Briefings in Bioinformatics, 2025, **26(5)**, bbaf474

https://doi.org/10.1093/bib/bbaf474

Problem Solving Protocol

MIF–DTI: a multimodal information fusion method for drug–target interaction prediction

Jiehong Shan[1,2], Jinchen Sun[1,2], Haoran Zheng[1,2,*]

[1]School of Computer Science and Technology, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China

[2]Anhui Key Laboratory of Software Engineering in Computing and Communication, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China

[*]Corresponding author: Haoran Zheng, School of Computer Science and Technology, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China. E-mail: hrzheng@ustc.edu.cn

Abstract

Drug–target interaction (DTI) prediction is essential for drug discovery and repurposing. To overcome the limitations of current DTI prediction methods that rely on single-source encoding and inadequately fuse multimodal information, this study proposes a DTI prediction method based on multimodal information fusion (MIF–DTI) and further designs an ensemble version (MIF–DTI-B). MIF–DTI encodes the SMILES sequences of drugs and the amino acid sequences of targets via a sequence encoding module to extract their 1D sequence features. It conducts dual-view representation encoding on the hierarchical molecular graphs of drugs and the contact graphs of targets through a graph encoding module, aiming to capture their 2D topological structure information. A decoding module is utilized to fuse information from different modalities. MIF–DTI-B ensembles several MIF–DTI models through cross-validation strategy to improve predictive accuracy. This study evaluates the proposed models on three publicly accessible DTI datasets. Experimental results demonstrate that fully integrating multimodal information enables both MIF–DTI and MIF–DTI-B to consistently outperform state-of-the-art methods.

Keywords: drug–target interaction; multimodal information fusion; dual-view representation learning; ensemble model; deep learning

Introduction

Drug discovery and drug repurposing are significant research avenues in the biomedical domain. Drugs exert their intended therapeutic effects by binding to corresponding targets, such as proteins, genes, and producing interactions. Therefore, drug–target interactions (DTIs) are the basis of pharmacological activity [1]. DTI prediction can indicate whether a drug compound will interact with a target, hence offering essential evidence for assessing pharmacological activity and anticipated therapeutic effects. Consequently, it has become a core task in drug discovery and repurposing research. Conventional DTI identification predominantly depends on an array of biochemical investigations [2]. While reliable, this method involves complex procedures, long experimental durations, and high costs, limiting its applicability to large-scale data analysis.

With the continuous accumulation of biomedical data and the rapid development of high-performance computing technologies, *in silico* methods have become important for DTI prediction [3–5]. These methods offer significant advantages in improving prediction accuracy and efficiency while reducing costs. Currently, *in silico* methods for DTI prediction can be classified into three categories: ligand-based methods, structure-based methods, and deep learning-based methods [6]. Ligand-based methods assume that drugs with similar structures or side-effect profiles tend to interact with the same targets, and vice versa [7]. These methods rely on prior knowledge of known active ligands and infer potential DTIs indirectly by computing similarities between drugs and between targets. However, their efficacy diminishes when only a limited number of active compounds are identified for a specific target [8].

Structure-based methods use the 3D structures of compounds and proteins to identify DTIs, including techniques such as molecular docking, molecular dynamics simulation, and binding free energy prediction. In 2020, Gentile *et al*. [9] proposed DeepDocking, using deep neural networks to accelerate 3D docking. Milon *et al*. [10] utilized Triangular Spatial Relationship (TSR) bonds to analyze drug and target 3D structures, predicting DTIs and drug–target binding sites through TSR bond overlap. Compared with methods based on 1D sequences or molecular fingerprints [11–13] and 2D images [14, 15], structure-based methods capture structural details more comprehensively, improving accuracy. However, they are limited when the 3D structure of the target protein is unknown.

Deep learning has recently shown strong performance in bioinformatics. Early deep learning-based methods usually extract information from the SMILES sequence or ECFP representation of drugs and the amino acid sequence of targets using shallow networks. In 2018, the DeepDTA model proposed by Ozturk *et al*. [16] uses the convolutional neural network (CNN) to encode SMILES sequences and amino acid sequences, subsequently utilizing fully

Received: May 31, 2025. **Revised:** August 5, 2025. **Accepted:** August 20, 2025

connected layers for prediction. In 2019, Lee *et al.* [17] proposed the DeepConv-DTI, which uses CNNs to encode ECFP and amino acid sequences, achieving more accurate DTI prediction.

With the rise of attention mechanisms and models like Transformer [18], deep learning-based methods have gradually focused on decomposing drugs and targets and mining their interaction information. In 2021, Huang *et al.* [19] proposed MolTrans, which constructs fragment libraries of drugs and targets from unlabeled data, encodes fragment sequences via Transformer, and computes interaction matrices between them. In 2023, Bian *et al.* [20] introduced MCANet, which utilizes 1D CNNs to encode the SMILES sequences and amino acid sequences, employing cross-attention to learn interaction information between them. In 2024, Zhang *et al.* [21] further proposed FMCA-DTI, which uses fragment mining and multi-head cross-attention to capture mutual information.

Meanwhile, other studies have explored encoding drugs and targets as 2D graphs to fully leverage topological structural information. In 2022, Li *et al.* [22] proposed MINN-DTI. It uses a message delivery network to encode the molecular graph of the drug, uses a message passing network to encode drug molecular graphs and amino acid distance graphs of targets, followed by a Transformer to capture the interaction information between them. In 2024, Koh *et al.* [23] proposed PSICHIC, which uses a graph neural network (GNN) to encode both drug and predicted target contact graphs, with attention mechanisms to learn interaction features.

Current research identifies prevalent information sources for DTI prediction as 1D sequences, 2D graphs, 3D structures, and external knowledge. Although 3D structures and external knowledge offer rich information, they encounter challenges like data scarcity, high computational cost, and integration difficulty. Thus, most studies focus on more accessible 1D sequences and 2D graphs. The former is simpler to encode but less expressive, while the latter provides richer structure but necessitates intricate encoding. In short, 1D sequences and 2D graphs are complementary in encoding efficiency and representational power, but most methods still adopt a single modality.

Recently, some recent models, such as BINDTI [24] and 3DProtDTA [25], have begun to explore multimodal fusion. However, their integration strategies, which frequently focus on representation enhancement or simple feature concatenation, still face limitations in achieving a truly effective multimodal fusion. To this end, this study proposes MIF–DTI, a framework that uniquely combines dual-view representation learning with a decision-focused co-attention module to achieve a deeper and more direct fusion of 1D sequence and 2D graph data. The main technical contributions of this method are as follows:

- **2D structural encoding**: for the 1D sequences of drugs and targets, a hierarchical molecular graph is constructed for drugs via a substructure extraction module, and a 2D contact graph is generated for targets using the ESM-2 pre-trained model, effectively capturing their topological structure.
- **Dual-view representation learning**: a dual-view representation learning mechanism is introduced in both the sequence and graph encoding modules to capture intra-molecular structural features and inter-molecular interaction information, which improves prediction accuracy.
- **Multimodal features encoding and fusion**: MIF–DTI extracts low-level adjacency features from 1D sequences and high-level topological features from 2D graphs, enabling dual-source information integration. A collaborative attention mechanism fuses sequence and graph modalities, and an interaction score matrix estimates DTI probabilities, fully leveraging multimodal information.

Based on MIF–DTI, we further developed an ensemble model named MIF–DTI-B, which enhances performance in DTI prediction. Extensive experimental results demonstrate that both MIF–DTI and MIF–DTI-B surpass existing state-of-the-art methods in overall performance.

Method

This section first presents a formal definition of the research problem, then introduces the main components of the overall MIF–DTI framework, including the sequence encoding module, the graph encoding module, and the MIF decoding module. The ensemble model MIF–DTI-B is also described, and finally explains the loss function used.

Problem definition

In recent years, some studies [23, 26] have treated DTI prediction as a regression task, aiming to predict continuous indicators such as binding affinity or half-maximal inhibitory concentration. However, most studies [22, 27, 28] have formulated it as a binary classification task, predicting whether an interaction exists between a drug and a target. To facilitate comparison with mainstream baseline methods such as PSICHIC [23], BINDTI [24], and MCANet [20], this study uses the same setting and formulates DTI prediction as a binary classification task.

