Computers and Artificial Intelligence; Vol.3, No.6; 2026
E-ISSN: 3006-3728 P-ISSN: 3006-371X
DOI: https://doi.org/10.70267/cai.v3n6.141151
Published by: Zeus Press

MFECL-DTI: Multiscale Feature Extraction and Cross-Branch Contrastive Learning for Drug–Target Interaction Prediction

Yexin Chen* and Run Zhong

School of Information, Guizhou University of Finance and Economics, Guiyang, Guizhou, China

**Corresponding author: Yexin Chen.*

Abstract

Accurate drug–target interaction prediction plays an important role in drug discovery and repositioning. However, existing methods often struggle to fully exploit the complementary information between molecular structures and sequence representations. To address this issue, we propose MFECL-DTI, a dual-branch prediction framework integrating graph-based structural encoding, multiscale sequence feature extraction, cross-branch contrastive learning, and adaptive gated fusion. Specifically, the graph branch employs residual graph convolutional networks to extract structural features from drug molecular graphs and protein contact graphs, while the sequence branch combines multiscale convolutional neural networks with a Transformer to capture local patterns and contextual dependencies. Furthermore, a bidirectional contrastive learning mechanism is introduced to align the representations of the same drug–target pair across different branches. An adaptive gating mechanism then dynamically integrates complementary features for interaction prediction. Experiments on BioSNAP, DrugBank, and C.elegans achieve AUROC values of 0.925, 0.915, and 0.994, respectively, with corresponding AUPRC values of 0.927, 0.913, and 0.994. Comparative and ablation experiments demonstrate the effectiveness of the proposed framework in improving DTI prediction performance.

Keywords

drug–target interaction prediction, multimodal learning, self-supervised contrastive learning, adaptive gated fusion, graph neural networks

1. Introduction

Drug–target interaction (DTI) prediction aims to determine whether a candidate drug may interact with a target protein. It is an important task in computer-aided drug discovery and drug repurposing [1]. Given the large number of possible compound–protein combinations, exhaustive experimental screening is often time-consuming and resource-intensive. Computational models can therefore assist in prioritizing candidate interactions. Early methods integrated chemical and genomic information to establish predictive models, laying the foundation for multi-source DTI prediction [2].

In recent years, deep learning has been increasingly used to represent drugs and proteins. DeepDTA employed convolutional networks to encode drug and protein sequences [3], while GraphDTA introduced

© 2026 The Author(s). Authors retain copyright in their articles.

graph neural networks to represent drug molecular topology [4]. MolTrans used Transformer-based architectures to model associations between drug substructures and protein subsequences [5], and DrugBAN adopted bilinear attention to enhance local drug–target interaction modeling [6]. These studies demonstrate the value of both structural and sequential information. However, DeepDTA and GraphDTA primarily address binding affinity prediction, which is not identical to the binary DTI classification task considered here.

Despite their promising performance, methods that rely on a single representation may fail to fully exploit the complementary information contained in drug topology, protein residue contacts, and sequence context. Moreover, representations produced by different encoding branches may occupy different feature spaces, and direct concatenation may not adequately capture their shared information. It is therefore useful to combine complementary encoders with representation alignment and adaptive feature fusion.

To address these issues, we propose MFECL-DTI. The model contains a graph branch and a sequence branch. The former uses residual graph convolution to encode drug molecular graphs and protein contact graphs, while the latter employs a multiscale CNN–Transformer to learn local patterns and contextual dependencies from pretrained sequence embeddings. Cross-branch contrastive learning encourages agreement between the two interaction representations, and adaptive gated fusion produces the final joint representation for binary prediction. We evaluate MFECL-DTI on the BioSNAP, DrugBank, and C.elegans datasets and conduct ablation studies to assess the contribution of its major components.

2. Related work

2.1 Sequence- and Graph-based Drug–target Prediction

Sequence-based methods typically convert drug SMILES strings and protein amino acid sequences into numerical representations and use convolutional or attention-based networks for encoding. DeepDTA is a representative CNN-based method, whereas MolTrans employs Transformers to capture relations between molecular substructures and protein subsequences. Graph convolutional networks aggregate information from neighboring graph nodes [7], and GraphDTA applies graph neural networks to drug molecular representation. Protein language models have also provided useful residue-level representations and structural information [8].

2.2 Multimodal Representation and Feature Integration

Recent multimodal DTI approaches have incorporated pretrained language models, graph representations, and cross-modal fusion. DrugLAMP integrates pretrained features into multimodal interaction modeling [9], CIPFN uses progressive cross-modal fusion [10], and DBCA-DTI explores dual-branch representations with gated attention [11]. TriDTI also investigates cross-modal alignment across multiple representations [12]. These studies highlight the continuing importance of combining heterogeneous information and coordinating different feature branches.

