MotifGT-DTI: Pivotal Motif-Based Graph Transformer Model Improves Drug–Target Interaction Prediction

Wen Tian, Min Zeng, *Member, IEEE*, Jianxin Wang, *Senior Member, IEEE*, and Chengqian Lu

***Abstract*—Predicting drug–target interactions (DTIs) plays an essential role in drug discovery and drug repurposing. Although significant performance improvements have been achieved in DTI prediction, existing methods have not fully explored the properties of protein and drug molecular structure to make results interpretable. In this study, we propose MotifGT-DTI, a novel motif-based model with a graph transformer (GT) for DTI prediction. Specifically, MotifGT-DTI captures complex molecular patterns of drug molecular graph motifs and protein 3-D pocket subgraphs with GT. To attain protein characteristics more comprehensively, MotifGT-DTI fuses 1-D sequence and 3-D structure features with cross-attention from two views. Then, the structural-level association patterns of drug molecules and proteins are connected via a bilinear attention network. Experimental results show that MotifGT-DTI achieves the best accuracy compared to state-of-the-art baselines on four public datasets. In the three cold-start scenarios, the prediction results provided by our method are competitive in accuracy, generalization ability, and stability, highlighting its promising potential for practical applications. Furthermore, the visualization study demonstrates that MotifGT-DTI finds functional molecular motifs and provides interpretability for predicted results. The datasets and codes are publicly available at https://github.com/Dimpleney/MotifGT-DTI**

***Index Terms*—Bilinear attention network (BAN), drug–target interaction (DTI), graph transformer (GT), molecular graph motif, protein pockets.**

I. Introduction

DRUG–TARGET interaction (DTI) refers to the complex molecular process in which drug molecules bind to specific biological targets, such as proteins. Identifying DTI pairs is essential in drug discovery, as it accelerates the development process and reveals promising therapeutic targets. Traditional biochemical experiments for detecting DTI are costly and time-consuming, which hinders application in large-scale data [1]. In contrast, computational methods identify potential DTI pairs faster to guide the in vitro validation. There are three main classes of computational methods: docking-based, ligand-based, and machine-learning-based. Docking-based methods use the target proteins' 3-D structures to simulate drug–target binding by predicting binding poses and affinity through geometric complementarity and energy calculations. Molecular docking frameworks employ force fields to position ligands into binding sites and estimate binding energies [2]. Molecular dynamics simulations refine poses and calculate free energies [3]. Pharmacophore approaches identify key interaction features to guide ligand placement [4]. Although these methods provide atomic-level insights, they are constrained by the requirement for known 3-D structures and the limited accuracy of scoring functions for affinity prediction [5]. Ligand-based methods operate on the principle that chemically similar ligands tend to bind similar targets, leveraging existing pharmacological data [6]. Quantitative structure-activity relationship frameworks correlate molecular descriptors with biological activity using statistical techniques [7]. Similarity search frameworks compare structural fingerprints of candidate ligands to known active compounds to identify potential binders [8]. These approaches are valuable when protein structures are unknown, as they only require ligand information. However, their performance declines when the number of known ligands is insufficient [9]. In summary, docking-based methods provide mechanistic insights but require structural data, while ligand-based methods suit high-throughput screening but need ligand data. These limitations have driven the development of machine learning approaches that integrate diverse data and handle limited information scenarios [10].

Deep learning revolutionized DTI prediction via sequence-based approaches and graph-based approaches. Sequence-based methods leverage protein amino acid sequences and drug SMILES strings as their fundamental inputs. For instance, DeepConv-DTI [11] incorporates amino acid sequences and ECFP4 drug fingerprints, leverages multiscale convolutional neural networks (CNNs) to process variable-length protein sequences, and captures residue patterns for DTI interaction prediction. DeepDTA [12] utilizes a multilayer 1-D CNN architecture to process primary protein sequences and compound SMILE strings, extracting critical features for continuous binding affinity regression. MolTrans [13] employs knowledge-mined discrete substructures from drugs and proteins, processed through hierarchical transformer modules, to model semantic-level substructure relationships and interac-

Received 13 December 2024; revised 12 July 2025 and 1 November 2025; accepted 9 December 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62002390 and Grant 62472445 and in part by the Scientific Research Foundation of Hunan Provincial Education Department under Grant 22B0153. *(Corresponding author: Chengqian Lu.)*

Wen Tian and Chengqian Lu are with the College of Computer Science, Xiangtan University, Xiangtan, Hunan 411105, China, and also with the Key Laboratory of Intelligent Computing and Information Processing, Ministry of Education, Xiangtan 411105, China (e-mail: tianwen@smail.xtu.edu.cn; chengqlu@xtu.edu.cn).

Min Zeng and Jianxin Wang are with the School of Computer Science and Engineering, Central South University, Changsha 410083, China (e-mail: zengmin@csu.edu.cn; jxwang@mail.csu.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2025.3646159, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2025.3646159

tions for explainable DTI prediction. In contrast, graph-based approaches exploit 2-D molecular structures converted into graphs for atomic-level binding insights [14], where atoms form nodes and bonds constitute edges. This representation allows graph neural networks (GNNs) to grasp intrinsic drug properties. For instance, GNN-CPI [15] processes protein sequences via CNN and encodes drug molecular graphs with GNNs, and identifies compound-protein interactions after attentive fusion. GraphDTA [16] extracts atomic-level topological features of drug molecules with GNNs and amino acid sequence features with CNNs, and predicts drug–target binding affinities. GraphCDR [17] integrates multiomics cancer cell profiles and drug molecular graph structures, and leverages GNNs enhanced by contrastive learning regularization to improve robustness and generalization in cancer drug response prediction. PhySIcoCHemICal graph neural network (PSICHIC) [18] utilizes protein sequences and ligand SMILES inputs through physicochemical-constrained GNNs to generate interpretable interaction fingerprints, enabling structure-free prediction of binding mechanisms and affinity. KGDRP [19] integrates diverse biological entities (drugs, proteins, and cell lines) via a heterogeneous biomedical graph, employing knowledge-guided GNNs with auxiliary constraints to enhance drug response prediction and target discovery in cold-start scenarios.

Despite these advances, two limitations persist. First, many methods only focus on drug molecules, while ignoring protein structures. Since DTIs typically occur in binding pockets on protein surfaces or interiors [20], spatial structural information is crucial for accurate prediction. AttentionSiteDTI [21] encodes the protein pockets and learns their correlation with self-attention. Nonetheless, the rich contextual information of the entire protein sequence is neglected. Second, special motifs in drugs that are closely related to DTI binding patterns are often overlooked, such as frequent fragments with specific structures or patterns [22].

To address these issues, we propose a novel motif-based model named MotifGT-DTI for DTI prediction. First, we construct drug molecular graph motifs and protein pocket graphs based on 2-D molecular structures and 3-D spatial structures. The drug encoder utilizes a graph transformer (GT) [23] to capture both node-level and edge-level attributes, enabling global aggregation of atomic and bond features while mitigating local smoothing. The protein encoder consists of two parts: the pocket subgraph encoding module and the feature fusion module. To characterize the proteins comprehensively, we fuse the features of the 1-D sequences and 3-D pocket structures. Finally, a bilinear attention network (BAN) [24] is employed to identify potential interactions. Experimental results on four publicly available datasets show the superior predictive performance of our model. The ablation experiments prove that the GT module is effective. To validate the model's generalization ability, we conduct three different cold-start experiments. The results demonstrate that our model maintains accuracy in unseen environments. In addition, visualization experiments show that MotifGT-DTI captures key molecular motifs in binding and provides interpretability for the prediction results. The main contributions of our work are summarized as follows.

1) For proteins, we integrate 1-D sequence features and 3-D structural features to characterize the protein fully. For drugs, we extract motif-based representations.
2) To capture the node and edge features, we employ a GT to encode both the drug motifs and protein pocket subgraphs, thereby aggregating global information and mitigating over-smoothing.
3) The robustness and efficiency of MotifGT-DTI have been validated in comparison, cold-start, and visualization experiments. Notably, MotifGT-DTI's interpretability allows it to decode important drug motifs in DTI, providing valuable insights into the interaction mechanism between drugs and target proteins, and guidance for motif-based drug design tasks.

II. Related Work

Graph embedding representation has emerged as a pivotal research focus in drug-related domains. Based on GNNs' neighborhood aggregation, molecular graph models update node features iteratively from neighbors to encode graph structure. GNN-CPI [15] constructs r-radius subgraphs to capture local chemical environments, then expands the receptive field through multilayer graph convolutional networks (GCNs) aggregation for fusion features. MGraphDTA [25] employs a 27-layer multiscale GNN to capture drug molecules' local and global structural information across different hierarchical levels, generating feature representations. Based on the foundation of GCNs, the graph attention network (GAT) utilizes attention mechanisms for dynamic weight aggregation over neighbor nodes. Graph-DTA [16] adopts a hybrid architecture that employs GAT in the first layer for attention weights and GCN in the second for feature smoothing, and achieves optimal drug–target affinity prediction performance. Similarly, self-attention graph drug-target affinity (SAGDTA) [26] first extracts graph structural features using GCN and then introduces a self-attention mechanism to assign importance scores to atom nodes, thereby highlighting more critical local structures involved in DTIs. FragXsiteDTI [27] utilizes two topology adaptive GCN layers to capture multiscale local structures and GAT to extract attention scores between neighboring graph nodes, preserving nuanced interaction patterns in drug–protein recognition. The GT is fundamentally a Transformer architecture adapted for graph structures, employing global attention mechanisms to model graph features. Our work uses global attention to alleviate over-smoothing and capture global topology, differing from GCNs and local attention. Its edge features dynamically modulate attention scores, incorporating structural semantics like bond strength. Its spectrally defined Laplacian positional encodings inject distance-aware priors into initial embeddings, mitigating homogenization and preserving distinct biochemical properties of protein pockets.

III. Methods

A. Problem Description

The interactions between drugs and targets are governed by fundamental physicochemical properties, such as hydrogen

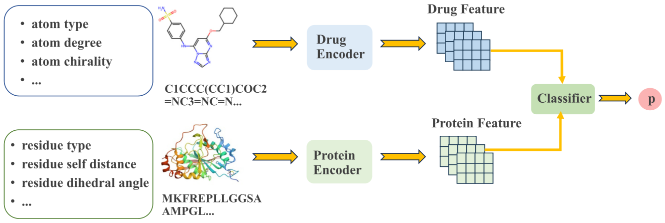


Fig. 1. Diagram of DTI prediction. The DTI pair is first represented as the model input in a certain feature form, such as SMILES and amino acid sequences. After feature extraction, the learned embeddings are fed into a classifier to calculate the candidate DTI pair probability p.

bonds, van der Waals forces, and other molecular-level interactions. These properties stem from the structural attributes of drug molecules (e.g., atomic compositions and bond characteristics) and target proteins (e.g., amino acid sequences and residue features). The DTI prediction task is therefore a structure-property relationship learning problem, where machine-learning models infer interactions from structural data. As shown in Fig. 1, the raw structures of drugs and proteins contain essential information relevant to DTI formation, including atomic features, edge features of the drug, and residue features of the protein. The deep learning model uses specialized encoders to extract complex structural features from both proteins and small molecules for candidate DTI predictions. Formally, let $\mathcal{D} = \{d_1, d_2, \ldots, d_{N_d}\}$ denote a finite set of drug molecules, and $\mathcal{P} = \{p_1, p_2, \ldots, p_{M_p}\}$ denote a finite set of target proteins. Here, N_d and M_p are the cardinalities of the drug and protein sets, respectively. $X_d \in R^{N_d \times f_d}$ denotes the feature matrix of drug molecule d_i, containing f_d-dimensional physicochemical attributes (e.g., as atom types, electronegativity. Similarly, let $X_t \in R^{M_p \times f_p}$ denote the residue feature matrix of target protein p_j, comprising f_p-dimensional properties (e.g., amino acid types, side chain polarity). The DTI prediction task aims to learn a mapping function $f : D \times P \to \{0, 1\}$ based on the feature matrices, where $f(d_i, p_j) = 1$ indicates an interaction and $f(d_i, p_j) = 0$ otherwise.

