.*, 2018). There are 7 subunits and 16 target proteins in GABARs. We download 11 172 compounds from the latest version 5.1.8 (released on Jan.3, 2021) of the DrugBank database. We filter out the compounds as described in subsection 2.1. As a result, 6708 drugs are used for this case study. We trained our model on the Drugbank dataset described in subsection 2.1. We calculated the interaction probabilities between 16 GABAR proteins and 6708 drugs and ranked them by their probabilities. The number of drugs divided in the training and test sets is described in Table 9. The last column of Table 9 shows the number of drugs predicted in the top 10 list. The results indicate that our model provides insights into potential DTIs and has potential applications in drug repositioning. More details about the top 10 candidate results of GABAR proteins are listed in Supplementary Table S9.

Furthermore, to explore the structural similarity between the top-ranked candidates and the approved drugs, we visualized the ECFP of compounds. We conducted t-distributed stochastic neighbor embedding (t-SNE) for dimension reduction and visualized the approved drugs, the top-ranked candidates and compounds with low interaction probabilities. Obviously, as shown in Supplementary Figure S1, the top-ranked candidates are more similar to the approved drugs than the compounds with low interaction probabilities.

3.7 Model interpretation

To demonstrate that the attention mechanism not only enhances the model's performance but also leads to more interpretability, we conduct two case studies, i.e. Crystal structure of HIV protease D545701 bound with GW0385 (PDB: 2FDD) and Crystal structure of type 2 PDF from *Streptococcus agalactiae* in complex with inhibitor AT018 (PDB: 5JF3). We first feed drug SMILES and amino acid sequence into our model, then get the protein attention matrix $A_p \in \mathbb{R}^{M \times f}$. We apply the mean operator to A_p to get the protein attention vector $a_p \in \mathbb{R}^M$, which reflects the distribution of attention values on amino acid sequences. Then, we map the attention vector *ap* to the 3D structure of the complex to visualize which regions in a protein have more effective roles for the interaction.

The attention weights of 2FDD and 5JF3 are shown in Figure 2. The amino acids in proteins, which get high attention weights, are highlighted in red in 3D structure visualization. As shown in Figure 2a, two of the twelve binding sites, ALA 28 and PRO 81, get high attention scores, especially PRO 81, which get the highest score. As for 5HF3, there are ten binding sites. As shown in Figure 2b, LEU 132 gets the fourth score and VAL 71 gets the highest score. These results indicate that our model can help researchers narrow the search space for binding sites. Meanwhile, we also notice that many non-binding sites are highlighted.

Table 8. Comparisons of our model with and without attention block

Methods	Accuracy (Std)	Precision (Std)	Recall (Std)	AUC (Std)	AUPR (Std)
No_Attention_DTI	0.770 (0.001)	0.783 (0.001)	0.705 (0.002)	0.826 (0.001)	0.832 (0.001)
Attention_DTI	0.775 (0.003)	0.784 (0.002)	0.722 (0.001)	0.830 (0.002)	0.841 (0.002)
Multi-Head_DTI	0.784 (0.001)	0.788 (0.001)	0.753 (0.001)	0.846 (0.001)	0.852 (0.001)
HyperAttentionDTI	**0.810** (0.002)	**0.799** (0.001)	**0.829** (0.001)	**0.889** (0.001)	**0.897** (0.001)

Table 9. Summary of the drug numbers in the case study

Name	UniProt ID	Train	Test	Top 10
GABAR subunit alpha-1	P14867	84	15	4
GABAR subunit alpha-2	P47869	76	16	5
GABAR subunit alpha-3	P34903	77	16	6
GABAR subunit alpha-4	P48169	66	21	5
GABAR subunit alpha-5	P31644	73	17	5
GABAR subunit alpha-6	Q16445	66	21	6
GABAR subunit beta-1	P18505	56	20	4
GABAR subunit beta-2	P47870	55	21	5
GABAR subunit beta-3	P28472	58	20	5
GABAR subunit delta	O14764	54	20	6
GABAR subunit epsilon	P78334	54	20	4
GABAR subunit gamma-1	Q8N1C3	58	19	5
GABAR subunit gamma-2	P18507	63	17	5
GABAR subunit gamma-3	Q99928	57	20	3
GABAR subunit pi	O00591	53	21	4
GABAR subunit theta	Q9UN88	49	24	6

