Proceedings of Machine Learning Research vol 278:1–13, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

HBADTI: Drug-target interaction prediction based on multi head attention and bidirectional cross attention

Jiaming Zhao 1048669736@QQ.COM
University of South China

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

The study of drug-target interactions (DTIs) holds critical importance in the drug development process. The core challenge in DTI prediction lies in accurately capturing the features of both drugs and proteins, as well as thoroughly understanding their interaction mechanisms. In light of this, we developed an end-to-end DTI prediction model called HBADTI. The model employs graph convolutional networks to encode drug features. For protein feature extraction, we designed a dedicated feature extraction module (ESAM) that combines convolutional neural networks (CNNs) with multi-head self-attention mechanisms to effectively capture protein sequence characteristics. Subsequently, a bidirectional cross-attention network is utilized to integrate the features of both drugs and proteins, followed by a multilayer perceptron to classify unknown drug-target pairs.Comparative experimental results demonstrate that HBADTI outperforms multiple baseline methods. Ablation studies further confirm that both the bidirectional attention network and the ESAM module significantly contribute to the improvement of DTI prediction performance.

Keywords: Drug target interactions, cross attention, self-attention mechanisms, drug discovery, deep learning

1. Introduction

In drug discovery, the effective identification of drug-target interactions (DTIs) (Gao et al., 2024) plays a crucial role in innovative drug development. While traditional in vitro experiments can reliably validate DTIs, they are time consuming, costly, and low-throughput (Li et al., 2024), making them inadequate for modern drug discovery needs. Computational approaches have gained significant attention due to their ability to efficiently identify potential DTIs, thereby reducing the cost and time required for experimental screening. Traditional computational methods primarily include molecular docking (Zhang et al., 2024a) and virtual screening (Jung et al., 2022). Molecular docking predicts interactions by simulating the three-dimensional binding patterns between small molecules and target proteins, but its effectiveness is limited by the scarcity of protein 3D structural data (Supriya et al., 2016). Virtual screening predicts novel active molecules based on the characteristics of known active compounds (He et al., 2020), yet its performance heavily depends on the quantity and quality of available active molecules - the predictive accuracy declines significantly when data is insufficient (Zhang et al., 2024b).

Recent years have witnessed remarkable advances in machine learning, particularly deep learning, within the field of bioinformatics. For drug-target interaction (DTI) prediction, chemical compounds are typically encoded as SMILES strings (Krenn et al., 2023), while proteins are represented as one-dimensional amino acid sequences. Current deep learning approaches generally follow a two-stage pipeline: feature extraction for both drugs and proteins (Huang et al., 2023), followed by

interaction modeling between the extracted features.In the feature extraction stage, graph convolutional networks (GCNs) (Chen et al., 2020), graph attention networks (GATs) (Zeng et al., 2023), and graph neural networks (GNNs) (Öztürk et al., 2018) have been widely adopted to encode two-dimensional molecular graphs of drugs. For protein sequence feature extraction, convolutional neural networks (CNNs) (Tsubaki et al., 2019), as well as hybrid architectures combining CNNs with long short-term memory (LSTM) networks or Transformer decoders, have demonstrated promising results. The interaction modeling paradigm has evolved significantly. Early approaches relied on simple concatenation or fusion of drug and protein feature vectors as input to prediction models. Recent research has shifted focus toward more sophisticated interaction-aware modeling techniques. Notable examples include: the TransformerCPI (Zeng et al., 2023) model which employs a modified Transformer architecture for sequence-based compound-protein interaction classification; DrugBAN (Nguyen et al., 2021) that incorporates a bilinear attention network to capture local structural relationships while utilizing conditional domain adversarial learning to align interaction representations across different distributions for better generalization to novel drug-target pairs; CATDTI (Bai et al., 2023), a cross-attention and Transformer-based model that combines CNNs with Transformers to encode distance relationships between amino acids in protein sequences while employing a cross-attention module to capture interaction features; and BINDTI (Brauwers and Frasincar, 2023), a bidirectional intent network prediction framework that integrates graph convolutional networks (GCNs), ACmix hybrid models, and attention mechanisms to achieve more precise representation of drug and protein characteristics.

We propose HBADTI, an innovative end-to-end framework for identifying drug-target interaction candidates, with two primary methodological contributions. First, we developed a novel feature extraction module called ESAM that creatively combines convolutional neural networks (CNNs) with multi-head self-attention mechanisms. This integrated approach synergizes local feature extraction with global dependency modeling to achieve efficient characterization of amino acid sequence features. Specifically, the ESAM module employs three 1D CNN layers to capture local sequential patterns, while the convolutional outputs are further processed through self-attention mechanisms to establish long-range dependencies within the sequence, thereby comprehensively enhancing the quality of protein feature representation.