3. Methods

3.1 Overall Framework

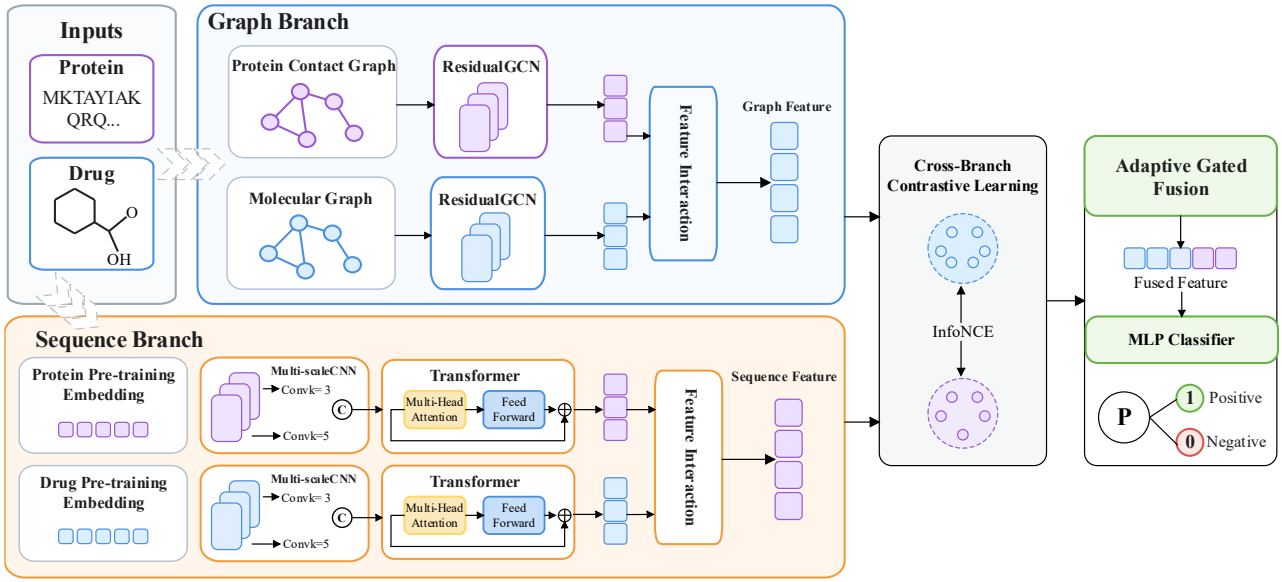
For DTI prediction, a single feature representation may not simultaneously capture molecular structural patterns and sequence context. We therefore develop a dual-branch framework integrating graph convolution, a CNN–Transformer encoder, cross-branch contrastive learning, and adaptive feature fusion. In the graph branch, drug molecular graphs and protein residue contact graphs are processed by residual graph convolutional networks to obtain their structural representations. Drug and protein features are then combined into a graph-based interaction representation.

In parallel, the sequence branch takes pretrained representations of drug SMILES strings and protein amino acid sequences as input. Multiscale convolution extracts local patterns, and a Transformer encoder captures contextual dependencies. The resulting drug and protein representations are combined to obtain a sequence-based interaction representation. Bidirectional contrastive learning aligns the graph and sequence representations of the same drug–target pair during training. Finally, an adaptive gating mechanism balances

© 2026 The Author(s). Authors retain copyright in their articles.

the two branches, and a multilayer perceptron predicts the interaction label of the corresponding drug–target pair.

Figure 1: Overall architecture of the MFECL-DTI model



3.2 Structural Feature Extraction with Residual Graph Convolution

Drug molecular topology and spatial contacts between protein residues provide complementary structural information relevant to DTI. In the graph branch, we construct a molecular graph for each drug and a residue contact graph for each protein. Both are encoded using graph convolutional networks with residual connections and multilevel feature aggregation.

3.2.1 3.2.1 Drug Molecular Graph Representation

For a drug represented by its SMILES string, RDKit is used to construct an undirected molecular graph $G_d = (V_d, E_d)$, where V_d denotes the set of atoms and E_d denotes chemical bonds. Each atom is described by a 75-dimensional feature vector containing atom type, degree, implicit valence, formal charge, hybridization, aromaticity, and hydrogen-related information.