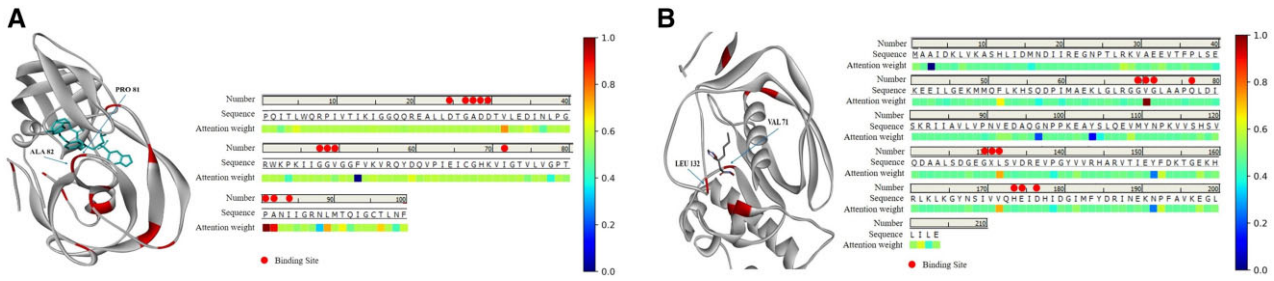

Fig. 2. Attention weights of protein sequences. (**a**) Attention weight of Protease (PDB: 2FDD). (**b**) Attention weight of Peptide deformylase (PDB: 5JF3)

4 Conclusion

In this work, we introduce an end-to-end bio-inspired deep learning-based model, HyperAttentionDTI, for DTI prediction. We design a novel attention mechanism between drugs and proteins to model the complex interaction between atoms and amino acids. To verify the effectiveness of our model, we compare our model with the state-of-the-art baselines on three benchmark datasets under four different settings. The results show that our model achieves significantly improved performance on AUC and AUPR under all settings. Moreover, the case study of the human Gamma-aminobutyric acid receptors demonstrate the ability of our model to improve the VS of drug discovery. Finally, we map attention weights to protein sequences, which could help us narrow the search space for binding sites and further explore how a drug binds to its target protein in the future.

Although HyperAttentionDTI has demonstrated effective performances in predicting DTIs, there is still room for improvement. First, the features of SMILES strings extracted by 1D-CNNs are abstract and difficult to analyze. Therefore, we will use the graph structure of drugs, which is more natural than SMILES string, and design molecular graph neural networks to achieve advanced performance in the future. Second, our model cannot precisely locate the binding sites. Introducing the label information of binding sites and leveraging multi-task learning have the potential to further push the limit of HyperAttentionDTI.

Funding

This work was supported by NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization [U1909208], the National Natural Science Foundation of China [61772552, 62072473], 111 Project [B18059] and the Hunan Provinvial Science and Technology Program [2018WK4001, 2020GK2019].

Conflict of Interest: none declared.

References

Abbasi,K. *et al.* (2020) Deepcda: deep cross-domain compound–protein affinity prediction through lstm and convolutional neural networks. *Bioinformatics*, **36**, 4633–4642.

Abdel-Basset,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.

Agamah,F.E. *et al.* (2020) Computational/in silico methods in drug target and lead prediction. *Brief. Bioinf.*, **21**, 1663–1675.

Bahdanau,D. *et al.* (2015) Neural machine translation by jointly learning to align and translate. In: 3rd *International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA*.

Chen,L. *et al.* (2020) TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, **36**, 4406–4414.

Chen,W. *et al.* (2021) Predicting drug–target interactions with deep-embedding learning of graphs and sequences. *J. Phys. Chem. A*, **125**, 5633–5642.

Davis,M.I. *et al.* (2011) Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.*, **29**, 1046–1051.

Ezzat,A. *et al.* (2019) Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Brief. Bioinf.*, **20**, 1337–1357.

Gao,K.Y. *et al.* (2018) Interpretable drug target prediction using deep neural representation. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, Stockholm, Sweden, pp. 3371–3377.