Second, we designed an innovative bidirectional cross-attention network architecture. This network implements deep fusion of drug and target features through multi-head attention mechanisms and cross-attention mechanisms. The technical implementation involves two key stages: initial enhancement of drug and target feature representations through a multi-head attention network, followed by bidirectional feature interaction via a cross-attention network. This dual-stage architecture not only enables precise capture of specific interaction patterns between drugs and targets but also significantly improves the model's capacity for bidirectional feature representation. The proposed framework demonstrates superior performance in capturing the complex interplay between molecular structures and protein targets, offering a more robust solution for drug-target interaction prediction compared to conventional approaches.

2. Materials and Methods

2.1. Data Set

The performance of the HBADTI model was evaluated on three widely used benchmark datasets: BindingDB, BioSNAP, and Human. Each dataset provides distinct characteristics for comprehen-

sive model assessment. The BindingDB dataset contains a large collection of experimentally validated protein-ligand binding data, serving as a reliable source of positive interaction samples. The BioSNAP dataset offers balanced biological network data covering multiple domains including gene regulation, metabolic pathways, and protein-protein interactions. The Human dataset integrates large-scale biomedical data with carefully curated negative DTI samples, providing crucial negative controls for model training and evaluation.In addition, in this study, we used the data grouping strategy to simulate the whole process of drug research and development, covering new drug discovery, new target identification and drug target interaction exploration, in order to comprehensively evaluate the performance of the model. We processed the datasets using three distinct grouping approaches: Including E_1: Warm start; E_2: Drug cold start; E_3: Target cold start.

2.2. HBADTI

In this study, we propose HBADTI, an end-to-end framework for drug-target interaction (DTI) prediction. The framework consists of four key computational components: (1) Molecular graph representation of drugs using SMILES strings followed by feature extraction through graph convolutional networks (GCNs); (2) Protein sequence encoding and feature extraction using our novel ESAM; (3) Feature fusion via a bidirectional cross-attention network that generates comprehensive drug-target pair representations; and (4) Final interaction prediction through multilayer perceptrons based on the fused features.The comprehensive operational flow of HBADTI is vividly illustrated in Figure 1.

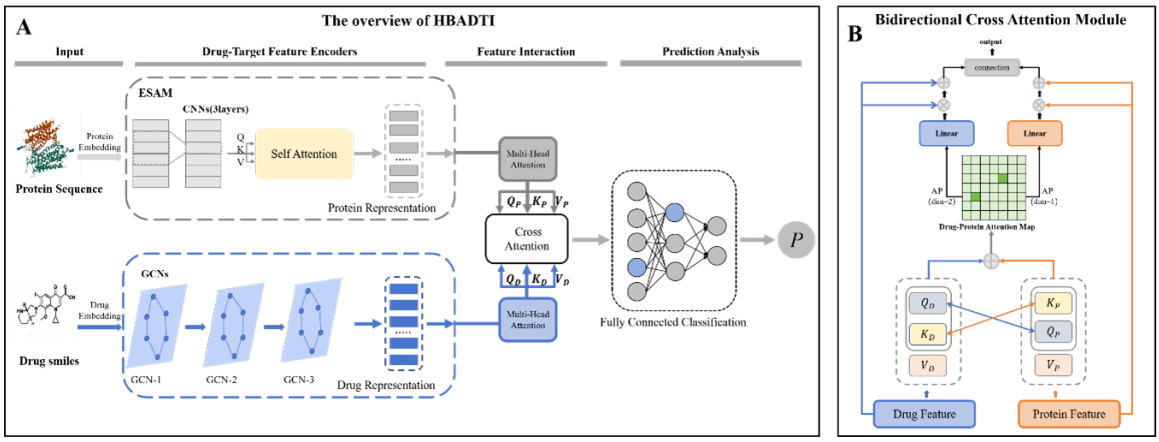

Figure 1: HBADTI framework diagram of the model.

(1) The key task in drug-target interaction (DTI) prediction is to construct a model M that can map the fused features of protein sequence P and drug molecular graph G to an interaction probability $p \in [0, 1]$. Specifically, proteins are represented by their amino acid sequences with a maximum length of 1,000 residues. Sequences exceeding this length are truncated, while shorter sequences are zero-padded to achieve uniform length. This standardized representation not only facilitates computational processing but also effectively preserves the critical features of proteins. For drug molecules, SMILES strings are used for initial representation and are subsequently converted into two-dimensional molecular graphs G. This graph-based representation overcomes the limitations of

traditional one-dimensional SMILES strings by more comprehensively capturing the spatial structure and topological characteristics of molecules.

(2) The GCN-based encoding of drug features leverages graph convolutional networks to effectively capture the inherent graph-structured information of drug molecules. Initially, drug-related SMILES strings are converted into two-dimensional molecular graphs $G = (V, E)$, where V denotes the set of atomic nodes and E represents the set of chemical bond edges. Each node's features are initialized based on its chemical properties, generating a 74-dimensional integer vector using the DGL package. For drug molecules with insufficient numbers of atoms, zero-padding is applied to node features to ensure uniform dimensional representation across all drugs. These node features undergo linear transformation to produce a dense matrix, which serves as input to the GCN.During the GCN encoding process, the molecular graph passes through three GCN layers. At each layer, convolutional operations aggregate neighboring information for each atom through the adjacency matrix, thereby updating node features. The drug encoder can be formally represented as:

$$\mathbf{H}_{\mathrm{G}}^{(1+1)} = \mathrm{ReLU}\left(\widehat{\mathbf{D}}^{-\frac{1}{2}}\widehat{\mathbf{A}}\widehat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}_{\mathrm{G}}^{1}\mathbf{W}^{1}\right) \tag{1}$$

In the formula, $\hat{A} = A + I$ denotes the normalized adjacency matrix with self-connections, where A represents the original adjacency matrix and I is the identity matrix H_G^1 represents the node feature representation at layer L, while denotes the activation function.