Given a set of drugs $\mathcal{D}$ and a set of targets $\mathcal{T}$, the DTI prediction task can be formalized as a function: $f : \mathcal{D} \times \mathcal{T} \to [0, 1]$. Herein, $f(D_x, T_y) = 1$ means that there exists an interaction between drug D_x and target T_y, and 0 otherwise. The goal of this study is to learn an approximate function of f that predicts the existence of interaction between a given drug and target.

MIF–DTI framework

MIF–DTI takes the drug's SMILES sequence and the target's amino acid sequence as inputs, and consists of a sequence encoding module, a graph encoding module, and an MIF decoding module. The overall architecture is shown in Fig. 1. Specifically, the sequence encoding module uses CNNs to extract sequence information from drugs and targets, and uses a cross-attention mechanism to capture 1D interaction information. The graph encoding module converts the drug's SMILES sequence into a hierarchical molecular graph and the target's amino acid sequence into a contact graph. It then utilizes graph attention networks (GATs) and dual-view representation learning to capture topological features and 2D interaction information. The MIF decoding module calculates multimodal fusion coefficients via a collaborative attention mechanism, computes the interaction score matrix through matrix multiplication, and finally aggregates the score matrix with a fully connected layer to estimate the DTI probability.

Sequence encoding module

The MIF sequence encoding module (Fig. 2) includes an embedding part and multiple MIF-1D encoding blocks to generate multi-depth drug and target representations.

For SMILES and amino acid sequences, characters are encoded into integers (1–64 for SMILES, 1–24 for amino acids), following the method in MCANet [20]. Sequences are aligned to fixed lengths (200 for drugs, 1500 for targets) by zero-padding or truncation, and

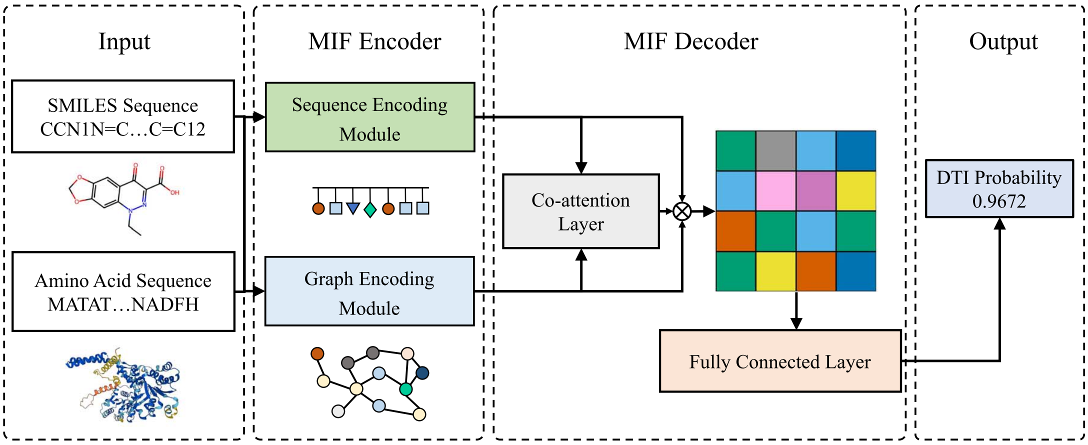

Figure 1. Overall architecture of MIF–DTI with sequence and graph encoders and a multimodal decoder for DTI prediction.

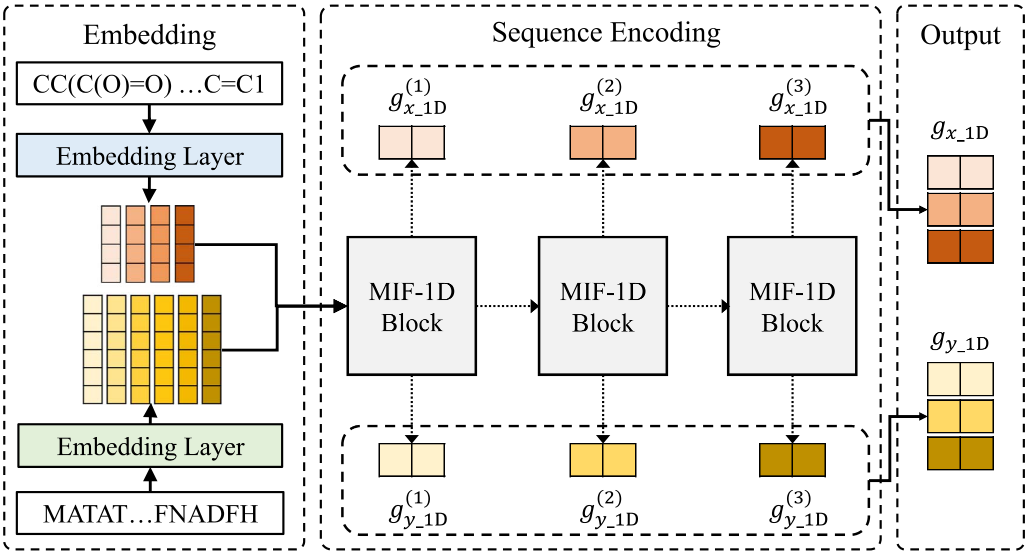

Figure 2. Structure of the MIF sequence encoding module with embedding and MIF-1D encoding blocks for multi-depth global representations of drugs and targets.

then mapped to high-dimensional representations via separate embedding layers.

The overall processing flow of the MIF-1D encoding block is shown in Fig. 3. For the embedded representations of SMILES sequences and amino acid sequences, multiple layers of 1D CNN are used to extract local features by sliding fixed-size convolution kernels along the sequences. Given an embedded representation $h \in \mathbb{R}^{L\times d}$, where L is the sequence length and d is the embedding dimension, a single-layer 1D CNN is computed as Equation (1):

$$h' = \text{ReLU}\left(\text{CNN}\left(h\right)\right) \tag{1}$$

By stacking multiple 1D CNN layers, the MIF-1D encoding block progressively extracts features from different regions. It then applies a shared-weight multi-head attention mechanism, using the drug sequence representation as the query and target sequence representation as key-value pairs to generate the interaction representation of the drug, and vice versa for the target.

Let L_D and L_T denote the encoded lengths of SMILES and amino acid sequences after CNN layers, and d_C the encoding dimension. Their representations are $h_x \in \mathbb{R}^{L_D\times d_C}$ and $h_y \in \mathbb{R}^{L_T\times d_C}$. The attention module computes the drug query $q_x^i \in \mathbb{R}^{L_D\times d_H}$, the target query $k_y^i \in \mathbb{R}^{L_T\times d_H}$, and the target value $v_y^i \in \mathbb{R}^{L_T\times d_H}$ through fully connected layers, as shown in Equation (2).

$$\begin{cases} q_x^i = h_x W_q^i \\ k_y^i = h_y W_k^i \qquad i = 1, 2, \ldots, N_H, \\ v_y^i = h_y W_v^i \end{cases} \tag{2}$$

where N_H is the number of attention heads, and W_q^i, W_k^i, $W_v^i \in \mathbb{R}^{d_C\times d_H}$ represent the projection weights for the query, key, and value in the ith attention head, respectively. d_H is the output dimension of each attention head, and $d_C = d_H \times N_H$.