The initial atom features are first mapped into a shared hidden space:

$$H_d^{(0)} = X_d W_{\text{in}} + b_{\text{in}} \tag{1}$$

Here, X_d is the atom feature matrix, and W_{in} and b_{in} are learnable parameters. Self-loops are added to the adjacency matrix and symmetric normalization is applied, with A denoting the adjacency matrix, I the identity matrix, and D the associated degree matrix. To mitigate feature degradation across graph convolution layers, residual connections are introduced:

$$H^{(l+1)} = \text{LN}\left(H^{(l)} + \text{GELU}(\hat{A}H^{(l)}W^{(l)} + b^{(l)})\right) \tag{2}$$

As different graph convolution layers capture complementary neighborhood information, the initial projected features and the outputs of successive layers are concatenated and aggregated through a linear transformation:

$$H_d = W_a[H_d^{(0)} \| H_d^{(1)} \| \cdots \| H_d^{(L)}] + b_a \tag{3}$$

Here, $\|$ denotes the concatenation operator and L represents the number of graph convolution layers. Global mean pooling is then used to obtain a graph-level drug representation:

© 2026 The Author(s). Authors retain copyright in their articles.

$$h_d^g = \frac{1}{|V_d|}\sum_{v\in V_d} H_d(v) \tag{4}$$

Residual connections preserve node-specific information, while multilevel aggregation integrates structural patterns from different neighborhood ranges.

3.2.2 Protein Contact Graph Representation

For target proteins, neighboring positions along the amino acid sequence do not fully describe spatial proximity between residues. We therefore construct a residue contact graph $G_p = (V_p, E_p)$ from protein contact probabilities predicted by ESM-2, where V_p and E_p denote the residue nodes and their connections. The graph contains both sequence-adjacent edges and edges derived from predicted contact probabilities. A sparse graph is constructed using a probability threshold and top-K selection. Candidate contacts with relatively high probabilities are selected from local and long-range residue groups, and edges exceeding the threshold of 0.1 are retained. The original contact probabilities are used as edge weights, after which the graph is symmetrized.

Each amino acid residue is first mapped to a learnable node embedding and subsequently encoded using residual graph convolution similar to that applied to drug molecular graphs. To accommodate the larger protein graphs, sparse edge-based aggregation is used:

$$m_i^{(l)} = \sum_{j\in\mathcal{N}(i)} \frac{a_{ij}}{\sqrt{d_i d_j}} h_j^{(l)} \tag{5}$$

Here, $\mathcal{N}(i)$ denotes the neighbors of node i, a_{ij} is the edge weight, and d_i and d_j are the weighted node degrees. After residual graph convolution, multilevel aggregation, and global mean pooling, the protein structural representation is obtained:

$$h_p^g = \mathrm{MeanPool}(H_p) \tag{6}$$

To model interactions between drug and protein structural representations, we combine their features using concatenation, element-wise multiplication, and the absolute difference:

$$h_g = \phi_g\left([h_d^g \| h_p^g \| h_d^g \odot h_p^g \| |h_d^g - h_p^g|]\right) \tag{7}$$

Here, $\odot$ denotes the element-wise multiplication, ϕ_g while the interaction mapping consists of a linear layer, a ReLU activation, and layer normalization. This yields the graph-branch interaction representation h_g.

3.3 Multiscale CNN–Transformer Sequence Feature Extraction

Drug SMILES strings and protein amino acid sequences contain local motifs as well as contextual dependencies. Convolutional neural networks are effective for extracting local sequence features, while Transformer self-attention can model dependencies between distant positions. We combine multiscale convolution with a Transformer encoder to integrate local and contextual information [13].

3.3.1 Multiscale Local Sequence Encoding

The sequence branch reads pretrained drug and protein representations from precomputed language-model feature caches. The drug embeddings have 768 dimensions and the protein embeddings have 1,024 dimensions. Separate linear projections map these input representations into a 128-dimensional hidden space:

$$X_s = EW_s + b_s \tag{8}$$

Here, E denotes the pretrained sequence embeddings and W_s and b_s are learnable parameters. Two parallel one-dimensional convolutions with kernel sizes of 3 and 5 are subsequently used to extract patterns at different local scales:

$$F_3 = \mathrm{GELU}(\mathrm{Conv}_3(X_s)) \tag{9}$$

$$F_5 = \mathrm{GELU}(\mathrm{Conv}_5(X_s)) \tag{10}$$

© 2026 The Author(s). Authors retain copyright in their articles.

The outputs of the two convolutional pathways are concatenated and combined through a linear mapping:

$$F_{\text{local}} = W_m[F_3 \| F_5] + b_m \tag{11}$$

This operation integrates local patterns at different receptive-field sizes and provides the input for subsequent context modeling.

3.3.2 Transformer-based Contextual Encoding

Although multiscale convolution extracts local patterns, it is less effective at capturing long-range dependencies. We therefore use a Transformer encoder with multihead self-attention. Fixed sinusoidal positional encodings are added to the multiscale CNN features:

$$X_t = F_{\text{local}} + PE \tag{12}$$

Here, PE denotes positional encoding. The resulting sequence features are processed by a Transformer encoder. Scaled dot-product attention is defined as follows:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{13}$$

Multihead attention learns contextual relationships in multiple feature subspaces. The Transformer output is combined with the multiscale CNN features through a residual connection and layer normalization:

$$H_s = \text{LN}(\text{Transformer}(X_t) + F_{\text{local}}) \tag{14}$$

Masked mean pooling over valid sequence positions yields the drug sequence representation h_d^s and protein sequence representation h_p^s. As in the graph branch, the two representations are combined using concatenation, element-wise multiplication, and absolute difference:

$$h_g^i = \phi_s\left([h_d^s \| h_p^s \| h_d^s \odot h_p^s \| | h_d^s - h_p^s |]\right) \tag{15}$$

The resulting feature h_s represents the drug–target pair in the sequence branch.