(3) Protein feature extraction in our model is achieved through a hybrid architecture that combines Convolutional Neural Networks (CNNs) with multi-head self-attention mechanisms. CNNs are renowned for their robust feature-learning capabilities. However, they face limitations when it comes to modeling long-range dependencies within amino acid sequences. To overcome this shortcoming, we incorporate a multi-head self-attention mechanism. This mechanism is designed to effectively capture global interdependencies across sequences. We propose an Enhanced ESCM approach to enhance the performance of feature extraction. ESCM enables parallel computation across multiple attention heads. This allows the model to concurrently focus on correlations between different positions within the sequence, Thus, there is a significant enhancement in the performance of feature extraction. Specifically, we rep-resent the protein encoder as $\mathrm{E_P}$, Building on the concept of embedding amino-acid sequence words, we initialize amino acids into a learnable embedding matrix $\mathrm{E_P} \in \mathrm{R}^{23*\mathrm{D_P}}$, Here, 23 represents the number of amino-acid types, and $\mathrm{D_P}$ is the spatial dimension. The protein feature matrix is then obtained through $\mathrm{X_p}$. In the convolutional layer, we utilize a 3x1 convolution kernel to extract amino-acid sequence features. Additionally, we expand the kernel size to capture more local features. The formula for the convolution operation is as follows:

$$\mathbf{H}_{\mathrm{P}}^{(1+1)} = \mathrm{ReLU}\left(\mathbf{CNN}\left(\mathbf{W}_{\mathrm{c}}^{1}, \mathbf{b}_{\mathrm{c}}^{1}, \mathbf{H}_{\mathrm{P}}^{1}\right)\right) \tag{2}$$

Among these elements, $\mathrm{H_P^1}$ serves as the input for the L layer, After that, it is run through the convolution operation. $\mathrm{W_c^1}$ represents the convolution kernel (weight matrix) of the L layer, and $\mathrm{b_c^1}$ is the bias term of the L layer. The outcome from the third convolutional layer is directed to a self-attention unit. This unit is designed to capture long-range dependencies. The specific calculation procedure is presented as follows:

$$\mathbf{P}_{\mathrm{i,j}} = \prod_{\mathrm{l}=1}^{\mathrm{M}} \left(\sum_{\mathrm{x,y} \in \mathrm{N_m(i)}} \mathbf{B}\left(\mathbf{W}_{\mathrm{q}}^{(1)}\mathbf{f}_{\mathrm{i,j}}, \mathbf{W}_{\mathrm{k}}^{(1)}\mathbf{f}_{\mathrm{x,y}}\right)\mathbf{W}_{\mathrm{v}}^{(1)}\mathbf{f}_{\mathrm{x,y}} \right) \tag{3}$$

Let the rows and columns of the protein representation be indexed by i and j respectively and $\mathbf{w}_{\mathrm{q}}^{(1)}$, $\mathbf{w}_{\mathrm{k}}^{(1)}$ are weight matrices applied to the feature $\mathbf{f}_{\mathrm{i,j}}$, $\mathbf{f}_{\mathrm{x,y}}$, $\mathbf{W}_{\mathrm{q}}^{(1)}$ que-ries at positions i, j and domain positions x, y, correspondingly. The projection matrices for keys $\mathbf{w}_{\mathrm{k}}^{(1)}$ and values $\mathbf{w}_{\mathrm{v}}^{(1)}$ are employed to compute the query vector, key vector, and value vector of the input features respectively.

Attention weight calculation:

$$\mathrm{C}\left(\mathbf{W}_{\mathrm{q}}^{(1)}\mathbf{f}_{\mathrm{i,j}}, \mathbf{W}_{\mathrm{k}}^{(1)}\mathbf{f}_{\mathrm{n,\,m}}\right) = \mathrm{softmax}\left(\frac{\left(\mathbf{w}_{\mathrm{q}}^{(1)}\mathbf{f}_{\mathrm{i,j}}\right)^{\mathrm{T}}\left(\mathbf{w}_{\mathrm{k}}^{(1)}\mathbf{f}_{\mathrm{n,\,m}}\right)}{\sqrt{\mathbf{z}}}\right) \tag{4}$$

Among them, z is the characteristic dimension of $\mathbf{W}_{\mathrm{q}}^{(1)}\mathbf{f}_{\mathrm{i,j}}$.