Meanwhile, the module uses the same weights to compute the target query q_y^i and the drug key-value vectors k_x^i and v_x^i, as shown

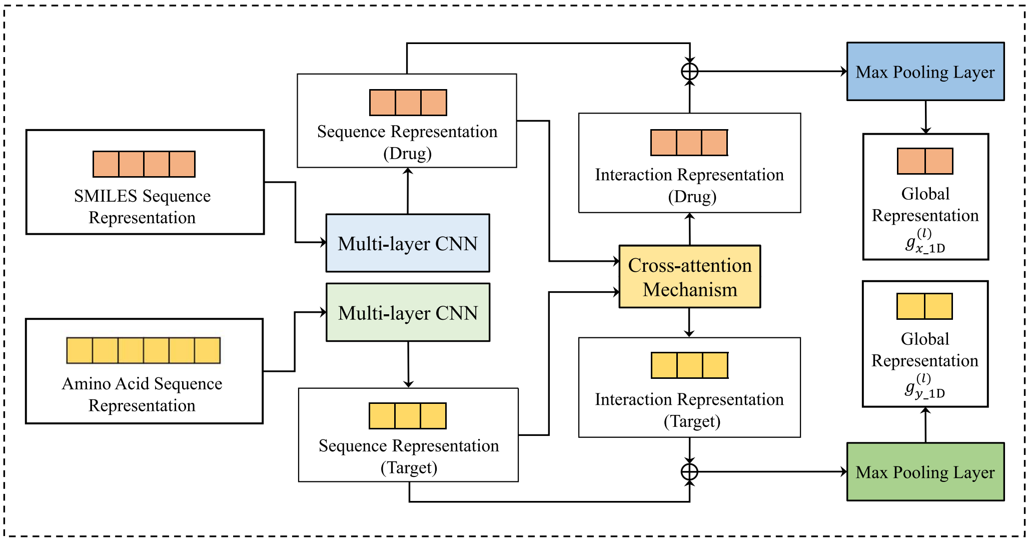
SMILES Sequence Representation
Sequence Representation (Drug)
Multi-layer CNN
Amino Acid Sequence Representation
Multi-layer CNN
Sequence Representation (Target)
Interaction Representation (Drug)
Cross-attention Mechanism
Interaction Representation (Target)
Max Pooling Layer
Global Representation $g_{x\_1D}^{(l)}$
Global Representation $g_{y\_1D}^{(l)}$
Max Pooling Layer

Figure 3. Overall processing flow of the MIF-1D encoding block, including multilayer 1D CNN for local dependencies, multi-head cross-attention for drug–target interactions, and max pooling to generate global representations for MIF decoding.

in Equation (3).

$$\begin{cases} q_y^i = h_y W_q^i \\ k_x^i = h_x W_k^i \qquad i = 1, 2, \ldots, N_H \\ v_x^i = h_x W_v^i \end{cases} \tag{3}$$

Subsequently, the attention module computes the feature matrices f_x^i and f_y^i of the drug and the target under the ith attention head using dot-product operations and the **softmax** function, as shown in Equation (4).

$$\begin{cases} f_x^i = \text{softmax}\left(\dfrac{q_x^i k_y^{iT}}{\sqrt{d_H}}\right) v_y^i = A_x^i v_y^i \\ f_y^i = \text{softmax}\left(\dfrac{q_y^i k_x^{iT}}{\sqrt{d_H}}\right) v_x^i = A_y^i v_x^i \end{cases} \quad i = 1, 2, \ldots, N_H, \tag{4}$$

where $A_x^i \in \mathbb{R}^{L_D \times L_T}$ and $A_y^i \in \mathbb{R}^{L_T \times L_D}$ are the normalized attention score matrices for the drug and the target, respectively. The final outputs satisfy $f_x^i \in \mathbb{R}^{L_D \times d_H}$ and $f_y^i \in \mathbb{R}^{L_T \times d_H}$.

After completing the multi-head operations, the attention module concatenates the output features from all attention heads and applies a linear transformation to generate the interaction representations of the drug and the target, denoted as h'_x and h'_y, as shown in Equation (5).

$$\begin{cases} h'_x = \text{concat}\left(f_x^1, f_x^2, \ldots, f_x^{N_H}\right) W_c \\ h'_y = \text{concat}\left(f_y^1, f_y^2, \ldots, f_y^{N_H}\right) W_c \end{cases}, \tag{5}$$

where $W_c \in \mathbb{R}^{d_C \times d_C}$ is the learnable weight matrix, $h'_x \in \mathbb{R}^{L_D \times d_C}$, $h'_y \in \mathbb{R}^{L_T \times d_C}$.

Through the above operations, the MIF-1D encoding block obtains the sequence representations of the drug and the target, denoted as h_x and h_y, respectively, as well as their interaction representations h'_x and h'_y. Denoting the current encoding block as the lth block, the integrated encoding representations of the drug and target are $h_x^{(l)} \in \mathbb{R}^{L_D \times d_C}$ and $h_y^{(l)} \in \mathbb{R}^{L_T \times d_C}$, respectively. The computation process is shown in Equation (6).

$$\begin{cases} h_x^{(l)} = 0.5 \cdot h_x + 0.5 \cdot h'_x, \\ h_y^{(l)} = 0.5 \cdot h_y + 0.5 \cdot h'_y. \end{cases} \tag{6}$$

Finally, the encoding block applies the **MaxPool** operation to obtain the global representations of the drug and the target, denoted as $g_{x\_1D}^{(l)}$ and $g_{y\_1D}^{(l)} \in \mathbb{R}^{1 \times d_C}$, respectively. The computation process is shown in Equation (7).

$$\begin{cases} g_{x\_1D}^{(l)} = \text{MaxPool}\left(h_x^{(l)}\right), \\ g_{y\_1D}^{(l)} = \text{MaxPool}\left(h_y^{(l)}\right). \end{cases} \tag{7}$$

Graph encoding module

The graph encoding module generates a 2D graph structure based on the 1D sequences of the drug and the target, followed by dual-view encoding. The overall process is shown in Fig. 4.

The graph encoding module generates a hierarchical molecular graph G_x from the SMILES sequence using RDKit [29] and the BRICS algorithm with refinement rules [30], comprising atom-, substructure-, and molecule-level nodes. Node features are defined in Supplementary Material Table S.1. For the target amino acid sequence, a 2D contact graph G_y is constructed using the ESM-2 model [31], which outputs a fully connected graph. To reduce over-smoothing, only edges with contact values greater than 0.5 are retained following Koh *et al.* [23]. Node features are defined in Supplementary Material Table S.2.

Subsequently, the graph encoding module uses multiple MIF-2D encoding blocks to learn drug and target representations with structural and interaction information. Residual connections are introduced between blocks to deepen the network and mitigate overfitting. Each MIF-2D encoding block is composed of structure-view layer, interaction-view layer and graph update layer, as shown in Fig. 5.

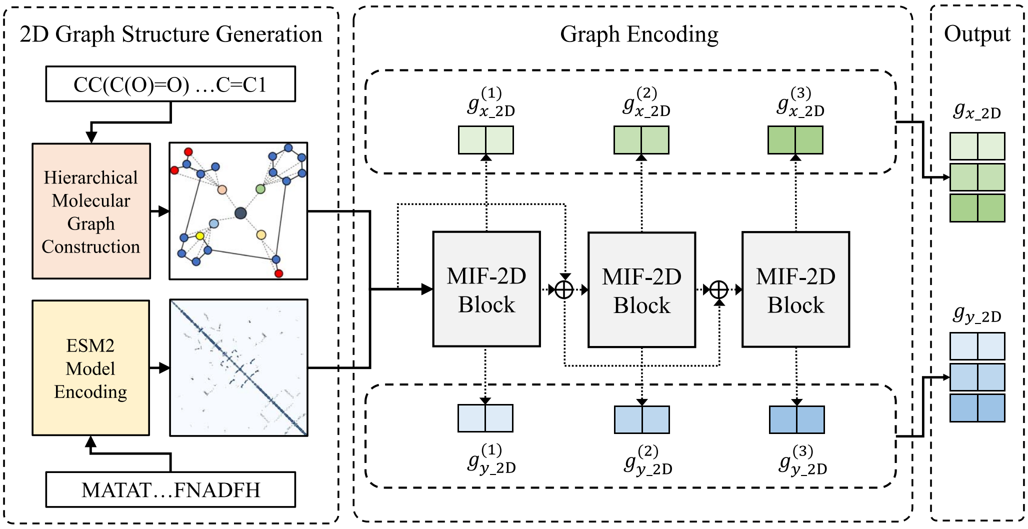

Figure 4. Structure of the MIF graph encoding module with 2D graph generation and MIF-2D encoding blocks for multi-depth global representations of drugs and targets.