3.4 Cross-branch Feature Alignment via Contrastive Learning

The graph and sequence branches describe the same drug–target pair from the perspectives of structural topology and sequence context. Because these features are generated by different encoders, their representation spaces may differ.

To promote cross-branch consistency, we introduce an InfoNCE-based contrastive objective [14]. For a minibatch of N drug–target pairs, the graph representation h_g^i and sequence representation h_s^i are projected into a shared contrastive space by two independent linear layers and then L2-normalized:

$$z_g^i = \text{Norm}(W_g h_g^i + b_g) \tag{16}$$

$$z_s^i = \text{Norm}(W_s h_s^i + b_s) \tag{17}$$

The normalized graph and sequence representations are denoted by z_g^i and z_s^i, respectively. For each drug–target pair, the two representations produced by the different branches form a positive pair. Cross-branch representations from different samples in the same minibatch are treated as negatives. We calculate the contrastive loss in both graph-to-sequence and sequence-to-graph directions. The graph-to-sequence contrastive loss is defined as:

$$\mathcal{L}_{g\to s} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(z_g^i, z_s^i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(z_g^i, z_s^j)/\tau)} \tag{18}$$

Similarly, the sequence-to-graph contrastive loss is:

© 2026 The Author(s). Authors retain copyright in their articles.

$$\mathcal{L}_{s\to g} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathrm{sim}(z_s^i, z_g^i)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(z_s^i, z_g^j)/\tau)} \tag{19}$$

Here, sim() denotes the dot-product similarity of normalized vectors, and τ is a temperature parameter controlling the smoothness of the similarity distribution. The final cross-branch contrastive loss is the average of the two directional losses:

$$\mathcal{L}_{\mathrm{CL}} = \frac{1}{2}\left(\mathcal{L}_{g\to s} + \mathcal{L}_{s\to g}\right) \tag{20}$$

By encouraging the representations of the same pair to approach each other in the contrastive space, this auxiliary objective improves cross-branch consistency without eliminating the complementary characteristics of the two branches. Contrastive learning is used only during training; prediction relies directly on the graph and sequence representations.

3.5 Adaptive Gated Fusion and Interaction Prediction

After feature extraction, each drug–target pair is represented by a graph feature h_g and a sequence feature h_s. As the relative contribution of structural and sequential information can vary across samples, simple concatenation or fixed weighting may not be sufficiently flexible. An adaptive feature-level gate is therefore learned from the concatenation of the two representations using two fully connected layers followed by a sigmoid activation:

$$\alpha = \sigma\left(W_2(W_1[h_g \| h_s] + b_1) + b_2\right) \tag{21}$$

Here, $\alpha \in (\mathbf{0},\mathbf{1})^{128}$ is the adaptive gating vector and σ denotes the sigmoid function. The two branch representations are then integrated through element-wise weighted fusion:

$$h_f = \alpha \odot h_g + (\mathbf{1} - \alpha) \odot h_s \tag{22}$$

The fused feature h_f is passed to a multilayer perceptron, and a softmax layer produces the probabilities of the two interaction classes. Class 1 denotes an interaction and class 0 denotes a non-interaction.

3.6 Objective Function

We formulate DTI prediction as a binary classification task and use cross-entropy as the primary supervised loss:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=0}^{1} y_{ic}\log P(y = c \mid d_i, p_i) \tag{23}$$

Here, y_{ic} is a binary indicator that equals 1 if sample i belongs to class c, and 0 otherwise. To jointly optimize interaction classification and cross-branch representation consistency, we employ the following objective:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{CE}} + \lambda\mathcal{L}_{\mathrm{CL}} \tag{24}$$

The coefficient λ controls the relative contribution of the contrastive loss to the classification objective.

4. Experiments and Results

4.1 Datasets

We evaluate MFECL-DTI on the BioSNAP, DrugBank, and C.elegans datasets. Each dataset contains drugs, proteins, and binary drug–target interaction labels, where 1 indicates an interaction and 0 indicates a non-interaction. The numbers of drug entities, protein entities, and interaction pairs are summarized in Table 1.

© 2026 The Author(s). Authors retain copyright in their articles.