(4) To effectively integrate drug and target features, we developed a bidirectional cross-attention network that combines multi-head attention (Vaswani et al., 2017) and cross-attention mechanisms (Chen et al., 2022). The multi-head attention module enhances feature representation for both drug molecules and target proteins through parallelized feature subspace learning and gradient stabilization. The drug molecular graph features are augmented as follows:

$$\mathbf{H}_{\mathrm{D}}^{\mathrm{e}} = \mathrm{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \mathrm{H}_{\mathrm{G}} \tag{5}$$

Q, K, V are generated via projection based on H_G. Analogously, The following equation is used to ex-press the features of the target protein sequence:

$$\mathbf{H}_{\mathbf{P}}^{e} = \mathrm{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \mathrm{H}_{\mathbf{P}} \tag{6}$$

The bidirectional network is designed to capture complex drug-protein interactions and accurately predict binding affinity. It employs bidirectional information flow and multi-head cross-attention mechanisms to globally identify critical binding regions and deepen understanding of drug-target interactions. Specifically, the architecture first utilizes two independent encoders to extract drug and protein features separately. These features are then enhanced through a multi-head attention layer before being processed by a cross-attention layer, enabling multi-subspace feature inter-action that captures hierarchical interaction patterns. The bidirectional information flow creates a dynamic fusion mechanism where drug and protein features mutually influence and optimize each other's representations.

$$\mathbf{Q_j} = \mathbf{Q}\mathbf{W}_{\mathrm{j}}^{\mathrm{Q}}, \mathbf{K}_{\mathrm{j}} = \mathbf{K}\mathrm{W}_{\mathrm{j}}^{\mathrm{K}}, (\mathrm{j} = 1, \ldots, \text{ heads }) \tag{7}$$

Among these, the drug features and target features serve as the initial values for the Q and K, and the values of Q and K for each attention head are generated through a linear transformation using the corresponding weight matrices. W_j^Q and W_j^K. At this point, "heads" is used to signify the number of attention heads.

$$A_j = \mathrm{Softmax}\left(\frac{Q_j K_j^T}{\sqrt{D_k}}\right), \quad (i = 1, \ldots, \text{ heads }) \tag{8}$$

Among them, A_j, Q_j and K_j respectively represent the similarity matrix between the substructures of each herb and the proton structure of the protein. The probability distribution is obtained by scaling the dot product ($\sqrt{D_k}$ to prevent gradient explosion) and Softmax normalization.

$A_j \in R^{n\times n}$ represents the attention weight matrix of the jth head, where each element reflects the strength of the combination of two substructures.

Once the bidirectional cross attention is completed, a matrix probability distribution map is produced, $a = A_{uk} + A_{Ku}$ ($a \in R^{n\times n}$); A_{uk} is an attention map with drug substructures as queries and protein structures as bonds, while A_{Ku} is an attention map with protein structures as queries and drug structures as bonds. Subsequently, a Sigmoid activation function is used to generate weights F_d and F_p, for the comprehensive matrix a, which are used to dynamically adjust the fusion ratio of features.The adjusted features G_p and G_d of the protein and the drug are described as follows:

$$G_p = (1 + F_p) \odot H_p^e \tag{9}$$

$$G_d = (1 + F_d) \odot H_d^e \tag{10}$$

Finally, feature dimensionality reduction is performed to extract global characteristics that aggregate binding information from all substructures, forming a comprehensive representation of the drug-target complex.

(5) The prediction results and output are obtained by formulating the DTI prediction task as a binary classification problem. To compute the final DTI probabilities, we employ ReLU activation functions and fully connected layers, while training the model using cross-entropy loss.

$$\text{Loss} = -\left[y \log\left(y'\right) + (1 - y) \log\left(1 - y'\right)\right] \tag{11}$$

Among these, y denotes the true label, which can take on a value of either 0 or 1. A value of 0 indicates the absence of an interaction between the drug-target pair, with a value of 1 signifying the presence of an interaction. y' represents the interaction probability predicted by the model, and its value ranges within the interval [0,1].

3. Experimental section

3.1. Experimental Setup

During the DTI (Drug-Target Interaction) prediction experiments, our model employed AUROC (area under the receiver operating characteristic curve (Wang et al., 2023)) and AUPRC (area under the precision-recall curve (Lu et al., 2025)). and Accuracy as the primary evaluation indicators. Additionally, Recall, Specificity, and Precision were incorporated as supple-mentary indicators to conduct a comprehensive assessment of the model's performance.

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}; \ \text{Preci-sion} = \frac{\text{TP}}{\text{TP} + \text{FP}}; \\ \text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}; \ \text{Re-call} = \frac{\text{TP}}{\text{TP} + \text{FN}} \end{aligned} \tag{12}$$

In the above equation, TP represents the interacting pairs correctly predicted by the classifier, while TN denotes the non-interacting pairs correctly identified. Similarly, FP indicates cases where the classifier mistakenly predicted non-interacting pairs as interacting ones, and FN refers to instances where truly interacting pairs were incorrectly classified as non-interacting.