B. Model Structure

The framework of MotifGT-DTI is illustrated in Fig. 2. First, molecular graph motifs and protein pocket subgraphs are constructed from the drug molecular graph and protein 3-D structure, respectively. Then, the constructed graphs are fed into dedicated encoders to obtain graph feature vectors. The protein encoder contains a pocket graph encoder and a feature fusion module. Finally, the encoded drug feature and protein feature are passed into a BAN to calculate the possibility of potential interactions. Specifically, the drug representation, protein representation, and prediction module are detailed as follows:

1) Drug Representation: Drug molecules are fundamentally represented as 2-D undirected graphs with atoms as nodes and bonds as edges, capturing structural connectivity, functional groups, and key physicochemical properties. MacFrag [28] partitions molecular graphs into chemically meaningful fragments with modified BRICS rules, producing Rule of Three-compliant motifs with enhanced diversity. Naturally, GNNs provide an effective framework for such graph-structured data. However, traditional GNNs exhibit two notable limitations. First, these models typically discard valuable edge features, focusing solely on node data despite critical connection-encoded information. Second, the key issue is over-smoothing, where iterative neighbor aggregation modeled as a Laplacian smoothing process causes node features to converge toward homogeneity after multiple layers. To address the neglect of edge features, the GT incorporates edge features by modulating implicit node attention scores through multiplication with linearly projected edge embeddings, enabling edge attributes to dynamically refine pairwise interactions. To mitigate over-smoothing, spectrally defined Laplacian positional encodings are projected and integrated into input node embeddings, injecting distance-structural priors that help preserve topological distinctiveness and reduce representation homogeneity.

Specifically, every drug molecule is composed of several molecular motifs (frequent fragments). Each molecular motif is expressed as a graph $G_d^m = (V_d, E_d)$ illustrated in Fig. 2(a), where V_d represents the set of atoms and E_d represents the set of chemical bonds in the motif. The initial node and edge features of drugs are derived from the chemical structure and properties of their molecules, specified in Table I. For node i, its feature is represented as a 41-D vector x_i. For the edge between node i and node j, its feature is represented as a 10-D vector y_{ij}. To align the node and edge reprsentations, x_i and y_{ij} are fed into linear layers to obtain hidden node representation $\hat{r}_i^0$ and edge representation e_{ij}^0 as follows:

$$\hat{r}_i^0 = w_A^0 x_i + b_A^0 \tag{1}$$

$$e_{ij}^0 = w_B^0 y_{ij} + b_B^0 \tag{2}$$

where w_A^0 and w_B^0 are the weight parameters, and b_A^0 and b_B^0 are the bias parameters.

Subsequently, Laplacian positional encodings integrate spectral eigenvectors to inject distance-structural priors, ensuring proximal nodes share similar features while distant nodes diverge. For a subgraph G_d^m with n nodes, we decompose its Laplacian matrix $\mathcal{L}$ to get its laplacian eigenvectors as follows:

$$\mathcal{L} = I - D^{-1/2} A D^{-1/2} = U^T \Lambda U \tag{3}$$

where I is the n-dimensional identity matrix, A is the $n \times n$ adjacent matrix of the subgraph G_d^m, D is the degree matrix, and Λ, U represent the diagonal eigenvalue matrix and the orthogonal eigenvector matrix, respectively. Laplacian positional encoding leverages the graph's spectral properties by selecting the smallest k nontrivial eigenvectors to form a multidimensional encoding λ_i for node i. This encoding is transformed via a learned linear layer and then integrated into each node's initial feature representation to add structural position information

$$\lambda_i^0 = W_C^0 \lambda_i + b_C^0 \tag{4}$$

$$r_i^0 = \hat{r}_i^0 + \lambda_i^0 \tag{5}$$

where λ_i^0 represents the transformed positional encoding obtained by applying a linear transformation to λ_i.

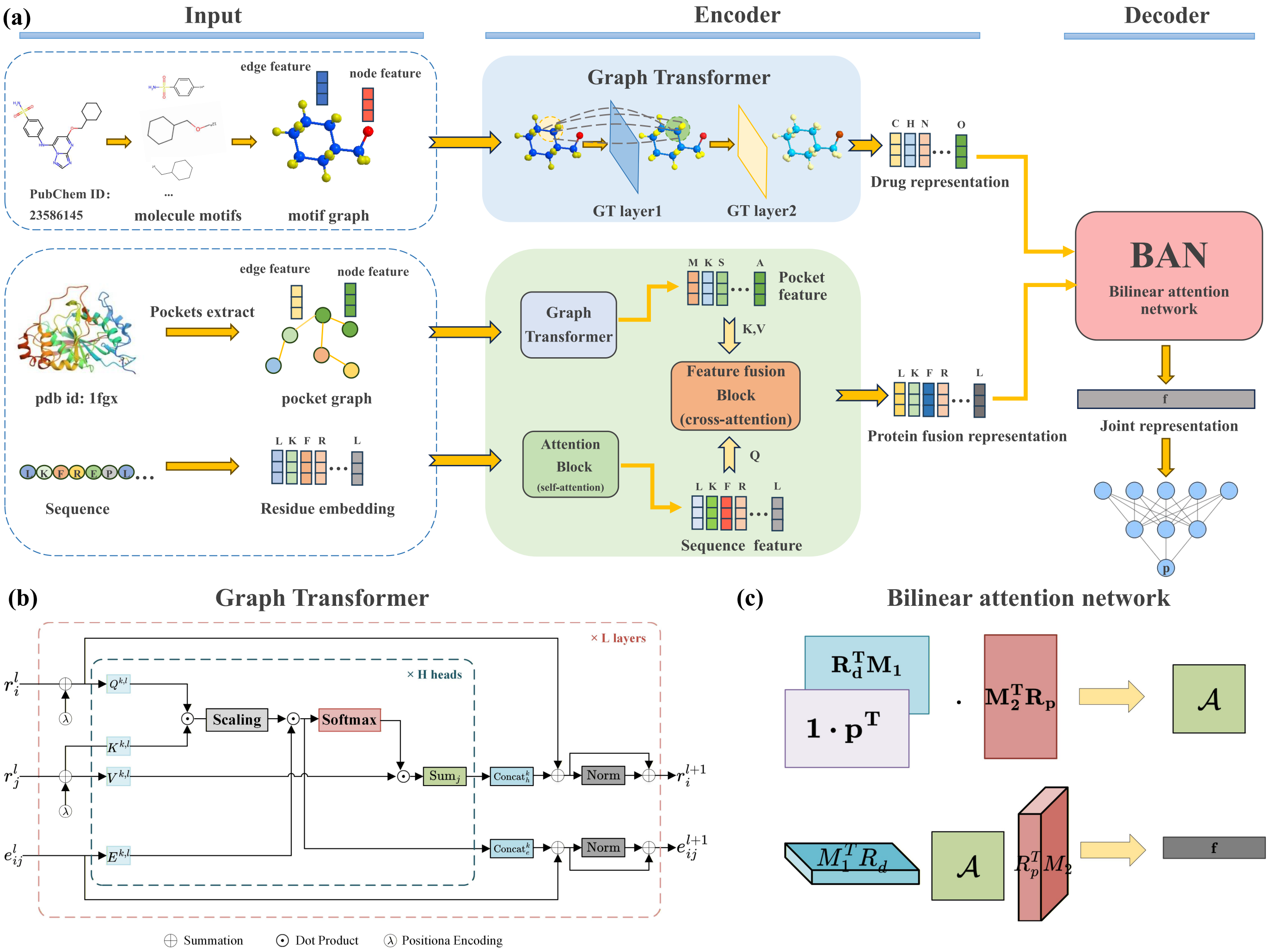


Fig. 2. Overview of the MotifGT-DTI framework. (a) Graphs are constructed for drug motifs and protein pockets, which are then encoded into feature vectors by their respective encoders. Subsequently, the drug motif features and the protein fusion features are fed into the BAN to calculate the final interaction prediction. (b) GT architecture. Based on the feature extraction of attention nodes, the GT extends the representation of edge features and aggregates global information through the attention mechanism to avoid excessive dependence on local features. (c) BAN architecture. The BAN meticulously captures bilinear interactions between input channels, using low-rank bilinear pooling to extract detailed joint representations to compute final probabilities.

TABLE I
INITIAL NODE AND EDGE FEATURES FOR DRUG GRAPH REPRESENTATION

	Features	Dimension	Description
Node	atom type	17	C, N, O, F, P, S, Cl, Br, I, B, Si, Fe, Zn, Cu, Mn, Mo, other
	atom degree	7	0, 1, 2, 3, 4, 5, 6
	atom formal charge	1	0 or 1
	atom num radical electrons	1	0 or 1
	atom hybridization	6	sp, sp2, sp3, sp3d, sp3d2, other
	atom is aromatic	1	0 or 1
	atom total num H	5	0,1,2,3,4
	atom chirality	3	R, S, other
Edge	bond type	4	single, double, triple, aromatic
	bond is conjugated	1	0 or 1
	bond is in ring	1	0 or 1
	bond stereo	4	stereonone, stereoany, stereoz, stereoe

By focusing on the relationship between a node and other nodes in the graph, the GT uses the self-attention mechanism to capture long-range dependencies and global information of the graph. For node i and its neighbor node j in the lth layer, the node features and edge features are normalized into the query $Q_{ij}^{h,l}$, key $K_{ij}^{h,l}$, and value $V_{ij}^{h,l}$. Inject node features and edge features to update and obtain accurate global attention. Specifically, the attention score between node i and node j for the hth head in the lth layer $W_{ij}^{h,l}$ is calculated based on both node features and edge features. The detailed updating steps for the lth layer are as follows:

$$Q_{ij}^{h,l} = Q^{h,l}\text{Norm}\left(r_i^l\right) \tag{6}$$

$$K_{ij}^{h,l} = K^{h,l}\text{Norm}\left(r_j^l\right) \tag{7}$$

$$V_{ij}^{h,l} = V^{h,l}\text{Norm}\left(r_j^l\right) \tag{8}$$

$$E_{ij}^{h,l} = E^{h,l}\text{Norm}\left(e_{ij}^l\right) \tag{9}$$

where $Q^{h,l}$, $K^{h,l}$, $V^{h,l}$, and $E^{h,l}$ are the parameters of the linear layers.

Then, pairwise attention scores are dynamically modulated by elementwise multiplication with node query–key similarities, allowing edge properties to amplify or suppress node interactions

$$w_{ij}^{h,l} = \text{softmax}\left(\left(\frac{Q_{ij}^{h,l} \cdot K_{ij}^{h,l}}{\sqrt{d_k}}\right) \cdot E_{ij}^{h,l} e_{ij}^l\right). \tag{10}$$

Through this global attention mechanism, the GT makes the node pay attention to all other nodes in the graph and successfully alleviates the problem of over-smoothing [29].

To capture the representation of different feature spaces, the features from multiattention heads are concatenated to update the $(l+1)$th layer as follows:

$$\hat{r}_i^{l+1} = r_i^l + O_h^l\left(\text{Concat}_{h=1}^{H_n}\left(\sum_{j \in r_i} w_{ij}^{h,l} V_{ij}^{h,l} r_j^l\right)\right) \tag{11}$$

$$\hat{e}_{ij}^{l+1} = e_{ij}^l + O_e^l\left(\text{Concat}_{h=1}^{H_n}\left(w_{ij}^{h,l}\right)\right) \tag{12}$$

where O_h^l and O_e^l represent the parameters of the linear layers, and H_n represents the number of attention heads.