Table 1: Statistics of the datasets

Dataset	Drugs	Proteins	Pairs	Positive	Negative
BioSNAP	4510	2181	27,464	13,830	13,634
DrugBank	6655	4294	37,314	18,657	18,657
C.elegans	1767	1876	7786	3893	3893

4.2 Experimental Settings

MFECL-DTI is implemented in PyTorch and trained using the AdamW optimizer. The graph branch comprises separate residual GCN encoders for drug molecular graphs and protein contact graphs; both use a hidden feature dimension of 128 and two graph convolution layers. The sequence branch consists of two parallel one-dimensional convolutions with kernel sizes 3 and 5, followed by a single Transformer encoder layer with four attention heads. Both branch-level interaction representations have 128 dimensions. For bidirectional InfoNCE learning, the temperature is set to 0.2 and the loss coefficient to 0.05.

Table 2: Model hyperparameters

Parameter	Value
Batch size	64
Epochs	40
Dropout rate	0.1
Learning rate	5e-4
Optimizer	AdamW
Weight decay	0.01
SMILES max length	100
Protein max length	1200

4.3 Evaluation Metrics

We assess predictive performance using six metrics: area under the receiver operating characteristic curve (AUROC), area under the precision–recall curve (AUPRC), accuracy, precision, recall, and F1-score. AUROC evaluates discrimination between positive and negative pairs across classification thresholds, and AUPRC summarizes precision at different recall levels. Accuracy measures the proportion of correct predictions; precision reflects the reliability of predicted positives; recall measures coverage of true positives; and F1-score is the harmonic mean of precision and recall.

4.4 Baseline Methods

We compare MFECL-DTI with nine representative DTI prediction approaches spanning convolutional, attention-based, graph-based, and multimodal modeling strategies.

(1) TransformerCPI employs Transformer self-attention to capture compound–protein relationships for interaction prediction [15].

(2) MolTrans separately encodes drug and protein subsequences and models their interactions through joint representations.

(3) HyperAttentionDTI combines convolutional sequence features with attention-based modeling of atom–amino acid associations [16].

(4) DrugBAN uses bilinear attention to capture local drug–protein interactions and improve joint representations.

(5) MCANet adopts shared-weight multihead cross-attention to capture drug–target associations [17].

(6) BINDTI combines graph-based drug encoding, an ACmix protein encoder, and bidirectional intention attention for feature interaction [18].

(7) MGNDTI uses multimodal representations and a gating mechanism to fuse information from different sources [19].

(8) MMDG-DTI combines pretrained language-model features with hybrid graph neural networks and domain-generalization strategies [20].

© 2026 The Author(s). Authors retain copyright in their articles.

(9) DACMF-DTI integrates one-dimensional convolution, dual attention, and cross-modal fusion to characterize drug–target substructure interactions [21].

4.5 Performance Comparison

For a systematic evaluation, the datasets are divided into training, validation, and test sets at a ratio of 7:1:2, following the partitions adopted in the source benchmark study. The training data are used for parameter optimization, the validation data for monitoring performance and selecting the best checkpoint, and the held-out test data for final evaluation.

To reduce the effect of stochastic variation, MFECL-DTI is trained and evaluated with five different random seeds. We report the mean and standard deviation of each metric over these runs, with the best values in each comparison table shown in bold. The baseline values are cited from the published benchmark study, whereas the MFECL-DTI results are obtained from our own experiments under the same dataset partitions. Differences in implementation and model tuning should still be considered when interpreting the comparisons.

The comparative results for BioSNAP, DrugBank, and C.elegans are presented in Tables 3–5.

On BioSNAP, MFECL-DTI achieves an AUROC of 0.925 ± 0.005 and an AUPRC of 0.927 ± 0.003, outperforming the listed baselines on both metrics. Compared with HyperAttentionDTI, the strongest baseline for these metrics, the absolute increases are 0.015 in AUROC and 0.008 in AUPRC. The model also achieves an accuracy of 0.859 ± 0.004 and an F1-score of 0.862 ± 0.004, indicating a relatively balanced classification performance.

On DrugBank, the AUROC and AUPRC of MFECL-DTI are 0.915 ± 0.002 and 0.913 ± 0.003, respectively. Relative to the strongest baseline, HyperAttentionDTI (0.890 and 0.895), the absolute improvements are 0.025 and 0.018. The model obtains accuracy, precision, recall, and F1-score values of approximately 0.834, 0.841, 0.850, and 0.843, respectively, demonstrating the utility of the combined structural and sequential representations.