Details of implementation: We implemented the HBADTI model with the Python programming language.In terms of model architecture configuration, The key hyperparameters we established are as follows: CNN_layers=3, Self-Attention-head=3, Attention_heads=8, hidden_size=512,

epoch_size=100, batch_size=32, learning_rate=5, Learning_decay=0.5. To conduct a more comprehensive evaluation of the model's generalization ability and accu-racywe carried out comparative experiments between the HBADTI model and existing baseline methods. Regarding the data partitioning strategy, each dataset was divided into a training set, a validation set, and a test set at a ratio of 7:1:2. Furthermore, statistical experiments were conducted on three datasets to evaluate the significant performance enhancement of our model compared to each baseline model.The configuration of our experimental environment is as follows: The Python version used is 3.8. The deep-learning framework is PyTorch 2.0.0. The parallel computing platform is CUDA 11.8, and an NVIDIA GeForce RTX 3090 GPU is employed for model training. This setup ensures high computational efficiency and substantial processing power.

3.2. Baseline method

To conduct a comprehensive evaluation of our model's performance, we carried out a detailed comparative analysis. In this process, we benchmarked the proposed model against several of the most recent deep-learning methods.

TransformerCPI (Zeng et al., 2023): Its core design combines the advantages of pretrained protein embeddings and the Trans-former architecture, effectively capturing the interactions between drugs and targets.

GraphDTA (Huang et al., 2021): This model innovatively combines Graph Neural Networks (GNNs) and one-dimensional CNNs. It extracts drug molecular graph features by exploring various GNN variants such as GCN, GAT, GIN, and the GCN-GAT combined model. Meanwhile, it utilizes one-dimensional CNNs to process protein sequence features. Finally, feature fusion and prediction are achieved through fully connected layers.

MolTrans (Peng et al., 2024): The model is based on the Transformer architecture. It first decomposes drugs and proteins into sets of sequences with explicit substructures using the frequent continuous subsequence method. Then, it employs a Transformer embedding module to generate enhanced contextual representations. Finally, it models the drug-target interaction graph through dot product operations and combines CNN with a fully connected network for prediction.

DrugBAN (Nguyen et al., 2021): The model employs a bilinear attention network to encode interaction features between drugs and proteins. Through an innovative feature fusion mechanism, it effectively integrates information from both modalities, significantly enhancing the model's capability to capture deep-level drug-target relationships.

BINDTI (Brauwers and Frasincar, 2023): The model introduces a bidirectional intent module to model drug-target interactions. For feature ex-traction, it employs a hybrid architecture combining GCN and Acmix.

4. Experimental Result

4.1. Comparative experiment

Our model demonstrated outstanding performance across all three benchmark datasets, consistently achieving state-of-the-art results on multiple evaluation metrics while attaining the highest recall and precision rates on each dataset. Specifically, HBADTI showed significant improvements over the best baseline models: on the BindingDB dataset, it achieved 0.8% higher AUROC and 1% higher AUPRC; on BioSNAP, the improvements reached 1.6% (AUROC) and 2.2% (AUPRC); and

on the Human dataset, performance gains were 0.3% (AUROC) and 0.2% (AUPRC). The results in Tables 1, 2, and 3 clearly demonstrate that HBADTI provides substantial performance enhancement for the drug-target interaction prediction task.

Table 1: Performance Comparison of Models on the BindingDB DataSet with E_1 Setting (5 Random Runs)

Datasets	Model	AUROC	AUPRC	Accuracy	Recall	Precision	Specificity
BindingDB	TransformerCPI	0.9293	0.9252	0.8791	0.8743	0.8605	0.8875
	GraphDTA	0.9475	0.9377	0.8923	0.8901	0.8763	0.8904
	MolTrans	0.9504	0.9397	0.8996	0.9012	0.8795	0.8813
	DrugBAN	0.9551	0.9452	0.9119	0.8981	0.8976	0.9103
	BINDTI	0.9564	0.9433	0.9022	0.9023	0.8612	0.8951
	HBADTI	**0.9641**	**0.9534**	**0.9094**	**0.9072**	**0.8923**	**0.9112**

Table 2: Performance comparison of the model on the Human dataset under E_1 setting (randomly run 5 experiments)

Datasets	Model	AUROC	AUPRC	Accuracy	Recall	Precision	Specificity
Human	TransformerCPI	0.9734	0.9753	0.9244	0.9365	0.9215	0.9241
	GraphDTA	0.9806	0.9776	0.9392	0.9358	0.9223	0.9413
	MolTrans	0.9829	0.9832	0.9433	0.9344	0.9517	0.9462
	DrugBAN	0.9825	0.9801	0.9371	0.9382	0.9354	0.9386
	BINDTI	0.9841	0.9838	0.9428	0.9463	0.9401	0.9381
	HBADTI	**0.9874**	**0.9851**	**0.9485**	**0.9473**	**0.9496**	**0.9472**

Table 3: Performance comparison of the model on the BioSNAP dataset under E_1 setting (randomly run 5 experiments)

Datasets	Model	AUROC	AUPRC	Accuracy	Recall	Precision	Specificity
BioSNAP	TransformerCPI	0.8705	0.8805	0.7989	0.8013	0.7808	0.7907
	GraphDTA	0.8883	0.8913	0.8274	0.8386	0.8275	0.8236
	MolTrans	0.8951	0.8993	0.8263	0.8517	0.8113	0.8059
	DrugBAN	0.9073	0.9114	0.8322	0.8415	0.8351	0.8223
	BINDTI	0.8971	0.8966	0.8313	0.8294	0.8303	0.8312
	HBADTI	**0.9136**	**0.9183**	**0.8402**	**0.8351**	**0.8336**	**0.8443**

Note: The best results are presented in bold.