Finally, the node and edge features after multiple iterations are fed into a feedforward network with residual connections and normalization operations [30]

$$r_i^{l+1} = \hat{r}_i^{l+1} + W_{n2}^l\left(\text{ReLU}\left(W_{n1}^l\,\text{Norm}\left(\hat{r}_i^{l+1}\right)\right)\right) \tag{13}$$

$$e_{ij}^{l+1} = \hat{e}_{ij}^{l+1} + W_{e2}^l\left(\text{ReLU}\left(W_{e1}^l\,\text{Norm}\left(\hat{e}_{ij}^{l+1}\right)\right)\right) \tag{14}$$

where W_{n1}^l, W_{e1}^l, W_{n2}^l, and W_{e2}^l are the linear parameters. Then, we get the node features, including edge features r_i^{l+1}. A drug motif m contains n nodes, and its feature R_d^m is a collection of features formed by multiple nodes, $R_d^m = [r_1^{l+1}, r_2^{l+1}, r_3^{l+1}, \ldots, r_n^{l+1}]$. Finally, each drug molecule d composed of multiple motifs is denoted as R_d. The specific process is illustrated in Fig. 2(b).

2) Protein Representation: There is a close relationship between the 1-D sequence and the 3-D structure of a protein. For example, the 1-D amino acid sequence of a protein dictates its 3-D spatial structure, and the 3-D structure determines the function and properties of the protein, such as the binding sites. To obtain a comprehensive protein characterization, we integrate the 1-D structure (sequences) with the 3-D structure (pockets) features.

For protein pockets, the 3-D structure files are first retrieved from the protein data bank (PDB) [31] using corresponding PDB IDs, excluding unavailable entries. Subsequently, a protein-clustering method [32] based on computational geometry is used to identify the best candidate pockets for each protein. Then, each protein is represented as multiple pocket subgraphs. Each pocket graph is represented as $G_p^p = (V_p, E_p)$, where V_p represents the set of residues and E_p represents the set of chemical bonds between residues. The GT is used for the node features and edge features of pocket graphs. The initial node and edge features of proteins are derived from the chemical structure and properties of their molecules, specified in Table II. Each residue node is represented by a 41-D vector encoding its biochemical properties. The edge between two residues is represented by a 5-D vector encoding connectivity and geometric distance between residues. Based on the constructed pocket graph and encoded initial vectors, we obtain the protein pocket node and edge features following a process analogous to that for drug molecules. Similarly, each protein composed of several pockets is a collection of pocket features and is denoted as R_{pp}.

For the 1-D sequence, each amino acid is encoded as a 20-D vector by one-hot encoding. Then, the feature matrix $P_s \in R^{l_{\text{seq}} \times 20}$ is obtained for each sequence, where l_{seq} represents the maximum sequence length. The initial vectors are then fed into a self-attention module to learn the internal correlations within the sequence, described as follows:

$$Q_s = W_Q^s P_s,\ K_s = W_K^s P_s,\ V_s = W_V^s P_s \tag{15}$$

$$\begin{aligned} R_{\text{ps}} &= \text{Attention}_s\left(Q_s, K_s, V_s\right) \\ &= \text{softmax}\left(\frac{Q_s K_s^T}{\sqrt{d_k}}\right) V_s. \end{aligned} \tag{16}$$

Cross-attention mechanism is successfully used in combining asymmetrically two separate embedding [33]. We utilize a cross-attention module to integrate 1-D sequence features with 3-D structural features, enabling the model to capture the complex relationships between sequence and spatial structure. The feature fusion process is as follows:

$$Q_p = W_Q R_{\text{ps}},\quad K_p = W_K R_{\text{pp}},\ V_p = W_V R_{\text{pp}} \tag{17}$$

$$\begin{aligned} R_p &= \text{Attention}_c\left(Q_p, K_p, V_p\right) \\ &= \text{softmax}\left(\frac{Q_p K_p^T}{\sqrt{d_k}}\right) V_p \end{aligned} \tag{18}$$

where R_{pp} represents protein pocket features, R_{ps} represents protein sequence features. Finally, we can obtain protein features R_p that contain both sequence information and structural information.

3) Prediction Module: DTIs primarily arise from localized interactions between specific drug substructures and protein binding pocket residues. BAN is introduced because of its strength in capturing local feature correlations, proven in visual question answering [24]. BAN's low-rank bilinear attention mechanism models these pairwise interaction strengths as shown in Fig. 2(c). Specifically, given the drug molecular feature matrix R_d and protein feature matrix R_p, BAN first generates a bilinear attention mapping matrix $\mathcal{A}$ defined as

$$\mathcal{A} = \text{softmax}\left(\left(\left(1 \cdot q^T\right) \circ {R_d}^T M_d\right) M_p^T R_p\right) \tag{19}$$

where M_d and M_p are learnable transformation matrices, q is a learnable weight vector, $\circ$ represents the Hadamard product

TABLE II
INITIAL NODE AND EDGE FEATURES FOR PROTEIN POCKET GRAPH REPRESENTATION

	Features	Dimension	Description
Node	residue type	22	G, A, V, L, I, P, F, Y, W,S, T, C, M, N, Q, D, E, K, R, H, metal, other
	residue self distance	5	the maximum and minimum values of the scaled distances among all atoms in a residue, as well as the scaled distances between the CA and O atoms, O and N atoms, and C and N atoms
	residue dihedral angle	4	phi, psi, omega, chi1
Edge	residue is connected	1	0 or 1
	residue CA distance	1	scaled distance between the CA atoms
	residue center distance	1	scaled distance between the center
	residue max distance	2	max and min values of the scaled distance

(elementwise multiplication). This formulation quantifies the interaction strength between drug substructures and protein residues via elementwise operations. Subsequently, the joint representation f is extracted via the bilinear pooling layer

$$f = \sigma\left((R_d)^T M_d\right)^T \cdot \mathcal{A} \cdot \sigma\left(\left(R_p\right)^T M_p\right) \tag{20}$$

where σ denotes the rectified linear unit (ReLU) activation function. This process achieves parameter efficiency through shared weight matrices M_d and M_p. Finally, the joint feature f is fed into the MLP layer to obtain the final attention score $\widehat{y}$, as

$$\widehat{y} = \text{Sigmoid}\,(Wf + b) \tag{21}$$

where W and b are the learnable weight matrix and bias vector, respectively.

During training, we optimize all learnable parameters by minimizing the cross-entropy loss $\mathcal{L}$ and backpropagating gradients as follows:

$$\mathcal{L} = -\sum_i \left(y_i \log\left(\widehat{y}_i\right) + (1 - y_i)\log\left(1 - \widehat{y}_i\right)\right) + \frac{\lambda}{2}\|\theta\|_2^2 \tag{22}$$

where θ is a vector containing all the learnable parameters in the model, λ is the regulation parameter, y_i is the ground truth of the ith drug–target pair, $\widehat{y}_i$ is corresponding probability predicted by the model, and $\mathcal{L}$ is the total loss.

IV. EXPERIMENT SETTING

A. Datasets

We downloaded four public datasets in DTI prediction: Human, Biosnap, Drugbank, and BindingDB. The Human dataset integrates DrugBank, Matador, and STITCH, applying confidence scoring and dissimilarity rules for negative samples, and includes 2726 drugs and 1787 proteins [34]. The BioSNAP dataset combines STITCH and DrugBank data, and offers balanced positive and negative pairs across 4510 drugs and 2181 proteins [35]. DrugBank serves as a comprehensive pharmaceutical knowledge base, documenting 6647 drugs and 4255 validated targets [36]. BindingDB is a public, large-scale, web-accessible database compiling experimentally validated drug–protein interactions from scientifically curated sources, composed of 49 199 interactions among 14 643 drugs and 2623 proteins [37]. The statistical information for the four datasets in our experiments is shown in Table III.

TABLE III
SUMMARY OF THE DATASETS

Dataset	Proteins	Drugs	Interactions	Active	Inactive
Human	1787	2726	6728	3364	3364
Biosnap	2181	4510	27476	13738	13738
Drugbank	4255	6647	35022	17511	17511
BindingDB	2623	14643	49199	20674	28525

TABLE IV
PARAMETER SETTINGS OF MOTIFGT-DTI

Parameter	Setting
Feature dimensions of nodes in drug graph	44
Feature dimensions of edges in drug graph	10
Feature dimensions of nodes in protein graph	41
Feature dimensions of edges in protein graph	5
The number of graph transformer layers	2
The number of graph transformer attention heads	[2,4,8]
The dimensions of drug embeddings	[64,128,256,512]
The dimensions of protein embeddings	[64,128,256,512]

To objectively evaluate model performance and mitigate data partitioning bias, we implemented a fivefold stratified cross-validation strategy [38]. The dataset was randomly partitioned into five mutually exclusive folds. In each iteration, the model was trained taking data from the combined training set (80% of the data) with a batch size of 32, and then evaluated on the single held-out test fold (20%). This process was repeated sequentially until each fold served as the test set once. The final results of the model performance evaluation were the mean ± standard deviations of the fivefold cross-validation.

B. Experimental Setup

The batch size affecting the training efficiency and stability refers to the number of samples per iteration or each update of the model weight during training. The optimizer determines how to update the model parameters according to the gradient, influencing the convergence speed of the model. Here, the batch size is set to 32, and the Adam optimizer is employed with a learning rate set to 10^{-4}. A total of 100 epochs are trained on each dataset. All benchmark experiments were conducted on an NVIDIA RTX 4090 GPU using the parameter configurations specified in Table IV.

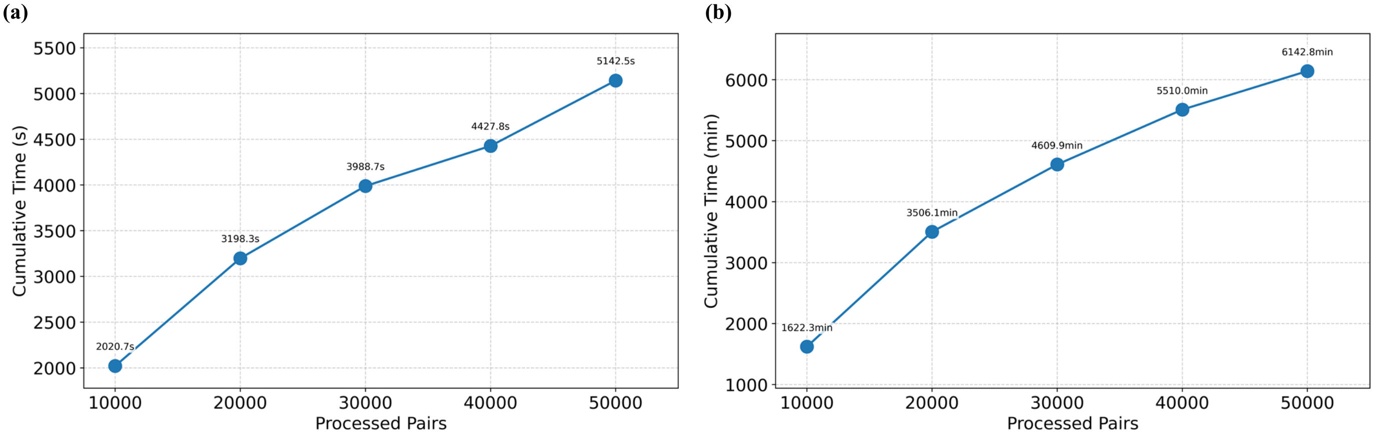


Fig. 3. Scalability and efficiency analysis of our model on the BindingDB dataset. (a) Load times for different scales of data. (b) Optimization time required by the model across varying data scales.