Table 3: Comparison of predictive performance on BioSNAP (mean ± standard deviation)

Method	AUROC	AUPRC	Accuracy	Precision	Recall	F1-score
TransformerCPI	0.881 ± 0.002	0.883 ± 0.008	0.812 ± 0.001	0.808 ± 0.002	0.822 ± 0.007	0.815 ± 0.002
MolTrans	0.880 ± 0.006	0.885 ± 0.009	0.803 ± 0.004	0.784 ± 0.007	0.838 ± 0.004	0.800 ± 0.003
HyperAttentionDTI	0.910 ± 0.005	0.919 ± 0.004	0.832 ± 0.015	0.815 ± 0.026	0.861 ± 0.010	0.837 ± 0.013
DrugBAN	0.904 ± 0.001	0.912 ± 0.002	0.836 ± 0.001	0.834 ± 0.003	0.840 ± 0.009	0.837 ± 0.002
MCANet	0.909 ± 0.002	0.918 ± 0.003	0.836 ± 0.001	0.834 ± 0.004	0.842 ± 0.011	0.838 ± 0.003
BINDTI	0.902 ± 0.002	0.900 ± 0.001	0.835 ± 0.001	0.827 ± 0.008	0.848 ± 0.011	0.837 ± 0.001
MGNDTI	0.901 ± 0.004	0.903 ± 0.009	0.828 ± 0.003	0.811 ± 0.015	0.852 ± 0.022	0.831 ± 0.006
MMDG-DTI	0.909 ± 0.001	0.912 ± 0.001	0.831 ± 0.004	0.838 ± 0.007	0.823 ± 0.015	0.830 ± 0.006
DACMF-DTI	0.900 ± 0.002	0.895 ± 0.006	0.830 ± 0.005	0.827 ± 0.017	0.830 ± 0.008	0.828 ± 0.004
MFECL-DTI	**0.925 ± 0.005**	**0.927 ± 0.003**	**0.859 ± 0.004**	**0.840 ± 0.004**	**0.861 ± 0.005**	**0.862 ± 0.004**

Table 4: Comparison of predictive performance on DrugBank (mean ± standard deviation)

Method	AUROC	AUPRC	Accuracy	Precision	Recall	F1-score
TransformerCPI	0.863 ± 0.001	0.871 ± 0.001	0.788 ± 0.001	0.779 ± 0.003	0.804 ± 0.005	0.791 ± 0.002
MolTrans	0.848 ± 0.004	0.851 ± 0.011	0.765 ± 0.002	0.733 ± 0.012	0.830 ± 0.024	0.771 ± 0.007
HyperAttentionDTI	0.890 ± 0.002	0.895 ± 0.001	0.812 ± 0.003	0.795 ± 0.008	0.833 ± 0.008	0.814 ± 0.000
DrugBAN	0.879 ± 0.002	0.885 ± 0.002	0.812 ± 0.002	0.810 ± 0.008	0.808 ± 0.006	0.809 ± 0.001
MCANet	0.885 ± 0.002	0.889 ± 0.006	0.810 ± 0.006	0.804 ± 0.012	0.817 ± 0.006	0.810 ± 0.005
BINDTI	0.876 ± 0.001	0.871 ± 0.000	0.809 ± 0.002	0.803 ± 0.013	0.817 ± 0.017	0.813 ± 0.002
MGNDTI	0.885 ± 0.003	0.886 ± 0.006	0.801 ± 0.004	0.773 ± 0.002	0.849 ± 0.006	0.810 ± 0.004
MMDG-DTI	0.887 ± 0.001	0.885 ± 0.002	0.803 ± 0.005	0.794 ± 0.015	0.812 ± 0.013	0.803 ± 0.001
DACMF-DTI	0.870 ± 0.001	0.872 ± 0.004	0.788 ± 0.002	0.778 ± 0.006	0.810 ± 0.010	0.793 ± 0.001
MFECL-DTI	**0.915 ± 0.002**	**0.913 ± 0.003**	**0.834 ± 0.003**	**0.841 ± 0.005**	**0.850 ± 0.009**	**0.843 ± 0.003**

On C.elegans, the compared methods all achieve high values of the primary ranking metrics under the present evaluation protocol. MFECL-DTI reaches an AUROC and an AUPRC of 0.994 ± 0.002, approximately 0.001 above the best listed baseline values. However, MFECL-DTI does not achieve the highest accuracy, precision, or F1-score. These results indicate that some trade-offs remain among evaluation metrics on this dataset.

© 2026 The Author(s). Authors retain copyright in their articles.

Table 5: Comparison of predictive performance on C.elegans (mean ± standard deviation)