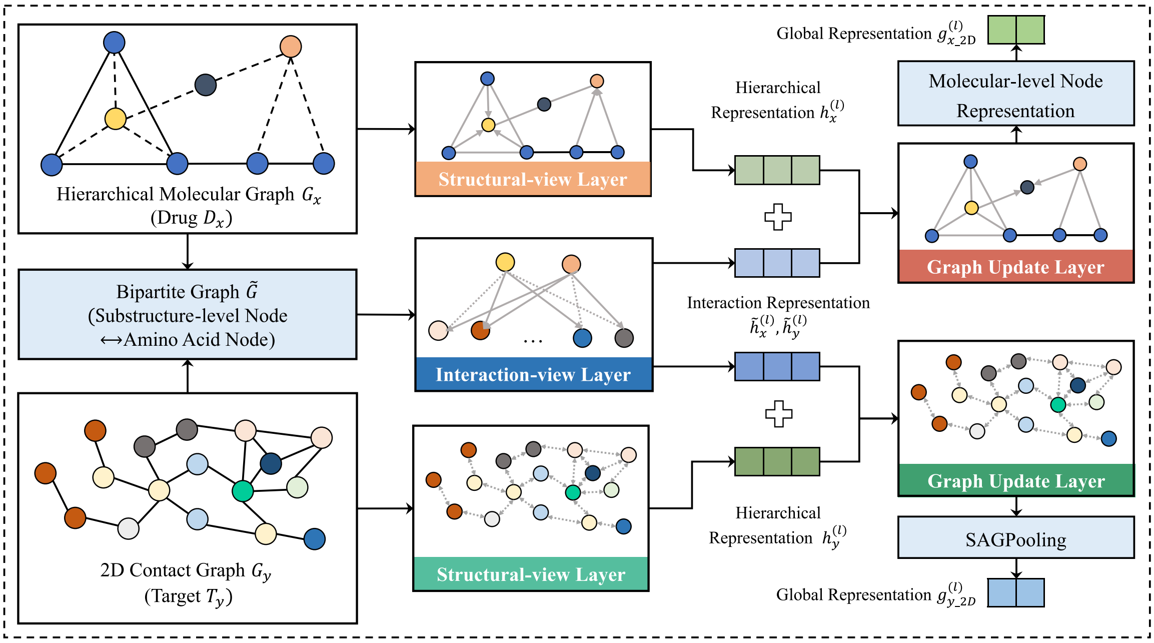
Hierarchical Molecular Graph G_x (Drug D_x)
Bipartite Graph $\tilde{G}$ (Substructure-level Node ↔Amino Acid Node)
2D Contact Graph G_y (Target T_y)
Structural-view Layer
Interaction-view Layer
Structural-view Layer
Global Representation $g_{x\_2D}^{(l)}$
Hierarchical Representation $h_x^{(l)}$
Molecular-level Node Representation
Graph Update Layer
Interaction Representation $\tilde{h}_x^{(l)}, \tilde{h}_y^{(l)}$
Graph Update Layer
Hierarchical Representation $h_y^{(l)}$
SAGPooling
Global Representation $g_{y\_2D}^{(l)}$

Figure 5. Structure of the MIF-2D encoding block, including structure-view, interaction-view, and graph update layers.

In the structure-view layer, two independent GATs are used to encode the drug's hierarchical graph G_x and the target's contact graph G_y, capturing their internal structural information. The representation of node i in G_x or G_y at the (l+1)th block is $h_i^{(l+1)}$, as computed in Equations (8) to (10).

$$h_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}_i\cup\{i\}} \alpha_{ij} W^{(l+1)} h_j^{(l)} + b^{(l+1)}\right) \tag{8}$$

$$\alpha_{ij} = \frac{\exp\left(\mathbf{LeakyReLU}(e_{ij})\right)}{\sum_{k\in\mathcal{N}_i\cup\{i\}} \exp\left(\mathbf{LeakyReLU}(e_{ik})\right)} \tag{9}$$

$$e_{ij} = a^{(l+1)\top}\left(W^{(l+1)} h_i^{(l)} + W^{(l+1)} h_j^{(l)} + W_e^{(l+1)} h_{ij}^{(l)}\right), \tag{10}$$

where $W^{(l+1)}$, $W_e^{(l+1)} \in \mathbb{R}^{d^{(l+1)}\times d^{(l)}}$, and $a^{(l+1)} \in \mathbb{R}^{d^{(l+1)}}$ are the trainable weights of GAT in the l+1th block, and $b^{(l+1)} \in \mathbb{R}^{d^{(l+1)}}$ is the bias term. **LeakyReLU**(·) [32] and $\sigma(\cdot)$ are activation functions. $d^{(l)}$ and $d^{(l+1)}$ are the input and output feature dimensions. $\mathcal{N}i$ is the neighbor set of node i, $h_j^{(l)}$ is the feature of node j, and αij is the attention coefficient between nodes i and j.

In the interaction-view layer, the MIF-2D encoding block constructs a bipartite graph $\widetilde{G} = V_s^{(x)} \times V^{(y)}$ between drug substructure-level nodes and all target nodes, and applies a single GAT to capture their interactions. Following prior work [21, 28], which shows that DTI is often triggered by key substructures and peptides, we randomly drop edges in $\widetilde{G}$ to encourage the model to focus on critical regions. The interaction representation $\tilde{h}_i^{(l+1)}$ is computed similarly to Equations (8) to (10).

For the atomic and molecular nodes in the hierarchical molecular graph G_x, the MIF-2D encoding block performs nonlinear encoding operations to ensure that all nodes in G_x have the same feature dimension. This operation shares weights with GAT in the interactive-view layer, and the calculation process is shown in equation (11).

$$\tilde{h}_k^{(l+1)} = \sigma\left(\widetilde{W}^{(l+1)} h_k^{(l)} + \tilde{b}^{(l+1)}\right), \tag{11}$$

where $\tilde{h}_k^{(l+1)}$ is the interaction representation of node k. $\widetilde{W}^{(l+1)}$ and $\tilde{b}^{(l+1)}$ are learnable parameters.

In the graph update layer, the MIF-2D encoding block first concatenates the structural and interaction representations of each node in G_x or G_y. It then updates them respectively using two independent GATs, shown in equation (12).

$$h_i^{(l+1)} = \mathrm{ELU}\left(\mathrm{LayerNorm}\left(h_i^{(l+1)} \| \tilde{h}_i^{(l+1)}\right)\right) \tag{12}$$

Finally, the MIF-2D encoding block extracts the updated representation of the molecular-level node v_{x_mol} as the global representation of the drug molecule $g_x^{(l+1)}$ and the global representation of the target $g_y^{(l+1)}$ via SAGPooling [33], as shown in Equations (13) and (14).

$$g_{x\_2D}^{(l+1)} = h_{x\_mol}^{(l+1)} \tag{13}$$

$$g_{y\_2D}^{(l+1)} = \mathrm{SAGPooling}\left(G_y\right) \tag{14}$$

MIF decoding module

The sequence and graph encoding modules extract multilevel global representations. The MIF decoding module then fuses these via co-attention [34, 35], matrix multiplication, and fully connected layers for DTI prediction. Specifically, it stacks the drug and target global representations into matrices g_x and g_y, where N_1 and N_2 are the numbers of MIF-1D and MIF-2D encoding blocks, and d is the representation dimension, as shown in Equation (15).

$$\begin{cases} g_x = \mathrm{Stack}\left(g_{x\_1D}^{(1)}, \ldots, g_{x\_1D}^{(N_1)},\ g_{x\_2D}^{(1)}, \ldots, g_{x\_2D}^{(N_2)}\right) \\ g_y = \mathrm{Stack}\left(g_{y\_1D}^{(1)}, \ldots, g_{y\_1D}^{(N_1)},\ g_{y\_2D}^{(1)}, \ldots, g_{y\_2D}^{(N_2)}\right) \end{cases}, \tag{15}$$

where **Stack** is the stacking operation, and $g_x, g_y \in \mathbb{R}^{(N_1+N_2)\times d}$.