C. Efficiency and Scalability

To evaluate computational efficiency, we conduct a detailed time-complexity analysis of core modules using Big O notation: the GT and the BAN. The GT's complexity primarily stems from input initialization, attention mechanism, and feature updates. For a single-layer incorporating edge features, the overall time complexity is $O(|V|d_n^2+|E|d_e^2+|E|Hd_e)$, where $|V|$ is the number of nodes, $|E|$ is the number of edges, d_n is the dimension of node feature, d_e is the dimension of edge feature, and H is the number of attention heads. BAN confines time complexity at $O(Kf_pM_p)$ via matrix chain multiplication and low-rank factorization, avoiding costly pairwise calculations while maintaining performance. Here, K denotes the rank, f_p denotes the dimension, and M_p denotes the number of residues. As the GT dominates among all of the modules, the overall time complexity is $O(|V|d_n^2 + |E|d_e^2 + |E|Hd_e)$.

To rigorously evaluate computational scalability, an in-depth analysis of data-loading efficiency and model optimization time is conducted. Fig. 3(a) illustrates the relationship between data-loading time and dataset scale on the largest dataset, BindingDB, showing an approximately linear correlation with the total number of drug–target pairs. This indicates that loading time remains predictable and scales sublinearly even with larger datasets, rather than exhibiting exponential growth. Similarly, Fig. 3(b) presents the model optimization time on varying data scales. Evaluation at 10 000-pair intervals reveals a scaling pattern: cumulative training-time growth progressively decelerates with increasing dataset size. This sublinear scaling pattern suggests that the model's convergence time stays within manageable bounds even with substantial increases in dataset size. Additional scalability analysis experiments on the other three datasets are provided in Section I of the Supplementary Information.

D. Baselines

To validate the performance of MotifGT-DTI, we compare it with six state-of-the-art baseline models on four public datasets. TransformerCPI [39] is a transformer-based model for predicting protein–compound interactions. It encodes proteins using Word2Vec on 3-g fragments followed by a three-layer gated CNN. Compounds are encoded as atomic-level graphs and processed through a single-layer GCN. DTIs are predicted by a three-layer transformer decoder with self-attention, attention aggregation, and fully connected layers. MolTrans [13] predicts DTIs using knowledge-inspired substructure mining and an interaction module. It encodes proteins as amino acid sequences and compounds as SMILES, mines frequent substructures via the frequent consecutive subsequence algorithm, processes them with an augmented two-layer Transformer, and computes interaction scores via dot products of substructure embeddings. DeepMGT-DTI [40] predicts DTIs by using a transformer to fuse multilayer graph convolutional features from drug molecules and a CNN for target sequences, achieving accuracy on benchmark datasets. DrugBAN [41] predicts DTIs through dual feature extraction pathways. Drugs undergo a three-layer GCN to encode molecular graphs, and proteins are processed by a three-layer CNN for sequence encoding. A bilinear attention mechanism models atom-subsequence interactions, generating a joint representation to predict binding affinity with enhanced interpretability. FragXsiteDTI [27] is a transformer-based model built upon the PerceiverIO framework [42] for DTI prediction. Its core innovation lies in simultaneously utilizing protein binding pockets and drug molecular fragments as inputs, generating embeddings through a two-layer topology adaptive GAT. A learnable latent query array employs cross-attention and self-attention mechanisms to capture interactions, predicting binding probability through MLP, achieving superior performance and interpretability. IMAEN [43] is an interpretable molecular augmentation model for DTI prediction, employing GNNs and attention mechanisms. It enhances molecular structure representation through a molecular structure augmentation module and processes protein sequences using an interpretable stack convolutional encoding module, achieving state-of-the-art performance on four DTI benchmark datasets. MMDG-DTI [44] predicts DTIs using multimodal feature fusion and domain generalization. It integrates textual features via pretrained LLMs (Prot-Bert, Smiles-Bert), structural features through a hybrid GNN, and employs DAT and contrastive learning

TABLE V

PERFORMANCE COMPARISON ON FOUR BENCHMARK DATASETS

Dataset	Model	AUROC ↑	AUPRC ↑	Precision ↑	F1 Score ↑	Recall ↑
Human	TransformerCPI	0.967 ± 0.012	0.970 ± 0.015	0.943 ± 0.014	0.927 ± 0.003	0.935 ± 0.002
	MolTrans	0.979 ± 0.002	0.982 ± 0.001	**0.964 ± 0.004**	0.913 ± 0.003	0.920 ± 0.004
	DeepMGT-DTI	0.975 ± 0.001	0.973 ± 0.004	0.938 ± 0.006	0.938 ± 0.012	0.952 ± 0.007
	DrugBAN	0.983 ± 0.011	0.974 ± 0.005	0.946 ± 0.004	0.924 ± 0.002	0.941 ± 0.002
	FragXsiteDTI	0.977 ± 0.002	0.979 ± 0.002	0.921 ± 0.003	0.908 ± 0.003	0.914 ± 0.002
	IMAEN	0.960 ± 0.004	0.939 ± 0.004	0.933 ± 0.011	0.941 ± 0.014	0.938 ± 0.002
	MMDG-DTI	0.984 ± 0.003	0.987 ± 0.003	0.959 ± 0.001	0.947 ± 0.001	0.953 ± 0.007
	MSI-DTI	0.980 ± 0.002	0.975 ± 0.003	0.947 ± 0.012	0.934 ± 0.007	0.940 ± 0.004
	MotifGT-DTI	**0.995 ± 0.001**	**0.994 ± 0.002**	0.954 ± 0.001	**0.971 ± 0.001**	**0.962 ± 0.002**
Biosnap	TransformerCPI	0.822 ± 0.004	0.854 ± 0.002	0.805 ± 0.003	0.700 ± 0.003	0.749 ± 0.001
	MolTrans	0.870 ± 0.003	0.881 ± 0.002	0.768 ± 0.032	0.827 ± 0.005	0.789 ± 0.005
	DeepMGT-DTI	0.888 ± 0.005	0.904 ± 0.003	0.825 ± 0.011	0.777 ± 0.011	0.800 ± 0.006
	DrugBAN	0.902 ± 0.004	0.902 ± 0.013	0.830 ± 0.001	0.843 ± 0.004	0.836 ± 0.003
	FragXsiteDTI	0.893 ± 0.001	0.903 ± 0.003	0.830 ± 0.001	0.814 ± 0.005	0.822 ± 0.002
	IMAEN	0.864 ± 0.002	0.841 ± 0.004	0.812 ± 0.011	0.802 ± 0.014	0.808 ± 0.002
	MMDG-DTI	0.872 ± 0.001	0.865 ± 0.001	0.803 ± 0.002	0.819 ± 0.001	0.811 ± 0.005
	MSI-DTI	0.875 ± 0.005	0.904 ± 0.004	0.816 ± 0.006	0.813 ± 0.007	0.814 ± 0.006
	MotifGT-DTI	**0.918 ± 0.001**	**0.926 ± 0.001**	**0.844 ± 0.001**	**0.852 ± 0.001**	**0.861 ± 0.002**
Drugbank	TransformerCPI	0.809 ± 0.005	0.807 ± 0.004	0.728 ± 0.005	0.728 ± 0.005	0.730 ± 0.003
	MolTrans	0.856 ± 0.005	0.863 ± 0.001	0.724 ± 0.002	0.843 ± 0.001	0.777 ± 0.002
	DeepMGT-DTI	0.874 ± 0.004	0.888 ± 0.002	0.771 ± 0.009	0.812 ± 0.010	0.801 ± 0.006
	DrugBAN	0.862 ± 0.003	0.868 ± 0.002	0.759 ± 0.004	0.823 ± 0.002	0.791 ± 0.002
	FragXsiteDTI	0.879 ± 0.003	0.881 ± 0.002	0.792 ± 0.004	0.801 ± 0.001	0.797 ± 0.005
	IMAEN	0.837 ± 0.004	0.825 ± 0.011	0.757 ± 0.012	0.795 ± 0.014	0.775 ± 0.002
	MMDG-DTI	0.853 ± 0.001	0.851 ± 0.001	0.771 ± 0.002	0.777 ± 0.001	0.774 ± 0.001
	MSI-DTI	0.849 ± 0.005	0.847 ± 0.005	0.778 ± 0.007	0.758 ± 0.008	0.767 ± 0.007
	MotifGT-DTI	**0.896 ± 0.001**	**0.902 ± 0.001**	**0.803 ± 0.001**	**0.853 ± 0.001**	**0.827 ± 0.001**
BindingDB	TransformerCPI	0.889 ± 0.002	0.861 ± 0.007	0.767 ± 0.006	0.777 ± 0.006	0.787 ± 0.006
	MolTrans	0.929 ± 0.004	0.895 ± 0.006	0.767 ± 0.006	0.815 ± 0.007	0.884 ± 0.008
	DeepMGT-DTI	0.935 ± 0.006	0.912 ± 0.005	0.845 ± 0.002	0.812 ± 0.011	0.830 ± 0.007
	DrugBAN	0.953 ± 0.001	0.929 ± 0.001	0.850 ± 0.002	**0.888 ± 0.005**	0.888 ± 0.008
	FragXsiteDTI	0.935 ± 0.003	0.900 ± 0.002	0.833 ± 0.007	0.817 ± 0.007	0.802 ± 0.004
	IMAEN	0.921 ± 0.009	0.871 ± 0.008	0.815 ± 0.007	0.826 ± 0.007	0.836 ± 0.007
	MMDG-DTI	0.949 ± 0.002	0.930 ± 0.003	0.873 ± 0.001	0.874 ± 0.008	0.875 ± 0.002
	MSI-DTI	0.939 ± 0.004	0.901 ± 0.005	0.817 ± 0.006	0.814 ± 0.007	0.815 ± 0.006
	MotifGT-DTI	**0.956 ± 0.001**	**0.932 ± 0.009**	**0.859 ± 0.002**	0.874 ± 0.001	**0.889 ± 0.004**

The results are presented as mean±standard deviation and the best results for the metrics on each dataset are highlighted in bold.

to enhance generalization across unseen domains. MSI-DTI [45] predicts DTIs by integrating multisource information like molecular fingerprints, protein sequences, and knowledge graph embeddings, using a multihead self-attention mechanism with residual connections to model feature interactions.

E. Evaluation Metrics

We employ the following five metrics to evaluate MotifGT-DTI performance on four datasets: area under the receiver operating characteristic curve (AUROC), area under the precision–recall curve (AUPRC), precision, $F1$-score, and recall.