Method	AUROC	AUPRC	Accuracy	Precision	Recall	F1-score
TransformerCPI	0.982 ± 0.003	0.977 ± 0.008	0.946 ± 0.005	0.942 ± 0.005	0.951 ± 0.008	0.947 ± 0.006
MolTrans	0.991 ± 0.002	0.991 ± 0.001	0.969 ± 0.002	0.977 ± 0.0063	0.959 ± 0.011	0.966 ± 0.003
HyperAttentionDTI	0.992 ± 0.001	0.992 ± 0.001	**0.970 ± 0.004**	0.971 ± 0.004	0.970 ± 0.004	0.970 ± 0.002
DrugBAN	0.990 ± 0.001	0.989 ± 0.002	0.967 ± 0.001	0.974 ± 0.006	0.959 ± 0.004	0.967 ± 0.001
MCANet	0.993 ± 0.002	0.993 ± 0.002	0.968 ± 0.004	0.976 ± 0.005	0.960 ± 0.004	0.968 ± 0.004
BINDTI	0.990 ± 0.001	0.989 ± 0.001	0.964 ± 0.002	0.965 ± 0.004	0.961 ± 0.004	0.963 ± 0.007
MGNDTI	0.992 ± 0.001	0.992 ± 0.001	**0.970 ± 0.002**	0.975 ± 0.004	0.970 ± 0.004	**0.972 ± 0.003**
MMDG-DTI	0.993 ± 0.001	0.992 ± 0.001	0.969 ± 0.005	**0.979 ± 0.003**	0.962 ± 0.011	0.969 ± 0.005
DACMF-DTI	0.991 ± 0.002	0.992 ± 0.002	0.967 ± 0.010	0.968 ± 0.016	0.967 ± 0.008	0.968 ± 0.009
MFECL-DTI	**0.994 ± 0.002**	**0.994 ± 0.002**	0.967 ± 0.003	0.965 ± 0.001	**0.971 ± 0.002**	0.967 ± 0.001

4.6 Ablation Study

To investigate the contribution of the principal components, we conduct three ablation studies against the full model. In the first variant (w/o Contrastive Learning), the cross-branch InfoNCE loss is removed while both branches and gated fusion are retained. In the second (w/o Graph Branch), the drug and protein graph encoders are removed and only the sequence branch is used. In the third (w/o Sequence Branch), the CNN–Transformer sequence encoders are removed and only the graph branch is retained. The single-branch variants do not use cross-branch contrastive learning or gated fusion.

On BioSNAP, the full model achieves an AUROC of 0.9270 and an AUPRC of 0.9271. Removing contrastive learning reduces the corresponding metrics to 0.9147 and 0.9206, absolute decreases of 0.0123 and 0.0065. Without the graph branch, the values become 0.9180 and 0.9154, whereas using the graph branch alone results in 0.8751 and 0.8859. These results suggest that the sequence branch provides a major source of predictive information, while structural features contribute complementary information and contrastive learning benefits the joint representations.

Table 6: Ablation results on BioSNAP

Variant	AUROC	AUPRC	Accuracy	Precision	Recall	F1-score
MFECL-DTI (full)	**0.9250**	**0.9271**	**0.8591**	0.8403	**0.8615**	**0.8624**
w/o contrastive loss	0.9147	0.9206	0.8351	0.8282	0.8457	0.8369
w/o graph branch	0.9180	0.9154	0.8489	**0.8406**	0.8614	0.8508
w/o sequence branch	0.8751	0.8859	0.7819	0.7515	0.8428	0.7945

Table 7: Ablation results on DrugBank

Variant	AUROC	AUPRC	Accuracy	Precision	Recall	F1-score
MFECL-DTI (full)	**0.9151**	**0.9129**	**0.8344**	**0.8416**	**0.8509**	**0.8462**
w/o contrastive loss	0.9054	0.9030	0.8287	0.8285	0.8290	0.8287
w/o graph branch	0.9072	0.9062	0.8315	0.8391	0.8204	0.8296
w/o sequence branch	0.8595	0.8643	0.7792	0.7737	0.7890	0.7813

On DrugBank, the full model reaches an AUROC of 0.9151 and an AUPRC of 0.9129. Removing contrastive learning reduces these values to 0.9054 and 0.9030; removing the graph branch yields 0.9072 and 0.9062; and removing the sequence branch further decreases them to 0.8595 and 0.8643. The trend is consistent with that observed on BioSNAP.

Table 8: Ablation results on C.elegans

Variant	AUROC	AUPRC	Accuracy	Precision	Recall	F1-score
MFECL-DTI (full)	**0.9944**	**0.9944**	**0.9673**	**0.9655**	**0.9692**	**0.9673**
w/o contrastive loss	0.9918	0.9916	0.9641	0.9605	0.9679	0.9642
w/o graph branch	0.9919	0.9915	0.9615	0.9627	0.9602	0.9614
w/o sequence branch	0.9857	0.9864	0.9512	0.9571	0.9448	0.9509

On C.elegans, the full model achieves AUROC and AUPRC values of 0.9944. The corresponding values are 0.9918 and 0.9916 without contrastive learning, 0.9919 and 0.9915 without the graph branch, and 0.9857 and 0.9864 without the sequence branch. Although the overall metrics are high, the full model performs best among the tested variants on AUROC and AUPRC. Across the three datasets, the results support the combination of structural and sequence features with cross-branch alignment. Since both contrastive learning

© 2026 The Author(s). Authors retain copyright in their articles.

and gated fusion are unavailable in the single-branch variants, their performance differences cannot be attributed solely to the removed encoder.