Then, the MIF decoding module uses the co-attention mechanism to iteratively compute multimodal fusion coefficients, resulting in the fusion coefficient matrix M, as shown in Equation (16). Based on this, the module performs matrix multiplication among the fusion coefficient matrix M, the global representation matrix of the drug g_x, and the global representation matrix of the target g_y to obtain the interaction score matrix S, as shown in Equation (17).

$$M_{ij} = \alpha^\top \tanh\left(W_x g_x^{(i)} + W_y g_y^{(j)}\right) \tag{16}$$

$$S = M \circ \left(g_x g_y^\top\right), \tag{17}$$

where $g_x^{(i)}, g_y^{(j)} \in \mathbb{R}^d$ are the ith row of matrix g_x and the jth row of matrix g_y, respectively. $M_{ij} \in \mathbb{R}$ is the element in the ith row and jth column of matrix M, where $1 \le i, j \le (N_1 + N_2)$. $W_x, W_y \in \mathbb{R}^{d\times d}$ are learnable weight matrices, $\alpha \in \mathbb{R}^d$ is a learnable weight vector, and $\alpha^\top \in \mathbb{R}^{1\times d}$. The symbol $\circ$ denotes element-wise multiplication, and $M \in \mathbb{R}^{(N_1+N_2)\times(N_1+N_2)}$.

Finally, the MIF decoding module flattens the score matrix S and computes the DTI probability p using a fully connected layer and the **Sigmoid** activation function, as shown in Equation (18).

$$p = \mathrm{Sigmoid}\left(W \cdot \mathrm{Flatten}(S) + b\right), \tag{18}$$

where $\mathrm{Flatten}(\cdot)$ is the matrix flattening operation, $W \in \mathbb{R}^{1\times[(N_1+N_2)\times(N_1+N_2)]}$ is a learnable weight vector, and $b \in \mathbb{R}$ is a learnable bias term.

Ensemble model

During five-fold cross-validation, the dataset is first split into training and test sets at an 8:2 ratio. Then, 20% of the training set is iteratively used as a validation set, and the rest for training, resulting in five MIF–DTI models. As each model is trained on different subsets and has different weights, they are integrated with equal weights to construct the ensemble model MIF–DTI-B. During prediction, each model computes the DTI probability, and the average is used as the final result. This approach fully utilizes the training data and improves model stability and robustness.

Loss function

The goal of this study is to predict whether there is an interaction between drugs and targets, which is essentially a binary classification problem. Therefore, model training adopts a binary cross-entropy loss function, and its calculation method is shown in the formula (19).

$$\mathrm{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log p_i + (1-y_i)\log(1-p_i)\right), \tag{19}$$

where N is the total number of samples in the training set, $p_i \in [0, 1]$ is the probability value of the ith sample output by the model, and $y_i \in 0, 1$ is the true label of the ith sample.

Experiment

Dataset

This study used three free and open public data sets in model training and evaluation: DrugBank [36], BioSNAP [37], and Davis [38]. DrugBank and BioSNAP are both comprehensive data sets, including drug–drug interactions and DTIs. This study only utilizes data in which DTI prediction tasks are related. The Davis dataset mainly records DTIs associated with kinases.

Regarding dataset size, BioSNAP includes 4510 drugs and 2181 targets, with 13 830 positive and 13 634 negative samples. Davis contains 68 drugs, 379 targets, 7320 positive, and 18 452 negative samples. DrugBank has 6655 drugs and 4294 targets but only positive samples. Following HyperAttentionDTI [39], we exclude small or unparsable drugs using RDKit toolkit, then randomly generate negative samples from valid drug–target pairs without known interactions on DrugBank before splitting the data into training, validation, and testing sets. The number of negative samples matches the positive ones, yielding 17 511 positive and 17 511 negative samples after processing.

Evaluation metrics

During the model evaluation phase, this study follows the evaluation protocol established in previous works [20, 39], and uses five metrics to compare different models: accuracy, precision, recall, the area under the receiver operating characteristic curve (AUROC), and the area under the precision-recall curve (AUPR). Among these, accuracy and AUROC are used to assess the overall predictive performance of the model, while precision, recall, and AUPR are used to evaluate the model's ability to identify positive samples.

Parameter setting

MIF–DTI consists of a sequence encoder, a graph encoder, and an MIF decoder. Hyperparameters are selected via random search.

The sequence encoder uses three MIF-1D encoding blocks, each with two CNN submodules (kernel sizes: 4,6,8 for drugs; 4,8,12 for targets) and one attention module with five heads. All outputs are 200D.

The graph encoder includes three MIF-2D encoding blocks, each containing two structure-view layers, one interaction-view layer, and two graph update layers, all implemented as single-layer, two-head GATs. Structure/interactions output 100D features; update layers output 200D.

Training uses batch size 64 and CyclicLR [40] with base LR and weight decay of 1×10^{-4}, and max LR multiplier of 10.

To ensure a fair comparison, we established a unified evaluation framework while respecting the model-specific training configurations of each baseline. Our unified framework mandated that all models were trained on the same data splits and employed an early stopping mechanism with a patience of 50 epochs. The model version that achieved the best performance on the validation set was selected for the final evaluation. Within this framework, for each baseline, we adhered to its core training configurations as suggested in the original publication, such as the choice of optimizer (e.g. Adam, SGD) and its corresponding learning rate.

Baselines

- HyperAttentionDTI [39] encodes drug and target sequences using CNNs and captures their interaction information through a hyper attention mechanism. It is one of the efficient sequence-based models in recent years.
- MCANet [20] encodes SMILES and amino acid sequences using 1D CNNs, and employs a cross-attention mechanism to capture bidirectional influences between drugs and targets. Its ensemble version, MCANet-B, was also proposed in the original publication. For a direct and fair comparison, the MIF–DTI-B model proposed in this study adopts the identical ensemble strategy.
- BINDTI [24] uses a GNN and an ACMix model to encode molecular graphs of drugs and sequences of targets, respectively. A bidirectional intent network is then applied to integrate both features, followed by a fully connected layer to predict DTIs.
- PSICHIC [23] decomposes drug molecules into multiple functional groups using a junction tree algorithm, generates target contact graphs using the ESM-2 model, and applies graph clustering to divide them into functional regions. A GNN is used to learn both intra-molecular and inter-molecular information.

Results

Results on benchmark datasets

Tables 1, 2, and 3, respectively, present the performance of MIF–DTI and baselines on the DrugBank, BioSNAP, and Davis datasets. A detailed comparison of model parameters and inference runtime is provided in the Supplementary Material Table S.3.

As shown in Table 1, MIF–DTI outperforms existing state-of-the-art models across multiple metrics on DrugBank. Compared with the best sequence model MCANet, it improves accuracy, AUROC, and AUPR by 2.86%, 3.06%, and 2.50%, respectively. Compared with the best graph model PSICHIC, the improvements are 1.87%, 2.30%, and 3.01%. Its ensemble version MIF–DTI-B enhances performance, surpassing MIF–DTI by 2.33%, 1.98%, and 2.07%, and outperforming the strongest ensemble baseline MCANet-B by 2.31%, 2.64%, and 2.18%.

On BioSNAP (Table 2), MIF–DTI and MIF–DTI-B also attain state-of-the-art performance. MIF–DTI attains results equivalent to the prior leading approach MCANet-B. MIF–DTI-B surpasses MIF–DTI with improvements of 2.25% in accuracy, 3.43% in precision, and 2.07% in AUPR.

On Davis (Table 3), MIF–DTI consistently outperforms both sequence- and graph-based models. It surpasses MCANet by 0.37%, 0.25%, and 0.40%, and surpasses PSICHIC by 1.22%, 1.59%, and 1.85%, respectively. These results demonstrate that both MIF–DTI and MIF–DTI-B consistently achieve superior performance across different datasets, validating the effectiveness of the proposed model design.

Results on cross-dataset validation

Within-dataset cross-validation under random split is an easier task that holds diminished practical significance. Therefore, we designed a more stringent cross-dataset validation to evaluate the genuine generalization performance of the MIF–DTI model. In this setting, we simulated the scenario of applying the model to novel biological and chemical spaces: the model was trained on the union of the DrugBank and BioSNAP datasets and then tested for performance on the unseen Davis dataset.