AUROC is an indicator that measures the classification ability of the model. The higher its value, the stronger the ability to distinguish positive and negative samples. The formulas for the horizontal true positive rate (TPR) and the vertical false positive rate (FPR) in the receiver operating characteristic (ROC) curves are as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{23}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \tag{24}$$

AUPRC demonstrates the relationship between precision and recall of the model under different thresholds, which effectively evaluates performance in imbalanced datasets. Precision refers to the proportion of positive samples predicted by the model that are actually positive, reflecting the performance in terms of false positives. Recall is a metric for evaluating the performance in terms of false negatives. The corresponding precision and recall formulas are as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{25}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{26}$$

$F1$-score is the harmonic mean of precision and recall, evaluating the overall performance of the model. The specific formula is as follows:

$$F1\ \mathrm{score} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{27}$$

TABLE VI
ABLATION STUDY ON FOUR DATASETS

Dataset	Model	AUROC ↑	AUPRC ↑	Precision ↑	F1 Score ↑	Recall ↑
Human	MotifGT-DTI$_{cat}$	0.986 ± 0.001	0.988 ± 0.001	0.946 ± 0.002	0.946 ± 0.002	0.946 ± 0.002
	MotifGT-DTI$_{ave}$	0.981 ± 0.001	0.984 ± 0.001	0.950 ± 0.001	0.916 ± 0.005	0.933 ± 0.003
	MotifGT-DTI$_{gcn}$	0.976 ± 0.002	0.976 ± 0.001	0.909 ± 0.006	0.924 ± 0.005	0.916 ± 0.004
	MotifGT-DTI$_{gat}$	0.977 ± 0.012	0.976 ± 0.004	0.906 ± 0.003	0.950 ± 0.011	0.928 ± 0.014
	MotifGT-DTI$_{sage}$	0.981 ± 0.004	0.982 ± 0.003	0.920 ± 0.008	0.950 ± 0.023	0.935 ± 0.074
	MotifGT-DTI$_{gin}$	0.977 ± 0.008	0.978 ± 0.009	0.903 ± 0.010	0.950 ± 0.003	0.926 ± 0.004
	MotifGT-DTI$_{ca}$	0.987 ± 0.001	0.989 ± 0.001	0.950 ± 0.011	0.965 ± 0.003	0.957 ± 0.001
	w/o sequence	0.983 ± 0.002	0.980 ± 0.001	0.913 ± 0.021	0.950 ± 0.003	0.931 ± 0.001
	w/o structure	0.960 ± 0.001	0.961 ± 0.002	0.901 ± 0.001	0.930 ± 0.002	0.915 ± 0.001
	MotifGT-DTI	**0.995 ± 0.001**	**0.994 ± 0.002**	**0.954 ± 0.001**	**0.971 ± 0.001**	**0.962 ± 0.002**
Biosnap	MotifGT-DTI$_{cat}$	0.908 ± 0.001	0.918 ± 0.001	0.817 ± 0.013	0.855 ± 0.018	0.836 ± 0.004
	MotifGT-DTI$_{ave}$	0.907 ± 0.001	0.917 ± 0.001	0.800 ± 0.003	**0.879 ± 0.002**	0.838 ± 0.003
	MotifGT-DTI$_{gcn}$	0.906 ± 0.003	0.919 ± 0.002	0.822 ± 0.009	0.870 ± 0.003	0.845 ± 0.005
	MotifGT-DTI$_{gat}$	0.901 ± 0.004	0.915 ± 0.003	0.826 ± 0.028	0.849 ± 0.023	0.837 ± 0.004
	MotifGT-DTI$_{sage}$	0.901 ± 0.008	0.914 ± 0.009	0.829 ± 0.010	0.842 ± 0.003	0.840 ± 0.004
	MotifGT-DTI$_{gin}$	0.907 ± 0.023	0.921 ± 0.001	0.830 ± 0.004	0.838 ± 0.011	0.844 ± 0.005
	MotifGT-DTI$_{ca}$	0.874 ± 0.002	0.878 ± 0.001	0.785 ± 0.005	0.836 ± 0.006	0.810 ± 0.003
	w/o sequence	0.909 ± 0.003	0.919 ± 0.012	0.829 ± 0.001	0.858 ± 0.001	0.843 ± 0.001
	w/o structure	0.888 ± 0.002	0.906 ± 0.003	0.797 ± 0.001	0.837 ± 0.002	0.816 ± 0.001
	MotifGT-DTI	**0.918 ± 0.002**	**0.926 ± 0.001**	**0.844 ± 0.002**	0.852 ± 0.003	**0.861 ± 0.001**
Drugbank	MotifGT-DTI$_{cat}$	0.895 ± 0.001	**0.903 ± 0.001**	**0.818 ± 0.005**	0.823 ± 0.001	0.821 ± 0.003
	MotifGT-DTI$_{ave}$	0.890 ± 0.001	0.893 ± 0.001	0.804 ± 0.005	0.811 ± 0.001	0.807 ± 0.007
	MotifGT-DTI$_{gcn}$	0.886 ± 0.003	0.899 ± 0.002	0.790 ± 0.002	0.831 ± 0.002	0.810 ± 0.002
	MotifGT-DTI$_{gat}$	0.887 ± 0.004	0.891 ± 0.003	0.801 ± 0.001	0.827 ± 0.003	0.814 ± 0.004
	MotifGT-DTI$_{sage}$	0.889 ± 0.002	0.892 ± 0.004	0.796 ± 0.010	0.829 ± 0.013	0.812 ± 0.004
	MotifGT-DTI$_{gin}$	0.890 ± 0.023	0.896 ± 0.011	0.795 ± 0.004	0.830 ± 0.001	0.812 ± 0.005
	MotifGT-DTI$_{ca}$	0.876 ± 0.001	0.878 ± 0.001	0.800 ± 0.002	0.834 ± 0.002	0.816 ± 0.001
	w/o sequence	0.888 ± 0.002	0.896 ± 0.001	0.845 ± 0.011	0.748 ± 0.002	0.794 ± 0.003
	w/o structure	0.862 ± 0.001	0.874 ± 0.002	0.765 ± 0.011	0.792 ± 0.013	0.778 ± 0.001
	MotifGT-DTI	**0.896 ± 0.002**	0.902 ± 0.001	0.803 ± 0.003	**0.853 ± 0.002**	**0.827 ± 0.002**
BindingDB	MotifGT-DTI$_{cat}$	0.945 ± 0.003	0.915 ± 0.004	0.837 ± 0.005	0.837 ± 0.005	0.837 ± 0.005
	MotifGT-DTI$_{ave}$	0.947 ± 0.011	0.916 ± 0.002	0.832 ± 0.014	0.834 ± 0.011	0.833 ± 0.002
	MotifGT-DTI$_{gcn}$	0.948 ± 0.002	0.917 ± 0.003	0.823 ± 0.004	0.838 ± 0.003	0.830 ± 0.004
	MotifGT-DTI$_{gat}$	0.940 ± 0.008	0.905 ± 0.009	0.801 ± 0.010	0.839 ± 0.007	0.819 ± 0.006
	MotifGT-DTI$_{sage}$	0.944 ± 0.003	0.912 ± 0.007	0.811 ± 0.008	0.838 ± 0.005	0.824 ± 0.004
	MotifGT-DTI$_{gin}$	0.948 ± 0.004	0.917 ± 0.006	0.823 ± 0.005	0.838 ± 0.003	0.830 ± 0.002
	MotifGT-DTI$_{ca}$	0.942 ± 0.003	0.911 ± 0.004	0.855 ± 0.003	0.824 ± 0.005	0.839 ± 0.004
	w/o sequence	0.948 ± 0.002	0.919 ± 0.003	0.817 ± 0.005	0.858 ± 0.002	0.837 ± 0.004
	w/o structure	0.937 ± 0.004	0.903 ± 0.005	0.811 ± 0.006	0.832 ± 0.004	0.821 ± 0.005
	MotifGT-DTI	**0.956 ± 0.001**	**0.932 ± 0.009**	**0.859 ± 0.002**	**0.874 ± 0.001**	**0.889 ± 0.004**

The results are presented as mean±standard deviation and the best results for the metrics on each dataset are highlighted in bold.

V. Experimental Results

A. Comparison With Baselines

Detailed comparison results with the other eight baseline methods on four datasets are shown in Table V. From the table, we can see that MotifGT-DTI overall outperforms other models on all four datasets. It consistently achieves higher recall on all four datasets with a lower false negative rate. This suggests a higher probability of identifying potential drug targets because fewer positive samples are misclassified as negative. This advantage is particularly pronounced on the Human dataset, where MotifGT-DTI dominates all metrics, including optimal AUROC and AUPRC, reflecting its exceptional sensitivity in complex biological environments. On the medium-sized Biosnap dataset, it not only achieves the highest AUROC and AUPRC but also maintains robust recall, demonstrating superior generalization by effectively capturing key interaction patterns. Even on the challenging Drugbank dataset, where all models exhibit performance degradation due to scale and complexity, MotifGT-DTI sustains leadership in AUROC and AUPRC while preserving competitive recall, confirming its architectural stability. Similarly, it extends this dominance to the BindingDB dataset, the largest and most imbalanced benchmark. Specifically, it surpasses DeepMGT-DTI and MSI-DTI, highlighting the advantage of motif and structure-aware design over their graph-transformer and multisource attention approaches. The consistent excellence of MotifGT-DTI likely stems from its unique integration of motif-based graph learning and transformer architectures, which enables comprehensive feature extraction from molecular structures and protein sequences.

B. Ablation Experiment

We conduct ablation studies on four datasets to validate the effectiveness of the following modules: the feature fusion block, GT, and BAN. The experimental results are shown in Table VI. For the feature fusion strategy, we compare

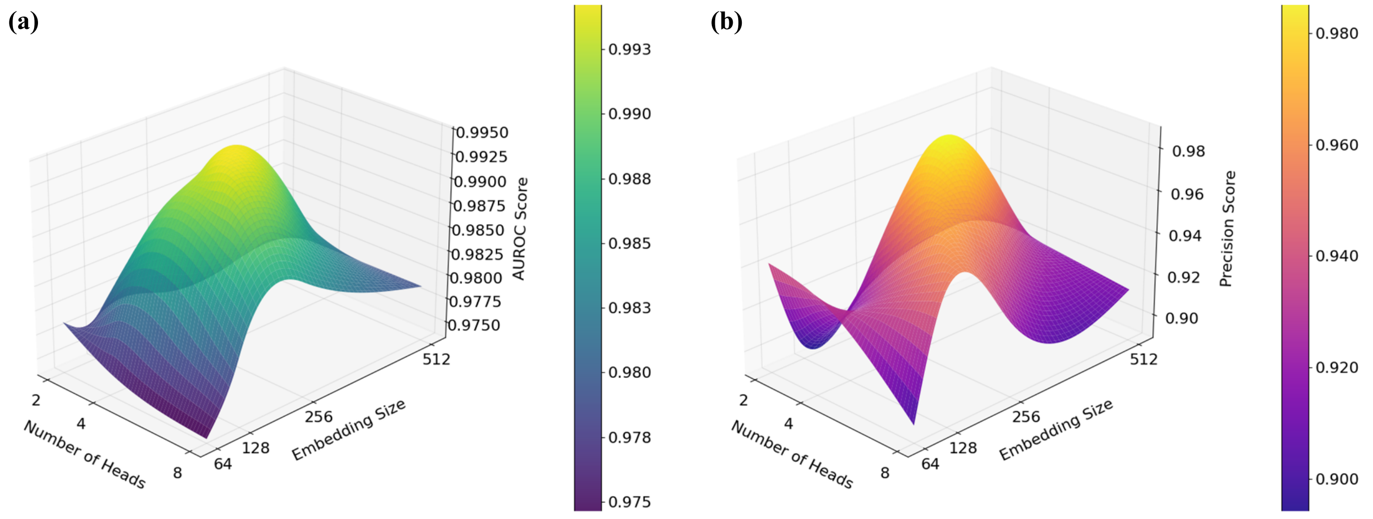


Fig. 4. Parameter analysis of heads and feature embeddings of drug and protein in the GT on the Human dataset. (a) AUROC values under different combinations of attention heads and feature embeddings. (b) Precision values under different combinations of attention heads and feature embeddings.

the performance of three different strategies: concatenation, averaging, and cross-attention. MotifGT-DTI$_{\text{cat}}$ is the model that concatenates the pocket features encoded by the GT and the sequence features encoded by self-attention as protein features. MotifGT-DTI$_{\text{ave}}$ is the model that takes the average of the pocket features and sequence features as the final protein features. As shown in Table VI, the results demonstrate that MotifGT-DTI with cross-attention outperforms all baseline methods across AUROC, AUPRC, recall, and $F1$-score metrics on all datasets, even on the large-scale BindingDB dataset with imbalanced active and inactive drug–target pairs. To further investigate the contribution of the GT module, we conduct an ablation study by replacing the original GT with several representative GNN architectures: GCN [46], GAT [47], GraphSAGE [48], and graph isomorphism network (GIN) [49]. GCN uses convolution to aggregate neighbor features for local graph structures. GAT employs attention to dynamically weight neighbors' importance. GraphSAGE leverages sampling and inductive learning for scalability to large, dynamic graphs. GIN combines MLPs with graph structures for high expressive power, proving robust in graph isomorphism testing. The results on all four datasets (Human, Biosnap, Drugbank, and BindingDB) are summarized in Table VI. A clear trend observed is that the MotifGT-DTI model consistently achieves the best performance. While GAT's attention mechanism and GraphSAGE's sampling strategy offer improvements over standard GCN in some cases, they still fall short of the proposed GT. This evidence leads us to conclude that the global self-attention mechanism and the modeling of edge features within the GT are crucial for capturing the complex topological and relational information necessary for accurate DTIs prediction, offering a distinct advantage over GNNs that rely predominantly on local neighborhood aggregation. Additionally, we conducted ablation experiments on the BAN module. We replaced the BAN module with a cross-attention module named MotifGT-DTI$_{ca}$. The comparison of the MotifGT-DTI$_{ca}$ and MotifGT-DTI demonstrates that the BAN captures drug–protein interactions more effectively than cross-attention.