In addition to performing comparative experiments under the hot-start setting of E_1, we also carried out com-parative investigations under the settings of E_2 and E_3. The objective was to more comprehensively assess the model's capacity for searching for novel drugs and targets. The results are presented in Tables 4 and 5.

Table 4: Performance Comparison of Binding DB under E_2 and E_3 (Randomly Run 5 Experiments)

Datasets	Setting	Model	AUROC	AUPRC	Accuracy
BindingDB	E_2	TransformerCPI	0.9021	0.8954	0.8492
		GraphDTA	0.9214	0.9017	0.8624
		MolTrans	0.9293	0.9093	0.8696
		DrugBAN	0.9356	0.9118	0.8713
		BINDTI	0.9254	0.9132	**0.8728**
		HBADTI	**0.9421**	**0.9176**	0.8722
BindingDB	E_3	TransformerCPI	0.5305	0.4295	0.4772
		GraphDTA	0.5494	0.4372	0.4913
		MolTrans	0.5523	0.4328	0.4955
		DrugBAN	0.5562	0.4420	**0.5171**
		BINDTI	0.5584	0.4461	0.5043
		HBADTI	**0.5647**	**0.4530**	0.5075

Table 5: Performance Comparison of BioSNAP under E_2 and E_3 (Randomly Running 5 Experiments)

Datasets	Setting	Model	AUROC	AUPRC	Accuracy
BioSNAP	E_2	TransformerCPI	0.7732	0.7835	0.6941
		GraphDTA	0.7855	0.7903	0.7227
		MolTrans	0.7911	0.7950	0.7264
		DrugBAN	0.8024	0.8139	0.7353
		BINDTI	0.7967	0.7916	0.7361
		HBADTI	**0.8108**	**0.8134**	**0.7412**
BioSNAP	E_3	TransformerCPI	0.5731	0.5857	0.4975
		GraphDTA	0.5853	0.5907	0.5268
		MolTrans	0.5982	0.5983	0.5274
		DrugBAN	0.6019	0.6185	**0.5394**
		BINDTI	0.5954	0.5941	0.5328
		HBADTI	**0.6109**	**0.6175**	0.5316

Experimental results from multiple datasets demonstrate that the HBADTI model achieves superior performance compared to other baseline models in drug-target interaction prediction tasks. Specifically, on both BioSNAP and BindingDB benchmark datasets, HBADTI exhibits significant advantages in two key evaluation metrics: AUROC and AUPRC. Further analysis reveals that HBADTI consistently attains optimal prediction performance across three different experimental settings (E_1, E_2 and E_3) in the BindingDB dataset, which robustly validates the model's stability and generalization capability. However, under E_2 and E_3 conditions in the BioSNAP dataset, HBADTI shows marginally lower AUPRC values than the MolTrans model, a phenomenon that may be attributed to certain intrinsic characteristics of the dataset.

Table 6: Statistical testing of Binding DB, Human, and BioSNAP datasets in five random tests

Model	BindingDB			Human			BioSNAP		
	D-value	t	P	D-value	t	p	D-value	t	p
TransformerCPI	0.0482	25.56	<0.001	0.0173	5.93	<0.005	0.0337	16.80	<0.001
GraphDTA	0.0323	15.21	<0.001	0.0123	9.59	<0.001	0.0313	11.71	<0.001
MolTrans	0.0331	6.31	<0.005	0.0131	10.31	<0.001	0.0328	12.93	<0.001
DrugBAN	0.0218	5.97	<0.005	0.0118	5.90	<0.005	0.0295	5.60	<0.005
BINDTI	0.0143	12.23	<0.001	0.0034	5.80	<0.005	0.0118	7.79	<0.001

Moreover, we carried out statistical tests to substantiate the model's performance. Table 6 showcases the t- test results for the AUROC values derived from five random data splits. The D-value signifies the disparity in AUROC be-tween the HBADTI model and the baseline models. As a general rule, higher t-values are associated with lower p-values, and a p-value less than 0.05 is commonly regarded as statistically significant. The t-test results clearly indicate that the HBADTI model significantly outperforms the baseline models across the BindingDB, BioSNAP, and Human datasets.

4.2. Ablation experiment

We performed comprehensive ablation studies on the BioSNAP and Human datasets to systematically evaluate the contributions of different model components. Our experimental design employed a stepwise removal approach to isolate and quantify the impact of each module. First, we examined feature representation methods by individually removing each component and comparing the performance against the complete model. This allowed precise assessment of how each feature extraction technique influenced the final predictions. Next, we established a baseline model without any feature interaction mechanisms to verify the importance of attention-based fusion.