The results of the cross-dataset validation are shown in Table 4. As anticipated, all models exhibited a notable performance decline compared with the within-dataset validation, owing to the inherent difficulty of the task and the distribution shift between datasets. Despite this challenge, MIF–DTI-B emerged as the top-performing model, achieving the highest scores in four key metrics: Accuracy (66.68%), Precision (33.32%), AUROC (53.08%), and AUPR (30.90%). Notably, its relatively low recall (17.27%) stems from a conservative prediction strategy that trades recall for the highest precision, an advantage that effectively reduces costly experimental validation in practical applications. Collectively, these results provide compelling evidence of our proposed model's strong generalization capabilities and application potential, proving its ability to address real-world drug development difficulties.

Ablation experiments

We conducted ablation experiments on the DrugBank and BioSNAP datasets to evaluate the impact of key modules in MIF–DTI. Specifically, we design three variants:

- wo-1D-encoder, which removes the sequence encoding module to verify its contribution;
- wo-2D-encoder, which removes the graph encoding module to assess its importance;
- with-attention, which replaces the co-attention and interaction score matrix in the decoding module with cross-attention and max pooling, validating the effectiveness of the original fusion mechanism.

Table 1. Results of the proposed models and baselines on the DrugBank dataset (%)

Model	Accuracy	Precision	Recall	AUROC	AUPR
HyperAttentionDTI	81.00	79.90	82.90	88.90	88.44
MCANet	82.60	82.42	82.74	89.71	90.45
MCANet-B	85.48	85.84	84.83	92.11	92.84
BIN-DTI	80.43	79.91	81.33	89.71	90.45
PSICHIC	83.59	83.05	83.33	90.47	89.94
MIF–DTI	85.46	85.14	85.60	92.77	92.95
MIF–DTI-B	**87.79**	**87.72**	**87.59**	**94.75**	**95.02**

Table 2. Results of the proposed models and baselines on the BioSNAP dataset (%)

Model	Accuracy	Precision	Recall	AUROC	AUPR
HyperAttentionDTI	84.20	82.80	86.61	91.10	91.89
MCANet	84.27	83.28	85.53	91.38	91.78
MCANet-B	86.55	86.11	86.94	93.41	93.69
BIN-DTI	81.39	80.14	83.06	88.25	87.84
PSICHIC	84.14	83.72	84.45	91.21	90.60
MIF–DTI	86.95	87.28	86.21	**93.96**	94.32
MIF–DTI-B	**88.80**	**89.54**	**87.67**	93.39	**95.76**

Table 3. Results of the proposed models and baselines on the Davis dataset (%)

Model	Accuracy	Precision	Recall	AUROC	AUPR
HyperAttentionDTI	86.36	76.02	76.19	92.10	84.36
MCANet	87.05	77.54	76.67	92.56	85.11
MCANet-B	**89.27**	82.65	78.76	**94.87**	**89.43**
BIN-DTI	79.77	60.32	**84.02**	88.44	79.24
PSICHIC	86.20	78.91	71.70	91.22	83.66
MIF–DTI	87.42	78.47	76.66	92.81	85.51
MIF–DTI-B	89.21	**82.80**	78.28	94.53	89.04

Table 4. Cross-dataset performance comparison of the proposed models and baselines(%)

Model	Accuracy	Precision	Recall	AUROC	AUPR
HyperAttentionDTI	58.78	30.60	**35.31**	52.03	29.80
MCANet	62.22	28.24	23.27	49.63	28.58
MCANet-B	65.81	28.65	13.66	28.38	49.79
BIN-DTI	63.18	32.34	27.14	52.05	30.04
PSICHIC	64.10	32.30	24.07	52.67	30.63
MIF–DTI	64.33	31.93	22.29	52.46	30.44
MIF–DTI-B	**66.68**	**33.32**	17.27	**53.08**	**30.90**

As shown in Table 5 and 6, the overall performance of MIF–DTI surpasses all its variants, indicating that the sequence encoding module, graph encoding module, and MIF decoding module all effectively contribute to improving the model's predictive performance. The specific analysis of each variant's performance are as follows:

(1) Both sequence and graph encoding modules are important for MIF–DTI. As shown in Tables 5 and 6, removing the sequence encoder (wo-1D-encoder) causes accuracy, AUROC, and AUPR to drop by 2.20%, 2.20%, and 2.59% on DrugBank, and 2.90%, 2.51%, and 3.15% on BioSNAP. while removing the graph encoder (wo-2D-encoder) results in larger declines of 6.06%, 5.52%, and 5.83% on DrugBank, and 5.08%, 4.48%, and 4.68% on BioSNAP. This indicates both modules contribute to DTI prediction, with the graph encoder being more critical. Figure 6 further shows that MIF–DTI requires fewer training epochs and achieves better performance than single-modality models, confirming the benefit of multimodal fusion.

(2) The co-attention mechanism and interaction score matrix are crucial in multimodal fusion. Tables 5 and 6 show that replacing them with traditional attention and pooling (with-attention) reduces accuracy, AUROC, and AUPR by 1.76%, 1.69%, and 1.75% on DrugBank, and 3.33%, 2.59%, and 2.82% on BioSNAP. This demonstrates the effectiveness of co-attention in integrating drug and target representations across modalities and depths.

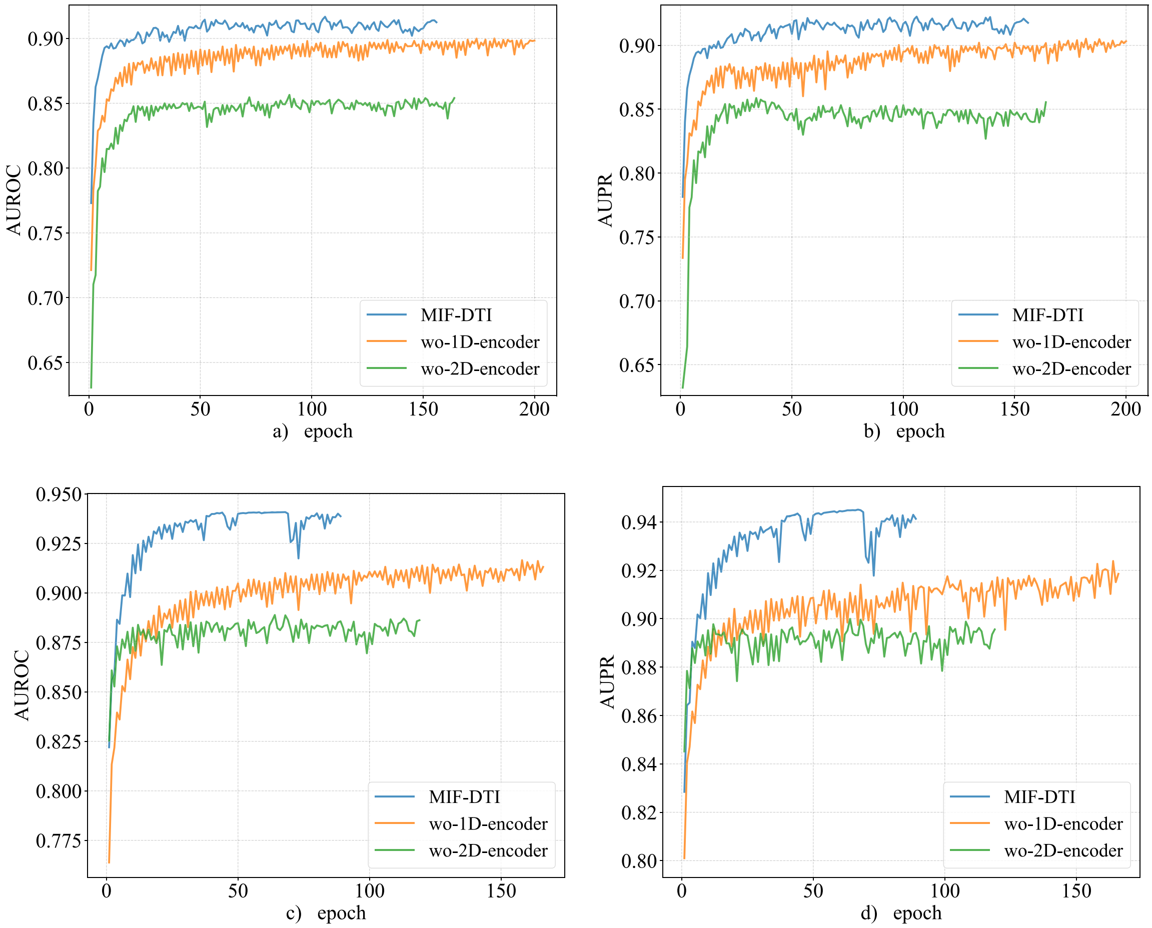

Figure 6. AUROC and AUPR curves of MIF–DTI and its partial variants during training, with (a, b) showing comparisons on the DrugBank dataset and (c, d) on the BioSNAP dataset.