To verify the impact of individual features, we conducted the following two experiments: 1) "w/o sequence" (removing 1-D sequence features while retaining 3-D structural features) and 2) "w/o structure" (removing 3-D structural features while retaining sequence features). From Table VI, it was clearly evident that performance consistently declined across all datasets (Human, Biosnap, Drugbank, and BindingDB) under both ablation scenarios. Critically, the decline is markedly more pronounced when structural features are omitted, particularly evident in key metrics like AUROC and AUPRC. It demonstrates that structural features contribute substantially more to protein representation than sequence features alone, likely due to their ability to capture critical spatial and biophysical properties essential for binding interactions. The overall experimental results confirm the necessity of multiview feature fusion, incorporating both sequence and 3-D structural features, for comprehensive characterization of proteins.

C. *Parameter Experiment*

In order to explore the parameters that affect the expressiveness and performance of the model, we perform grid search experiments on the following parameters: 1) the number of attention heads of the GT, which affects the model to capture the correlation between features and 2) the embedding dimension that determines the richness of the feature representation of the model. Specifically, we assess the effect of varying the number of attention heads within the discrete set [2, 4, 8], subject to the condition that the embedding dimension is divisible by the number of heads, and we explore embedding dimensions in [64, 128, 256, 512]. As shown in Fig. 4, AUROC and precision values improve as the number of heads increases from 2 to 4, peaking at 4. For embedding dimensions, the results showed that AUROC and precision consistently improve as sizes increase to 256 but

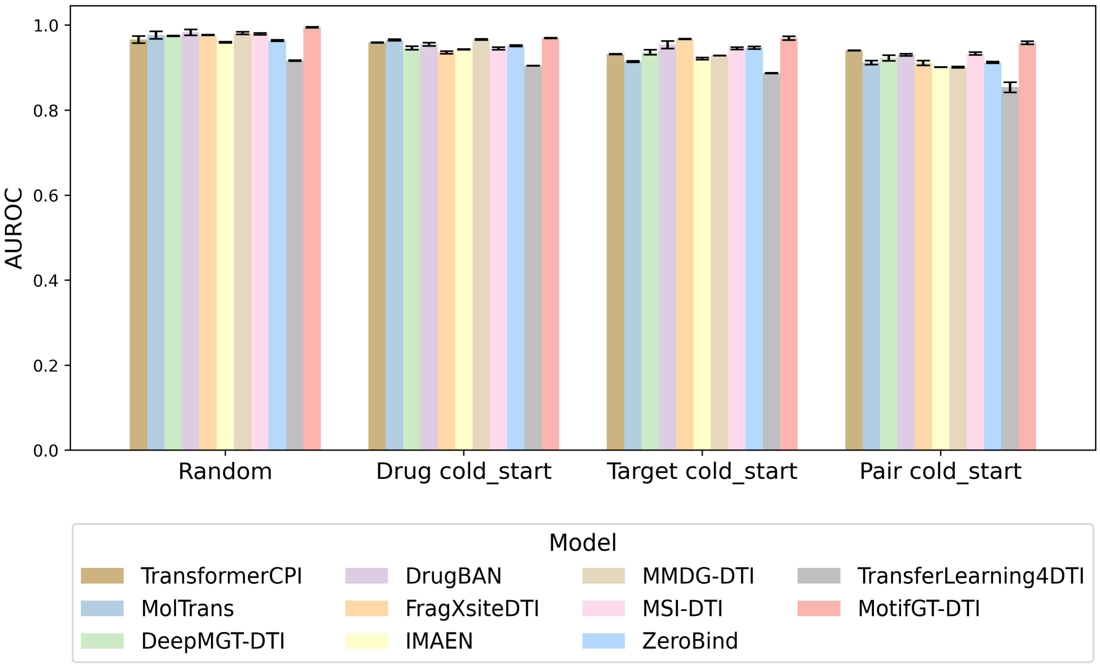


Fig. 5. Performance comparison on the Human dataset with random splitting strategy and three cold-start splitting strategies. Random split means randomly dividing the data to generate training data and test data. Drug cold-start, Target cold-start, and Pair cold-start represent three different cold-start splitting strategies, respectively. The bar chart above is plotted with the mean and standard deviation.

decline at 512, peaking at 256 dimensions, independent of attention heads. This unimodal trend highlights 256 as optimal, with higher dimensions causing overparameterization issues like noise or optimization hurdles, confirming robust findings across both metrics. The other parameter settings of MotifGT-DTI are shown in Table IV.

D. Cold-Start Experiment

In real-world drug development, DTI prediction models often face the challenge of cold-start, where some drugs or proteins have never been seen before. Whether the model can still provide accurate predictions is a key issue for the generalization ability and robustness. To evaluate the performance of MotifGT-DTI in the face of these special cases, we conduct cold-start experiments on the Human dataset. Specifically, we follow the previous work [41] and conduct experiments with three different data-splitting strategies.

1) *Drug Cold-Start:* The test set includes novel drugs absent from the training set.
2) *Target Cold-Start:* The test set includes novel proteins absent from the training set.
3) *Pair Cold-Start:* The test set includes novel drug–target pairs where both the drug and target are absent from the training set.

Drug cold-start and target cold-start scenarios, respectively, evaluate the sensitivity of the model to novel drugs and targets. Pair cold-start measures the generalization ability of the model when facing completely unknown drug–target pairs without prior knowledge. The effects of cold-start splits and random splits were tested on the human dataset for all models as shown in Fig. 5. In the cold-start scenario, all models experienced varying degrees of performance degradation. In the paired cold-start scenario that often occurs in practical applications, the performance of each prediction drops most significantly. It indicates that the cold-start scenario requires the model to have higher generalization ability for unseen proteins or drugs.

To further evaluate the generalization capability under data-scarce cold-start scenarios, we have additionally incorporated two state-of-the-art baseline methods: ZeroBind and Transferlearning4DTI. ZeroBind [50] leverages GNNs with a subgraph information bottleneck on protein and drug graphs to identify binding pockets for zero-shot DTI prediction. Transferlearning4DTI [51] utilizes transfer learning and pretrained compound representations to fine-tune feedforward neural networks for accurate DTI prediction. We used its publicly accessible pretrained models during testing. All methods were evaluated using identical data splits across four datasets: Human, Biosnap, DrugBank, and BindingDB. The detailed experimental results on the Human dataset are illustrated in Table VII. In the Drug cold-start scenario, it achieves the highest AUROC (0.970) and AUPRC (0.968), indicating exceptional generalization for novel drugs. ZeroBind achieves the best performance in terms of AUPRC. For the Target cold-start, MotifGT-DTI leads in AUROC (0.968) and precision (0.844), suggesting MotifGT-DTI effectively captures generalized protein features. Most notably, in the challenging pair cold-start scenario where all models show significant performance drops, MotifGT-DTI dominates all metrics (AUROC: 0.957, AUPRC: 0.872, precision: 0.860), substantially exceeding runner-up models. In cold-start scenarios, MotifGT-DTI demonstrates superior generalization over DeepMGT-DTI and MSI-DTI, particularly in the stringent pair cold-start setting,

TABLE VII
RESULTS ON HUMAN UNDER DIFFERENT COLD-START SETTINGS

Setting	Model	AUROC ↑	AUPRC ↑	Precision ↑
Drug cold-start	TransformerCPI	0.968 ± 0.001	0.968 ± 0.004	0.920 ± 0.008
	MolTrans	0.966 ± 0.003	0.940 ± 0.011	0.938 ± 0.015
	DeepMGT-DTI	0.947 ± 0.005	0.931 ± 0.004	0.875 ± 0.012
	DrugBAN	0.956 ± 0.006	0.948 ± 0.001	0.906 ± 0.009
	FragXsiteDTI	0.936 ± 0.004	0.927 ± 0.003	0.871 ± 0.005
	IMAEN	0.943 ± 0.001	0.938 ± 0.005	0.859 ± 0.013
	MMDG-DTI	0.967 ± 0.001	0.958 ± 0.001	**0.939 ± 0.002**
	ZeroBind	0.952 ± 0.003	**0.970 ± 0.002**	0.877 ± 0.004
	TransferLearning4DTI	0.905 ± 0.001	0.816 ± 0.004	0.802 ± 0.003
	MSI-DTI	0.954 ± 0.003	0.946 ± 0.004	0.897 ± 0.006
	MotifGT-DTI	**0.970 ± 0.002**	0.968 ± 0.001	0.902 ± 0.003
Target cold-start	TransformerCPI	0.932 ± 0.001	0.889 ± 0.004	0.778 ± 0.001
	MolTrans	0.914 ± 0.002	0.730 ± 0.053	0.780 ± 0.053
	DeepMGT-DTI	0.937 ± 0.006	0.912 ± 0.008	0.805 ± 0.009
	DrugBAN	0.954 ± 0.007	0.921 ± 0.003	0.786 ± 0.006
	FragXsiteDTI	0.968 ± 0.001	**0.968 ± 0.002**	0.839 ± 0.004
	IMAEN	0.924 ± 0.007	0.935 ± 0.021	0.835 ± 0.012
	MMDG-DTI	0.928 ± 0.001	0.858 ± 0.001	0.802 ± 0.013
	ZeroBind	0.947 ± 0.005	0.934 ± 0.004	0.787 ± 0.007
	TransferLearning4DTI	0.887 ± 0.002	0.778 ± 0.003	0.716 ± 0.001
	MSI-DTI	0.946 ± 0.004	0.879 ± 0.006	0.838 ± 0.007
	MotifGT-DTI	**0.968 ± 0.005**	0.951 ± 0.003	**0.844 ± 0.002**
Pair cold-start	TransformerCPI	0.942 ± 0.001	0.798 ± 0.001	0.731 ± 0.015
	MolTrans	0.911 ± 0.005	0.790 ± 0.027	0.682 ± 0.019
	DeepMGT-DTI	0.924 ± 0.007	0.801 ± 0.011	0.734 ± 0.012
	DrugBAN	0.931 ± 0.002	0.837 ± 0.001	0.671 ± 0.008
	FragXsiteDTI	0.909 ± 0.007	0.838 ± 0.003	0.730 ± 0.005
	IMAEN	0.904 ± 0.007	0.748 ± 0.004	0.716 ± 0.025
	MMDG-DTI	0.902 ± 0.002	0.819 ± 0.011	0.772 ± 0.001
	ZeroBind	0.913 ± 0.003	0.843 ± 0.002	0.805 ± 0.003
	TransferLearning4DTI	0.854 ± 0.019	0.805 ± 0.001	0.733 ± 0.016
	MSI-DTI	0.934 ± 0.004	0.834 ± 0.007	0.784 ± 0.008
	MotifGT-DTI	**0.957 ± 0.004**	**0.872 ± 0.002**	**0.860 ± 0.001**

The results are presented as mean±standard deviation. Bold values indicate the best performance in each metric group.

underscoring the robustness of its learned molecular and structural representations. This consistent superiority stems from its motif-based GT architecture, which extracts robust molecular and protein representations that generalize to unseen entities better than metalearning (ZeroBind) or transfer-learning (TransferLearning4DTI) approaches, ensuring stability across diverse cold-start conditions. The experimental results on the Biosanp, DrugBank, and BindingDB are detailed in the supplementary (Tables 1–3).