To further analyze the attention mechanisms, we created two specific model variants: one removing the multi-head self-attention while retaining only CNN-based protein feature extraction, and another replacing the bidirectional cross-attention with unidirectional attention. The experimental results (Table 7) demonstrate that the full model integrating both the ESAM module and bidirectional cross-attention network achieves superior performance across all evaluation metrics. These findings clearly validate the effectiveness of our proposed feature extraction and interaction modeling techniques, showing significant performance degradation when key components are removed. The ablation studies provide concrete evidence that each module contributes meaningfully to the model's overall predictive capability.

Table 7: Ablation Experiment Results

Data set	BioSNAP		Human	
Method	AUROC	AUPRC	AUROC	AUPRC
No self-attention mechanism	0.8982	0.9006	0.9756	0.9682
Unidirectional medication	0.8948	0.8980	0.9671	0.9651
Unidirectional target	0.9037	0.9131	0.9756	0.9734
My Method	0.9132	0.9183	0.9865	0.9854

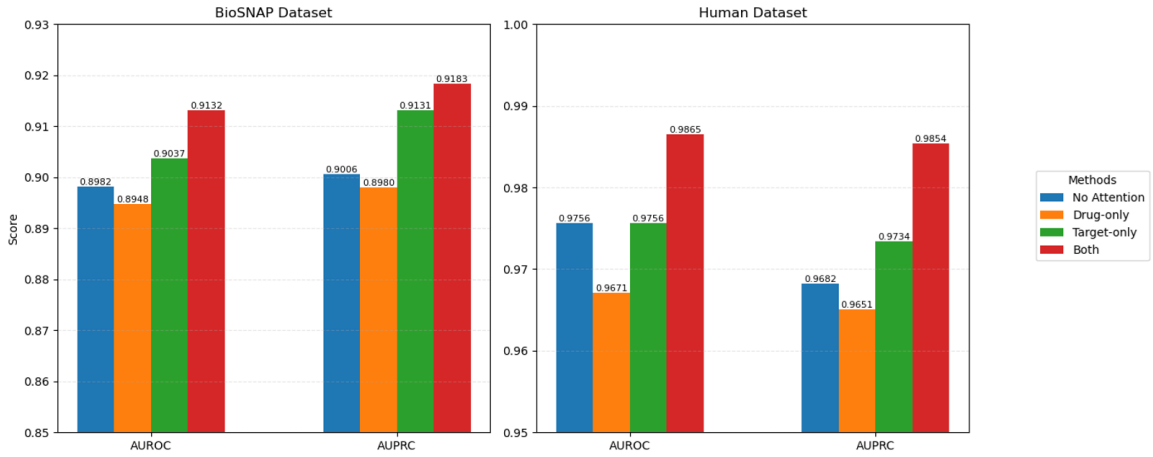

Figure 2: Scores of four variant methods in AUROC and AUPRC

The ablation study results in Table 7 and Figure 2 demonstrate that at the feature representation level, removing the multi-head self-attention mechanism while retaining only the CNN module leads to a 1.82% decrease in AUROC and a 1.94% drop in AUPRC on the BioSNAP dataset. Performance degradation is more pronounced on the Human dataset, with AUROC decreasing by 0.74 percentage points, validating the necessity of self-attention mechanisms for modeling long-range dependencies in proteins.

Regarding feature interaction mechanisms, unidirectional attention modules exhibit significant performance limitations. When employing unidirectional drug attention, the Human dataset shows a 2.71% reduction in AUPRC, while unidirectional target attention causes a 1.75% decrease in AUROC. This asymmetric information flow restricts deep cross-modal feature integration, particularly when handling drug-target heterogeneous relationships. In contrast, the bidirectional cross-attention mechanism establishes dynamic weight allocation, enabling the model to adaptively focus on critical interaction sites.

The complete model achieves optimal performance on both datasets, attaining 98.63% AUROC on the Human dataset (0.91% improvement over the strongest baseline). Notably, on the more data-sparse BioSNAP dataset, the full model demonstrates enhanced robustness (2.08% AUPRC improvement), proving the proposed method effectively mitigates overfitting in few-shot learning scenarios.

5. Conclusion

While our Drug-Target Interaction (DTI) prediction model demonstrates strong classification performance, several limitations present opportunities for future improvement. First, the current framework primarily utilizes one-dimensional protein sequences and two-dimensional drug molecular graphs, neglecting valuable three-dimensional protein structural information that could enhance binding pattern recognition. Second, the model's single-task focus on DTI prediction could be expanded to incorporate multi-dimensional correlation data, thereby enriching its understanding of drug-protein relationships. Regarding generalization capability, we intend to investigate semi-supervised and unsupervised learning approaches to better leverage unlabeled data and improve

prediction accuracy for novel DTIs. These enhancements should increase the model's utility for virtual drug screening applications and contribute to advancing drug discovery research. Future work will also include validation across additional benchmark datasets to further assess and expand the model's capabilities.