5. Conclusion

We propose MFECL-DTI, a drug–target interaction prediction model integrating multiscale feature extraction, cross-branch contrastive learning, and adaptive gated fusion. A residual GCN encodes the structural information in drug molecular graphs and protein contact graphs, while a multiscale CNN–Transformer encodes sequence representations. Cross-branch contrastive alignment and adaptive fusion combine the two sources of information. On BioSNAP, DrugBank, and C.elegans, MFECL-DTI achieves competitive AUROC and AUPRC scores, and ablation results support the complementary roles of the two branches and the contribution of contrastive learning.

The present evaluation is based on existing dataset partitions and does not fully assess generalization to previously unseen drugs or proteins. Future work may explore cold-start settings, computational efficiency, and richer protein structural information to evaluate the method in more challenging practical scenarios.

References

[1] Shi W, Yang H, Xie L, et al. A review of machine learning-based methods for predicting drug-target interactions. Health Information Science and Systems. 2024; 12: 30.

[2] Yamanishi Y, Araki M, Gutteridge A, et al. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008; 24(13): i232–i240.

[3] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics. 2018; 34(17): i821–i829.

[4] Nguyen T, Le H, Quinn T P, et al. GraphDTA: predicting drug–target binding affinity with graph neural networks. Bioinformatics. 2021; 37(8): 1140–1147.

[5] Huang K, Xiao C, Glass L M, et al. MolTrans: Molecular Interaction Transformer for drug–target interaction prediction. Bioinformatics. 2021; 37(6): 830–836.

[6] Bai P, Miljković F, John B, et al. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. Nature Machine Intelligence. 2023; 5: 126–136.

[7] Kipf T N, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. International Conference on Learning Representations, 2017.

[8] Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023; 379(6637): 1123–1130.

[9] Luo Z, Wu W, Sun Q, et al. Accurate and transferable drug-target interaction prediction with DrugLAMP. Bioinformatics. 2024; 40(12): btae693.

[10] Cao Z, Xie J, Xu J, et al. Cross-Modal Interaction-Aware Progressive Fusion Network for Drug-Target Interaction Prediction. Journal of Chemical Information and Modeling. 2025; 65(19): 10713–10726.

[11] Peng J, Liu X, Zhou X, et al. DBCA-DTI: A dual-branch multimodal framework based on bidirectional adaptive gated cross-attention mechanism for drug-target interaction prediction. Knowledge-Based Systems. 2026; 338: 115463.

[12] Yun G H, Park J H, Cho Y R. TriDTI: tri-modal representation learning with cross-modal alignment for drug-target interaction prediction. Briefings in Bioinformatics. 2026; 27(1): bbag034.

[13] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. Advances in Neural Information Processing Systems. 2017; 30.

[14] Tian Z, Yu Y, Ni F, et al. Drug-target interaction prediction with collaborative contrastive learning and adaptive self-paced sampling strategy. BMC Biology. 2024; 22: 216.

© 2026 The Author(s). Authors retain copyright in their articles.

[15] Chen L, Tan X, Wang D, et al. TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics. 2020; 36(16): 4406–4414.

[16] Zhao Q, Zhao H, Zheng K, et al. HyperAttentionDTI: improving drug-protein interaction prediction by sequence-based deep learning with attention mechanism. Bioinformatics. 2022; 38(3): 655–662.

[17] Bian J, Zhang X, Zhang X, et al. MCANet: shared-weight-based MultiheadCrossAttention network for drug-target interaction prediction. Briefings in Bioinformatics. 2023; 24(2): bbad082.

[18] Peng L, Liu X, Yang L, et al. BINDTI: A Bi-Directional Intention Network for Drug-Target Interaction Identification Based on Attention Mechanisms. IEEE Journal of Biomedical and Health Informatics. 2025; 29(3): 1602–1612.

[19] Peng L, Liu X, Chen M, et al. MGNDTI: A Drug-Target Interaction Prediction Framework Based on Multimodal Representation Learning and the Gating Mechanism. Journal of Chemical Information and Modeling. 2024; 64(16): 6684–6698.

[20] Hua Y, Feng Z, Song X, et al. MMDG-DTI: Drug-target interaction prediction via multimodal feature fusion and domain generalization. Pattern Recognition. 2025; 157: 110887.

[21] Wei Z, Zheng X, Li C, et al. DACMF-DTI: Dual attention embedded cross-modality fusion for drug-target interaction prediction. Knowledge-Based Systems. 2025; 326: 114063.

Funding

This research was funded by Postgraduate Education and Teaching Reform Project of Guizhou Province in 2025 (No. 2025YJSKYJJ223).

Conflicts of Interest

The author(s) declare no conflicts of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

© 2026 The Author(s). Authors retain copyright in their articles.