E. Feature Visualization for Encoder

To analyze the effect of the encoder, we conducted a visualization experiment on the feature vectors before and after encoding. Taking the Human dataset as an example, we randomly selected 500 DTI pairs and projected the 256-D embeddings into a 2-D embedding space using the t-distributed neighbor embedding (t-SNE) [52] method. The results are shown in Fig. 6. In the left figure, the distribution of positive and negative samples is relatively dispersed, and there is no obvious clustering or separation trend before encoding. It suggests that the original features have a limited ability to differentiate between positive and negative samples. From the right figure, it is obvious that although some positive and negative samples are still not completely separated, the encoded positive and negative samples have been clustered into clearly different classes. By comparing the t-SNE visualization results before and after encoding, we infer that the encoder successfully extracts the significant features of positive and negative samples. Therefore, MotifGT-DTI transforms the nondiscriminative original features into more informative low-dimensional representations that can encode features effectively.

F. Interpretability With Attention Visualization

To assess the interpretability of DTI predictions, we performed a series of visualization experiments. Human l-lactate dehydrogenase A (LDH-A) is a key enzyme that promotes the growth of cancer cells; inhibiting its activity is of great significance for cancer treatment [53]. We downloaded the complex formed by LDH-A and ligand 9YA from the PDB (PDB ID: 5W8L), which is not present in the training data. It is a typical pair code start case. It is noteworthy that our model successfully identified this unseen protein–ligand interaction pair. It shows that the model can provide accurate predictions for data that is not present in the training data, indicating good generalization performance.

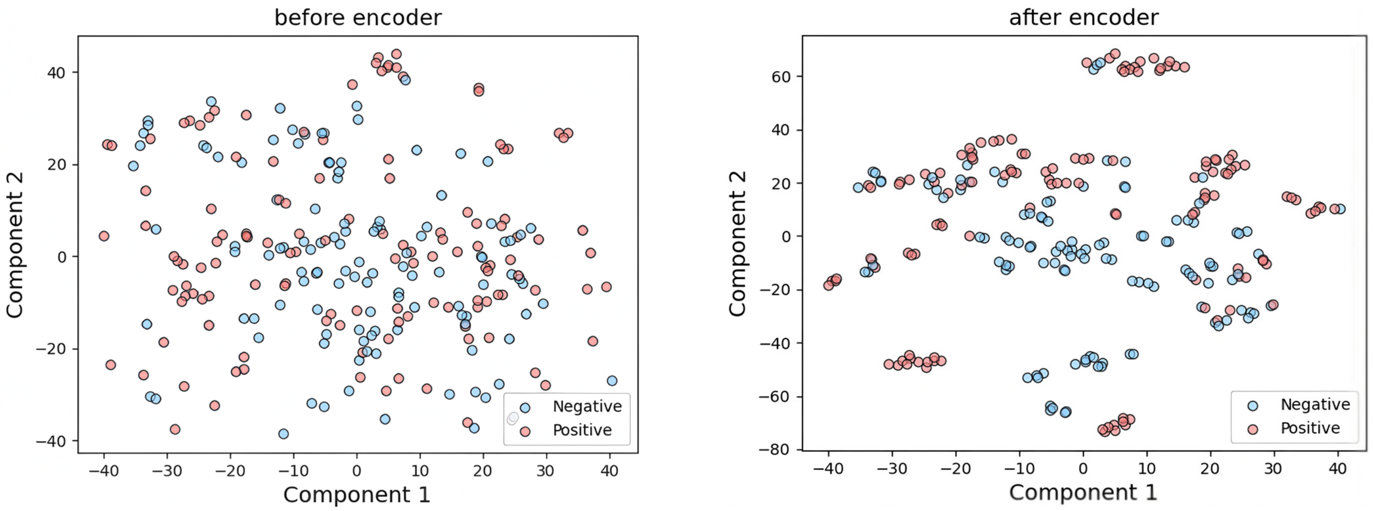


Fig. 6. Visualizations of drug–target pair embeddings on the Human dataset. Positive (left) and negative (right) pair features before and after encoding, respectively.

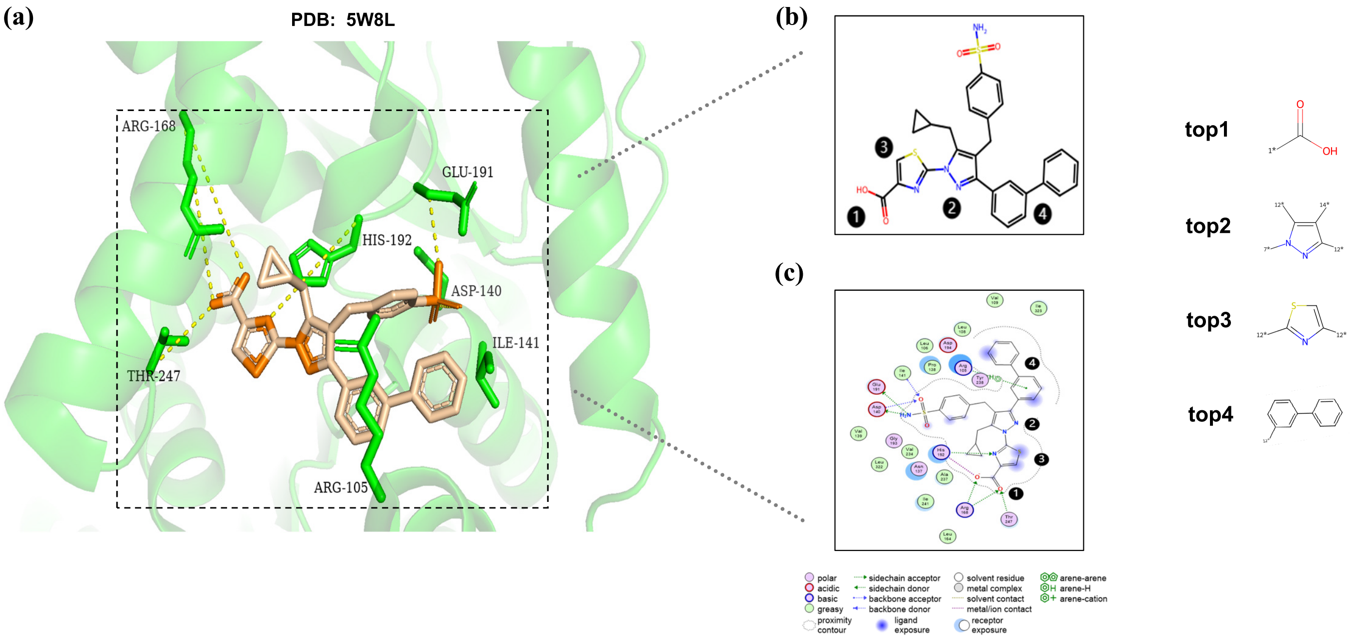


Fig. 7. Visualization study for interpretability. (a) Binding pocket structure's interpretability. 3-D representations of the binding sites in the 5W8L (ligand 9YA binding with human LDH-A) complex structure. (b) Highest probability motifs of 9YA. We highlight the top four drug motifs with the highest attention scores in different colors. (c) Visualization of binding sites in the corresponding crystal structures using the molecular operating environment (MOE) software.

Furthermore, to provide detailed interpretability, we conducted experiments on attention visualization. The bilinear attention map visualizes each substructure, providing insights and explanations on the key aspects of drug design work at the molecular level. The top four drug motifs with the highest predicted attention scores in the attention map are visualized and marked with dashed circles as shown in Fig. 7(b), providing empirical evidence for the necessity of motif-based techniques in achieving interpretability and biological relevance. Referring to known real protein–ligand sites, MotifGT-DTI correctly predicted 3 of them as shown in Fig. 7(c), and the remaining one is found to be highly related to the action site. Specifically, the bound key motif carboxylic acid group is correctly identified (acting as hydrogen bond receptors interacting with residues Arg168, His192, and Thr247). The thiazole structure interacting with residue His192 is correctly identified. The biphenyl ring structure interacting with residues Arg105 and Asn137 is also correctly predicted to be an important drug motif involved in the binding process. Although the predicted pyrazole structure is not a currently known site, studies [54] show that it is a key chemical feature that interacts with LDH-A in related inhibitors and is closely associated with reducing LDH-A enzyme activity and inhibiting cancer cell growth. All the results indicate that MotifGT-DTI identifies the functional molecular motifs contributing to the interaction and provides interpretability for predicted results.

VI. Conclusion

DTI prediction is a key step in virtual screening, which is applied in various fields, including new drug development and drug repositioning. Although traditional virtual screening biological experiments are reliable, they require a lot of time and money, which hinders their application to large-scale datasets. Recently, deep learning-based methods have made great progress, speeding up drug–target prediction and allowing the processing of large datasets. However, existing methods have their shortcomings in exploring the properties of protein and drug molecular structures. From the perspective of biological mechanisms, DTIs typically occur through hydrogen bonds or other types of interactions between the functional groups of the drug and the residues of the protein. Moreover, these interactions are closely related to the pocket structure of the protein. Therefore, the substructure of the drug and the pocket structure of the protein are crucial for predicting whether a drug–target pair will interact.

In this article, we propose a novel motif-based method with GT for DTI prediction, named MotifGT-DTI. Given the importance of protein structure in DTI prediction, we first extract the pocket information of the protein to serve as input from a 3-D structural perspective. Fusing 1-D sequence and 3-D structure, MotifGT-DTI provides a multiview protein representation. To fully leverage molecular properties for prediction, MotifGT-DTI adopts the motif-based method with edge features as the drug encoding strategy. GT effectively captures intricate topological information within the graph, thereby deepening the understanding of the structural characteristics of drug molecules and protein pockets.

The comparison experimental results show that MotifGT-DTI outperforms the state-of-the-art methods in overall prediction performance, illustrating the effectiveness of our strategies. Cold-start experiments across three different scenarios confirm the strong generalization ability of our model in complex real-world environments. Attention visualization study demonstrates that MotifGT-DTI captures key molecular motifs, providing interpretability for prediction results. In the above studies, we find that the special substructures of drugs (motifs) and proteins (pockets) are crucial for predicting DTIs. Despite MotifGT-DTI's promising results, limitations exist. The model's dependence on experimentally derived protein structures (PDB files) restricts its applicability, as these are unavailable for many proteins. Furthermore, while effectively capturing motifs and pockets, the current focus on these specific elements potentially overlooks other relevant molecular features critical for DTI understanding. Future work will prioritize integrating computationally predicted structures (e.g., AlphaFold3 [55]) to broaden the scope beyond PDB-reliant proteins. We will also explore incorporating a wider range of molecular descriptors beyond motifs and pockets, enhancing both universality and biological insight into interaction mechanisms.