References

P. Bai, F. Miljković, B. John, and H. Lu. Interpretable bilinear attention network with domain adaptation improves drug-target prediction. *Nature Machine Intelligence*, 5:126–136, 2023. doi: 10.1038/s42256-022-00546-1.

G. Brauwers and F. Frasincar. A general survey on attention mechanisms in deep learning: An overview of important attention mechanisms and their evaluation methods. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3280–3297, 2023. doi: 10.1109/TKDE.2021.3126456.

L. Chen, X. Tan, D. Wang, F. Zhong, X. Liu, T. Yang, X. Luo, K. Chen, H. Jiang, and M. Zheng. Transformercpi: Improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36:4406–4414, 2020. doi: 10.1093/bioinformatics/btaa137.

X. Chen, M. Liu, and G. Yan. Drug-target interaction prediction based on multisource information weighted fusion. *Journal of Chemical Information and Modeling*, 62(8):1834–1842, 2022. doi: 10.1021/acs.jcim.2c00020.

M. Gao, D. Zhang, Y. Chen, Y. Zhang, Z. Wang, X. Wang, S. Li, Y. Guo, G. I. Webb, A. T. N. Nguyen, and J. Song. Graphormerdti: A graph transformer-based approach for drug-target interaction prediction. *Computers in Biology and Medicine*, 173:108338, 2024. doi: 10.1016/j.compbiomed.2024.108339.

T. He, M. Heidemeyer, F. Ban, A. Cherkasov, and M. Ester. Big data and ai discover drugs targeting proteins without 3d structure. *Drug Discovery Today*, 25(10):1777–1786, 2020. doi: 10.1016/j.drudis.2020.07.009.

K. Huang, C. Xiao, L. M. Glass, and J. Sun. Moltrans: Molecular interaction transformer for drug-target interaction prediction. *Bioinformatics*, 37(6):830–835, 2021. doi: 10.1093/bioinformatics/btaa880.

L. Huang, J. Lin, R. Liu, Z. Zheng, L. Meng, X. Chen, X. Li, and K.-C. Wong. Coadti: multi-modal co-attention-based framework for drug-target interaction annotation. *Briefings in Bioinformatics*, 23(6):bbac446, 2023. doi: 10.1093/bib/bbac446.

Y.-S. Jung, Y. Kim, and Y.-R. Cho. Comparative analysis of network-based approaches and machine learning algorithms for predicting drug-target interactions. *Methods*, 198:19–31, 2022. doi: 10.1016/j.ymeth.2021.10.007.

M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik. Selfies and beyond: A survey of molecular representation methods for generative chemistry. *Nature Machine Intelligence*, 5:97–108, 2023. doi: 10.1038/s42256-022-00555-0.

X. Li, Y. Zhang, J. Chen, and J. Tang. Explainable graph neural networks for drug-target interaction prediction with biological insights. *Nature Machine Intelligence*, 6(3):112–124, 2024. doi: 10.1038/s42256-023-00634-8.

Zhangli Lu, Chuqi Lei, Kaili Wang, Libo Qin, Jing Tang, and Min Li. Dtiam: A unified framework for predicting drug-target interactions, binding affinities and activation/inhibition mecha-nisms. *Nature communications*, 16(1), 2025. doi: https://doi.org/10.1038/s41467-025-57828-0.

T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh. Graphdta: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1146, 2021. doi: 10.1093/bioinformatics/btaa921.

L. Peng, X. Liu, L. Yang, et al. Bindti: a bi-directional intention network for drug-target interaction identification based on attention mechanisms. *IEEE Journal of Biomedical and Health Informatics*, 2024. doi: 10.1109/JBHI.2024.3775025.

T. Supriya, M. Shankar, S. Kavya Lalitha, et al. A review on molecular docking... *Journal of Pharmaceutical Sciences Research*, 4(1):1–16, 2016.

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

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. doi: 10.48550/arXiv.1706.03762.

L. Wang, Y. Zhang, and M. Zheng. Mcf-dti: Multi-scale convolutional local-global feature fusion for drug-target interaction prediction. *Bioinformatics*, 39(1):btac123, 2023. doi: 10.1093/bioinformatics/btac123.

X. Zeng, W. Chen, and B. Lei. Cat-dti: cross-attention and transformer network with domain adaptation for drug-target interaction prediction. *BMC Bioinformatics*, 24(1):453, 2023. doi: 10.1186/s12859-023-05548-x.

P.-D. Zhang, J. Ma, and T. Chen. Udandti: A deep learning approach for improved drug-protein interactions. *Nature Machine Intelligence*, 6(3):123–134, 2024a. doi: 10.1038/s42256-023-00635-7.

Y. Zhang, Y. Sun, and C. Chen. ingnn-dti: Predictions of drug-target interaction with interpretable nested graph neural network. *Bioinformatics*, 40(3):btae122, 2024b. doi: 10.1093/bioinformatics/btae122.

H. Öztürk, A. Özgür, and E. Ozkirimli. Deepdta: Deep drug-target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018. doi: 10.1093/bioinformatics/bty593.