Case Study

To further validate the reliability of MIF–DTI, we analyzed the prediction accuracy for selected drugs and targets. First, we randomly selected two drugs and their corresponding targets from the DrugBank dataset, completely removed them from the training data, and tested the model's performance on predicting interactions for these unseen entities. The remaining data were used to train the model, which was then evaluated on the test set. As shown in Table 7, MIF–DTI achieved an accuracy of 91.7% in predicting interactions between Ofloxacin, SNX-5422, and their related targets. Similarly, we randomly selected two targets, MADH and MdfA, along with their associated drugs to form another test set. As shown in Table 8, the model again achieved a prediction accuracy exceeding 90% in this scenario.

Discussion

This study proposes more accurate and robust DTI prediction models, namely MIF–DTI and MIF–DTI-B. This section provides an in-depth discussion and analysis of the technical contributions of both models.

To explore the relationship between the 1D sequence and 2D graph, we visualized the adjacency matrix for target P35228 (Supplementary Figure S.1). In the matrix, adjacent residues (spatial distance <6 Å) cluster near the diagonal, indicating that sequentially close amino acids are often spatially proximal. This suggests the 1D sequence efficiently reflects local structure and is efficient to process. In contrast, the 2D graph derived from the adjacency matrix encodes richer global spatial features at the cost of higher complexity. The complementarity between the efficiency of the 1D sequence and the richness of the 2D graph validates our multimodal fusion strategy and inspires the future integration of more structural and functional modalities.

We analyzed model performance across datasets. As shown in Table 1 and Table 2, the proposed MIF–DTI and MIF–DTI-B outperform other models on DrugBank and BioSNAP datasets. However, on the Davis dataset, MIF–DTI shows only slight improvement over the sequence model MCANet, and MIF–DTI-B performs similarly to MCANet-B. In contrast, MCANet significantly outperforms PSICHIC and other graph-based models.

This may be due to the large number of nodes and sparse adjacency relationships in 2D graph, which demands larger datasets and stronger graph encoders. Davis includes far fewer drugs and targets than DrugBank and BioSNAP. In this low-sample scenario, the sequence encoding module in MIF–DTI can still learn effectively, while the graph encoding module may be undertrained, resulting in lower performance.

This hypothesis is supported by our ablation experiments, which confirm the graph module's dominant contribution on

Table 5. Ablation experiments of MIF–DTI and its variants on the DrugBank dataset (%)

Model	Accuracy	Precision	Recall	AUROC	AUPR
wo-1D-encoder	83.26	82.22	84.52	90.57	90.36
wo-2D-encoder	79.40	80.27	77.69	87.25	87.12
with-attention	83.70	83.40	83.87	91.08	91.20
MIF–DTI	**85.46**	**85.14**	**85.60**	**92.77**	**92.95**

Table 6. Ablation experiments of MIF–DTI and its variants on the BioSNAP dataset (%)

Model	Accuracy	Precision	Recall	AUROC	AUPR
wo-1D-encoder	84.05	83.30	84.82	91.45	91.17
wo-2D-encoder	81.87	80.38	83.95	89.48	89.64
with-attention	83.62	82.90	84.38	91.37	91.50
MIF–DTI	**86.95**	**87.28**	**86.21**	**93.96**	**94.32**

Table 7. Prediction results for drugs Ofloxacin and SNX-5422

Drug	Target	True label	Predicted label
DB01165-Ofloxacin	A0A024R811	False	False
	P00918	False	True
	Q14003	False	False
	P43700	True	True
	P0C0T5	False	False
	Q13219	False	False
	P11388	True	True
	P43702	True	True
DB06070-SNX-5422	P07900	True	True
	P02775	False	False
	P08238	True	True
	Q9HBA0	False	False
Accuracy			91.7%

Table 8. Prediction results for targets MADH and MdfA

Target	Drug	True label	Predicted label
P29894-MADH	DB03780	False	False
	DB03905	False	False
	DB07764	False	False
	DB07781	False	False
	DB07795	False	True
	DB08646	True	True
C9EH48-MdfA	DB00759	True	True
	DB02030	False	False
	DB07169	False	False
	DB07610	False	False
	DB08314	False	False
Accuracy			90.9%

larger datasets and show that the full model significantly outperforms any single-modality variant, underscoring the importance of the fusion strategy. This is the core of our architectural innovation. Some models, like 3DProtDTA [25], rely on simple feature concatenation, a passive method that leaves the task of interaction identification to a downstream classifier. More advanced models, such as BINDTI [24], adopt cross-attention for representation enhancement. In contrast, our MIF–DTI combines dual-view representation learning with a decision-focused co-attention module, enabling a more direct and powerful fusion by computing a final interaction score matrix from the rich and multi-depth features of both modalities.

Conclusion

This study proposes a DTI prediction method MIF–DTI based on multimodal information fusion. This method encodes the SMILES sequences of drugs and the amino acid sequences of targets through the sequence encoding module to extract their 1D sequence features. It then performs dual-view representation encoding on the hierarchical molecular graphs of drugs and the contact graphs of targets via the graph encoding module to capture their 2D topological structure information. Finally, MIF–DTI uses a co-attention mechanism to calculate the multimodal fusion coefficients and obtain the interaction score matrix through matrix operations, achieving the fusion of different modalities and global representations at different depths and improving the accuracy of DTI prediction. Based on MIF–DTI, this study further proposes an ensemble version, MIF–DTI-B by incorporating a cross-validation training strategy. Experimental results show that both MIF–DTI and MIF–DTI-B achieve better performance on three datasets and exhibit strong generalization in cross-dataset validation. They demonstrate higher performance upper bounds when data are sufficient and maintain favorable lower bounds under limited data conditions. Ablation experiments further confirm the effectiveness of each module in MIF–DTI. Through multimodal information fusion and model ensemble, this study provides a more accurate DTI prediction method, offering a reliable computational model for downstream tasks such as drug discovery and drug repurposing.

Key Points

- MIF–DTI introduces a multimodal information fusion strategy that integrates 1D sequences with 2D molecular and contact graphs to capture both sequence-level and structural features.
- A dual-view representation mechanism enhances drug–target interaction prediction.
- The ensemble model MIF–DTI-B further improves performance and robustness.
- Ablation studies confirm the critical contribution of each module, particularly the 2D graph encoder.

Author contributions

Jiehong Shan (Conceptualization, Methodology, Investigation, Formal analysis, Writing-original draft, Writing-review & editing), Jinchen Sun (Conceptualization, Methodology, Investigation, Formal analysis, Writing-original draft), and Haoran Zheng (Conceptualization, Methodology, Project administration, Supervision, Writing-review & editing)

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by the National Key Technologies R&D Program of China [2017YFA0505502]; and the Strategic Priority Research Program of the Chinese Academy of Sciences [XDB38000000].

Data availability

The codes and datasets are available online at https://git-hub.com/sjh126/MIF-DTI.