Hazra,A. *et al.* (2021) Recent advances in deep learning techniques and its applications: an overview. In: Rizvanov, A.A. *et al* (eds) *Advances in Biomedical Engineering and Technology.Springer Singapore, Singapore, pp. 103-122.*

He,K. *et al.* (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: *2015 IEEE International*

Conference on Computer Vision, ICCV 2015, Santiago, Chile, pp. 1026–1034.
Himmat,M. *et al.* (2016) Adapting document similarity measures for ligand-based virtual screening. *Molecules*, **21**, 476.
Hu,J. *et al.* (2020) Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, **42**, 2011–2023.
Huang,K. *et al.* (2021) MolTrans: molecular Interaction Transformer for drug–target interaction prediction. *Bioinformatics*, **37**, 830–836.
Kim,Y. (2014) Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, Doha, Qatar*, pp. 1746–1751.
Landrum,G. (2006) RDKit: Open-Source Cheminformatics Software. *http//www.rdkit.org/ (22 October 2021, date last accessed)*.
LeCun,Y. *et al.* (2015) Deep learning. *Nature*, **521**, 436–444.
Lee,I. *et al.* (2019) DeepConv-DTI: prediction of drug–target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol.*, **15**, e1007129.
Lim,J. *et al.* (2019) Predicting drug–target interaction using a novel graph neural network with 3d structure-embedded graph representation. *J. Chem. Inf. Model.*, **59**, 3981–3988.
Loshchilov,I. and Hutter,F. (2019) Decoupled weight decay regularization. In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA*.
Maia,E.H.B. *et al.* (2020) Structure-based virtual screening: from classical to artificial intelligence. *Front. Chem.*, **8**, 343.
Nguyen,T. *et al.* (2020) GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics 2020, 37, 1140-1147*.
Öztürk,H. *et al.* (2018) DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, **34**, i821–i829.
Paszke,A. *et al.* (2019) *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, volume 32 of *Advances in Neural Information Processing Systems, Vancouver, BC, Canada. Curran Associates, Inc.*
Quan,Z. *et al.* (2019) GraphCPI: graph neural representation learning for compound–protein interaction. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA*. IEEE, pp. 717–722.
Rifaioglu,A.S. *et al.* (2019) Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings in Bioinformatics*, **20**, 1878–1912.
Rogers,D. and Hahn,M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, **50**, 742–754.
Shin,B. *et al.* (2019) Self-attention based molecule representation for predicting drug–target interaction. In: *Proceedings of the 4th Machine Learning for Healthcare Conference*, Ann Arbor, Michigan, USA, Vol. 106, pp. 230–248.
Sigel,E. and Steinmann,M.E. (2012) Structure, function, and modulation of gabaa receptors. *J. Biol. Chem.*, **287**, 40224–40231.
Tang,J. *et al.* (2014) Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J. Chem. Inf. Model.*, **54**, 735–743.
Tian,K. *et al.* (2016) Boosting compound–protein interaction prediction by deep learning. *Methods*, **110**, 64–72.
Torng,W. and Altman,R.B. (2019) Graph convolutional neural networks for predicting drug-target interactions. *J. Chem. Inf. Model.*, **59**, 4131–4149.
Tsubaki,M. *et al.* (2019) Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, **35**, 309–318.
Vaswani,A. *et al.* (2017) Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA*, pp. 6000–6010.
Wang,J. *et al.* (2020) GNN-PT: enhanced prediction of compound-protein interactions by integrating protein transformer. *arXiv, preprint arXiv: 2009.00805 2020*.
Wen,M. *et al.* (2017) Deep-learning-based drug–target interaction prediction. *J. Proteome Res.*, **16**, 1401–1409.
Wishart,D. *et al.* (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acid Res.*, **34**, D668–D672.
Zhao,Q. *et al.* (2019) AttentionDTA: prediction of drug–target binding affinity using attention model. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA*, pp. 64–69.
Zhao,Q. *et al.* (2021) Biomedical data and deep learning computational models for predicting compound-protein relations. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, doi: 10.1109/TCBB.2021.3069040.
Zheng,S. *et al.* (2020) Predicting drug–protein interaction using quasi-visual question answering system. *Nat. Mach. Intell.*, **2**, 551–551.
Zhu,S. *et al.* (2018) Structure of a human synaptic gabaa receptor. *Nature*, **559**, 67–72.