References

1. Iqbal AB, Shah IA, Injila AA. *et al.* A review of deep learning algorithms for modeling drug interactions. *Multimedia Syst* 2024;**30**:124. https://doi.org/10.1007/s00530-024-01325-9
2. Paul SM, Mytelka DS, Dunwiddie CT. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 2010;**9**:203–14. https://doi.org/10.1038/nrd3078
3. Zeng X, Wang F, Luo Y. *et al.* Deep generative molecular design reshapes drug discovery. *Cell Rep Med* 2022;**3**:100794. https://doi.org/10.1016/j.xcrm.2022.100794
4. Zeng X, Zhu S, Liu X. *et al.* deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics.* 2019;**35**:5191–8. https://doi.org/10.1093/bioinformatics/btz418
5. Zeng X, Zhu S, Lu W. *et al.* Target identification among known drugs by deep learning from heterogeneous networks. *Chem Sci* 2020;**11**:1775–97. https://doi.org/10.1039/C9SC04336E
6. Bai P, Miljković F, John B. *et al.* Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nat Mach Intell.* 2023;**5**:126–36. https://doi.org/10.1038/s42256-022-00605-1
7. Napolitano F, Zhao Y, Moreira VM. *et al.* Drug repositioning: a machine-learning approach through data integration. *J Chem* 2013;**5**:1–9. https://doi.org/10.1186/1758-2946-5-30
8. Keiser MJ, Roth BL, Armbruster BN. *et al.* Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 2007;**25**:197–206. https://doi.org/10.1038/nbt1284
9. Gentile F, Agrawal V, Hsing M. *et al.* Deep docking: a deep learning platform for augmentation of structure based drug discovery. *ACS Cent Sci* 2020;**6**:939–49. https://doi.org/10.1021/acscentsci.0c00229
10. Milon TI, Wang Y, Fontenot RL. *et al.* Development of a novel representation of drug 3D structures and enhancement of the TSR-based method for probing drug and target interactions. *Comput Biol Chem* 2024;**112**:108117. https://doi.org/10.1016/j.compbiolchem.2024.108117
11. Shi H, Liu S, Chen J. *et al.* Predicting drug–target interactions using lasso with random forest based on evolutionary information and chemical structure. *Genomics.* 2019;**111**:1839–52. https://doi.org/10.1016/j.ygeno.2018.12.007
12. Zhan X, You ZH, Cai J. *et al.* Prediction of drug–target interactions by ensemble learning method from protein sequence and drug fingerprint. *IEEE Access* 2020;**8**:185465–76. https://doi.org/10.1109/ACCESS.2020.3026479
13. Ahn S, Lee SE, Kim M. Random-forest model for drug–target interaction prediction via Kullback–Leibler divergence. *J Chem* 2022;**14**:67.
14. Rifaioglu AS, Nalbat E, Atalay V. *et al.* DEEPScreen: high performance drug–target interaction prediction with convolutional neural networks using 2-D structural compound representations. *Chem Sci* 2020;**11**:2531–57. https://doi.org/10.1039/C9SC03414E
15. Zeng X, Xiang H, Yu L. *et al.* Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nat Mach Intell* 2022;**4**:1004–16. https://doi.org/10.1038/s42256-022-00557-6
16. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics.* 2018;**34**:i821–9. https://doi.org/10.1093/bioinformatics/bty593
17. Lee I, Keum J, Nam H. DeepConv-DTI: prediction of drug–target interactions via deep learning with convolution on protein sequences. *PLoS Comput Biol* 2019;**15**:e1007129. https://doi.org/10.1371/journal.pcbi.1007129

18. Vaswani A, Shazeer N, Parmar N. *et al.* Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds.), *Advances in Neural Information Processing Systems*, Vol **30**. Red Hook, NY, USA: Curran Associates, Inc., 2017, 6000–10.
19. Huang K, Xiao C, Glass LM. *et al.* MolTrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics.* 2021;**37**:830–6. https://doi.org/10.1093/bioinformatics/btaa880
20. Bian J, Zhang X, Zhang X. *et al.* MCANet: shared-weight-based MultiheadCrossAttention network for drug–target interaction prediction. *Brief Bioinform* 2023;**24**:bbad082.
21. Zhang Q, Zuo L, Ren Y. *et al.* FMCA-DTI: a fragment-oriented method based on a multihead cross attention mechanism to improve drug–target interaction prediction. *Bioinformatics.* 2024;**40**:btae347.
22. Li F, Zhang Z, Guan J. *et al.* Effective drug–target interaction prediction with mutual interaction neural network. *Bioinformatics.* 2022;**38**:3582–9. https://doi.org/10.1093/bioinformatics/btac377
23. Koh HY, Nguyen AT, Pan S. *et al.* Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data. *Nat Mach Intell* 2024;**6**:673–87. https://doi.org/10.1038/s42256-024-00847-1
24. Peng L, Liu X, Yang L. *et al.* BINDTI: a bi-directional intention network for drug–target interaction identification based on attention mechanisms. *IEEE J Biomed Health Inform* 2024;**29**:1602–12.
25. Voitsitskyi T, Stratiichuk R, Koleiev I. *et al.* 3DProtDTA: a deep learning model for drug–target affinity prediction based on residue-level protein graphs. *RSC Adv* 2023;**13**:10261–72. https://doi.org/10.1039/D3RA00281K
26. Wu H, Liu J, Jiang T. *et al.* AttentionMGT-DTA: a multi-modal drug–target affinity prediction using graph transformer and attention mechanism. *Neural Netw* 2024;**169**:623–36. https://doi.org/10.1016/j.neunet.2023.11.018
27. Dou L, Zhang Z, Qian Y. *et al.* BCM-DTI: a fragment-oriented method for drug–target interaction prediction using deep learning. *Comput Biol Chem* 2023;**104**:107844. https://doi.org/10.1016/j.compbiolchem.2023.107844
28. Zheng S, Li Y, Chen S. *et al.* Predicting drug–protein interaction using quasi-visual question answering system. *Nat Mach Intell.* 2020;**2**:134–40. https://doi.org/10.1038/s42256-020-0152-y
29. Landrum G. *et al.* RDKit: a software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* 2013;**8**:5281.
30. Sun J, Zheng H. HDN-DDI: a novel framework for predicting drug–drug interactions using hierarchical molecular graphs and enhanced dual-view representation learning. *BMC Bioinformatics* 2025;**26**:28. https://doi.org/10.1186/s12859-025-06052-0
31. Lin Z, Akin H, Rao R. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science.* 2023;**379**:1123–30. https://doi.org/10.1126/science.ade2574
32. Maas AL, Hannun AY, Ng AY. *et al.* Rectifier nonlinearities improve neural network acoustic models. *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, Vol **30**. Atlanta, GA, USA, 2013, 3.
33. Lee J, Lee I, Kang J. In: International conference on machine learning. Self-Attention Graph Pooling In: Chaudhuri K, Salakhutdinov R (eds.), *Proceedings of the 36th International Conference on Machine Learning*, Vol **97**. Brooklyn, NY, USA: PMLR, 2019, 3734–43.
34. Nyamabo AK, Yu H, Shi JY. SSI–DDI: substructure–substructure interactions for drug–drug interaction prediction. *Brief Bioinform* 2021;**22**:bbab133.
35. Li Z, Zhu S, Shao B. *et al.* DSN-DDI: an accurate and generalized framework for drug–drug interaction prediction by dual-view representation learning. *Brief Bioinform* 2023;**24**:bbac597.
36. Wishart DS, Feunang YD, Guo AC. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82. https://doi.org/10.1093/nar/gkx1037
37. Leskovec J, Sosič R. Snap: general-purpose network analysis and graph-mining library. *ACM Trans Intell Syst Technol* 2016; **8**:1–20.
38. Davis MI, Hunt JP, Herrgard S. *et al.* Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol* 2011;**29**:1046–51. https://doi.org/10.1038/nbt.1990
39. 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**:655–62. https://doi.org/10.1093/bioinformatics/btab715
40. Smith LN. Cyclical learning rates for training neural networks. In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE, 2017, 464–72.

Briefings in Bioinformatics, 2025, **26(5)**, bbaf474
https://doi.org/10.1093/bib/bbaf474
Problem Solving Protocol