References

[1] M. Bagherian, E. Sabeti, K. Wang, M. A. Sartor, Z. Nikolovska-Coleska, and K. Najarian, "Machine learning approaches and databases for prediction of drug–target interaction: A survey paper," *Briefings Bioinf.*, vol. 22, no. 1, pp. 247–269, Jan. 2021.

[2] N. Shaikh, M. Sharma, and P. Garg, "An improved approach for predicting drug–target interaction: Proteochemometrics to molecular docking," *Mol. BioSystems*, vol. 12, no. 3, pp. 1006–1014, 2016.

[3] W. J. Allen et al., "DOCK 6: Impact of new features and current docking performance," *J. Comput. Chem.*, vol. 36, no. 15, pp. 1132–1156, Jun. 2015.

[4] D. Schaller et al., "Next generation 3D pharmacophore modeling," *WIREs Comput. Mol. Sci.*, vol. 10, no. 4, p. 1468, Jul. 2020.

[5] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[6] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nature Biotechnol.*, vol. 25, no. 2, pp. 197–206, Feb. 2007.

[7] Z. Du, D. Wang, and Y. Li, "Comprehensive evaluation and comparison of machine learning methods in QSAR modeling of antioxidant tripeptides," *ACS Omega*, vol. 7, no. 29, pp. 25760–25771, Jul. 2022.

[8] F. Napolitano et al., "Drug repositioning: A machine-learning approach through data integration," *J. Cheminformatics*, vol. 5, no. 1, p. 30, Dec. 2013.

[9] M. Srinivasarao and P. S. Low, "Ligand-targeted drug delivery," *Chem. Rev.*, vol. 117, no. 19, pp. 12133–12164, Oct. 2017.

[10] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities," *Inf. Fusion*, vol. 50, pp. 71–91, Oct. 2019.

[11] I. Lee, J. Keum, and H. Nam, "DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences," *PLOS Comput. Biol.*, vol. 15, no. 6, Jun. 2019, Art. no. e1007129.

[12] H. Öztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: Deep drug–target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, Sep. 2018.

[13] K. Huang, C. Xiao, L. M. Glass, and J. Sun, "MolTrans: Molecular interaction transformer for drug–target interaction prediction," *Bioinformatics*, vol. 37, no. 6, pp. 830–836, May 2021.

[14] L. David, A. Thakkar, R. Mercado, and O. Engkvist, "Molecular representations in AI-driven drug discovery: A review and practical guide," *J. Cheminformatics*, vol. 12, no. 1, p. 56, Dec. 2020.

[15] M. Tsubaki, K. Tomii, and J. Sese, "Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences," *Bioinformatics*, vol. 35, no. 2, pp. 309–318, Jan. 2019.

[16] T. Nguyen, H. Le, T. P. Quinn, T. M. Nguyen, T. D. Le, and S. Venkatesh, "GraphDTA: Predicting drug–target binding affinity with graph neural networks," *Bioinformatics*, vol. 37, no. 8, pp. 1140–1147, 2020.

[17] X. Liu, C. Song, F. Huang, H. Fu, W. Xiao, and W. Zhang, "GraphCDR: A graph neural network method with contrastive learning for cancer drug response prediction," *Briefings Bioinf.*, vol. 23, no. 1, p. 457, Jan. 2022.

[18] H. Y. Koh, T. Nguyen, S. Pan, L. T. May, and G. I. Webb, "Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data," *Nat. Mach. Intell.*, vol. 6, no. 6, pp. 673–687, May 2024.

[19] Q. Ye et al., "A Knowledge-guided graph learning approach bridging phenotype-and target-based drug discovery," *Adv. Sci.*, vol. 12, no. 16, Apr. 2025, Art. no. 2412402.

[20] A. Stank, D. B. Kokh, J. C. Fuller, and R. C. Wade, "Protein binding pocket dynamics," *Acc. Chem. Res.*, vol. 49, no. 5, pp. 809–815, 2016.

[21] M. Yazdani-Jahromi et al., "AttentionSiteDTI: An interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification," *Briefings Bioinf.*, vol. 23, no. 4, p. 272, Jul. 2022.

[22] Z. Zhang, Q. Liu, H. Wang, C. Lu, and C. Lee, "Motif-based graph self-supervised learning for molecular property prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 15870–15882.

[23] V. Prakash Dwivedi and X. Bresson, "A generalization of transformer networks to graphs," 2020, *arXiv:2012.09699*.

[24] J.-H. Kim, J.-H. Jun, and B. Zhang, "Bilinear attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018.

[25] Z. Yang, W. Zhong, L. Zhao, and C. Yu-Chian Chen, "MGraphDTA: Deep multiscale graph neural network for explainable drug–target binding affinity prediction," *Chem. Sci.*, vol. 13, no. 3, pp. 816–833, 2022.

[26] S. Zhang, M. Jiang, S. Wang, X. Wang, Z. Wei, and Z. Li, "SAG-DTA: Prediction of drug–target affinity using self-attention graph network," *Int. J. Mol. Sci.*, vol. 22, no. 16, p. 8993, Aug. 2021.

[27] A. K. Yalabadi, M. Yazdani-Jahromi, N. Yousefi, A. Tayebi, S. Abdidizaji, and Ö. Ö. Garibay, "FragXsiteDTI: Revealing responsible segments in drug-target interaction with transformer-driven interpretation," in *Proc. Int. Conf. Res. Comput. Mol. Biol.*, 2023, pp. 68–85.

[28] Y. Diao, F. Hu, Z. Shen, and H. Li, “MacFrag: Segmenting large-scale molecules to obtain diverse fragments with high qualities,” *Bioinformatics*, vol. 39, no. 1, p. 012, Jan. 2023.

[29] L. Rampášek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, and D. Beaini, “Recipe for a general, powerful, scalable graph transformer,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 14501–14515.

[30] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, “Layer normalization,” 2016, *arXiv:1607.06450*.

[31] H. M. Berman, “The protein data bank,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000.

[32] S. M. Saberi Fathi and J. A. Tuszynski, “A simple method for finding a protein s ligand-binding pockets,” *BMC Structural Biol.*, vol. 14, pp. 1–9, Jul. 2014.

[33] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, “Multi-modality cross attention network for image and sentence matching,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10938–10947.

[34] H. Liu, J. Sun, J. Guan, J. Zheng, and S. Zhou, “Improving compound–protein interaction prediction by building up highly credible negative samples,” *Bioinformatics*, vol. 31, no. 12, pp. i221–i229, Jun. 2015.

[35] S. M. M. Zitnik, R. Sosič, and J. Leskovec. *BioSNAP Datasets: Stanford Biomedical Network Dataset Collection*. Accessed: Nov. 2025. [Online]. Available: http://snap.stanford.edu/biodata

[36] D. S. Wishart et al., “DrugBank: A knowledgebase for drugs, drug actions and drug targets,” *Nucleic Acids Res.*, vol. 36, no. 1, pp. D901–D906, Jan. 2008.

[37] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, “BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1045–D1053, Jan. 2016.

[38] S. Zheng, Y. Li, S. Chen, J. Xu, and Y. Yang, “Predicting drug–protein interaction using quasi-visual question answering system,” *Nat. Mach. Intell.*, vol. 2, no. 2, pp. 134–140, 2020.

[39] L. Chen et al., “TransformerCPI: Improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments,” *Bioinformatics*, vol. 36, no. 16, pp. 4406–4414, Aug. 2020.

[40] P. Zhang, Z. Wei, C. Che, and B. Jin, “DeepMGT-DTI: Transformer network incorporating multilayer graph information for drug–target interaction prediction,” *Comput. Biol. Med.*, vol. 142, Mar. 2022, Art. no. 105214.

[41] P. Bai, F. Miljković, B. John, and H. Lu, “Interpretable bilinear attention network with domain adaptation improves drug–target prediction,” *Nature Mach. Intell.*, vol. 5, no. 2, pp. 126–136, Feb. 2023.

[42] A. Jaegle et al., “Perceiver IO: A general architecture for structured inputs & outputs,” 2021, *arXiv:2107.14795*.

[43] J. Zhang, Z. Liu, Y. Pan, H. Lin, and Y. Zhang, “IMAEN: An interpretable molecular augmentation model for drug–target interaction prediction,” *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 121882.

[44] Y. Hua, Z. Feng, X. Song, X.-J. Wu, and J. Kittler, “MMDG-DTI: Drug–target interaction prediction via multimodal feature fusion and domain generalization,” *Pattern Recognit.*, vol. 157, Jan. 2025, Art. no. 110887.

[45] W. Zhao et al., “MSI-DTI: Predicting drug-target interaction based on multi-source information and multi-head self-attention,” *Briefings Bioinf.*, vol. 25, no. 3, Mar. 2024.

[46] T. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *Proc. Int. Conf. Learn. Represent.*, 2017.

[47] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lió, and Y. Bengio, “Graph attention networks,” in *Proc. Int. Conf. Learn. Represent.*, 2018.

[48] W. L. Hamilton, R. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1024–1034.

[49] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How powerful are graph neural networks?,” in *Proc. Int. Conf. Learn. Represent.*, 2019.

[50] Y. Wang, Y. Xia, J. Yan, Y. Yuan, H.-B. Shen, and X. Pan, “ZeroBind: A protein-specific zero-shot predictor with subgraph matching for drug-target interactions,” *Nature Commun.*, vol. 14, no. 1, p. 7861, Nov. 2023.

[51] A. Dalkıran et al., “Transfer learning for drug–target interaction prediction,” *Bioinf.*, vol. 39, no. 1, pp. i103–i110, 2023.

[52] L. V. D. Maaten and G. E. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

[53] V. R. Fantin, J. St-Pierre, and P. Leder, “Attenuation of LDH—A expression uncovers a link between glycolysis, mitochondrial physiology, and tumor maintenance,” *Cancer Cell*, vol. 10, no. 2, p. 172, Aug. 2006.

[54] G. Rai et al., “Discovery and optimization of potent, cell-active pyrazole-based inhibitors of lactate dehydrogenase (LDH),” *J. Medicinal Chem.*, vol. 60, no. 22, pp. 9184–9204, Nov. 2017.

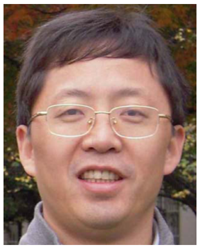
[55] J. Abramson et al., “Accurate structure prediction of biomolecular interactions with AlphaFold 3,” *Nature*, vol. 630, no. 8016, pp. 493–500, 2024.

Wen Tian is currently pursuing the master’s degree with the College of Computer Science, Xiangtan University, Xiangtan, Hunan, China.

Her research interests include bioinformatics, machine learning, and deep learning, especially in artificial intelligence-driven drug discovery (AIDD).

Min Zeng (Member, IEEE) received the B.Eng. degree from Lanzhou University, Lanzhou, China, in 2013, and the M.Eng. and Ph.D. degrees from Central South University, Changsha, China, in 2016 and 2020, respectively.

He is currently an Associate Professor at the School of Computer Science and Engineering, Central South University. His main research interests include bioinformatics, machine learning, and deep learning.



Jianxin Wang (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees in computer engineering from Central South University, Changsha, China, in 1992, 1996, and 2001, respectively.

He is currently the Chair and a Professor at the Department of Computer Science, Central South University. His current research interests include algorithm analysis and optimization, parameterized algorithms, bioinformatics, and computer networks.

Chengqian Lu received the Ph.D. degree in computer science from Central South University, Changsha, China, in 2019.

He is currently an Assistant Professor at Xiangtan University, Xiangtan, China. His research interests include bioinformatics, machine learning, and deep learning, especially in noncoding ribonucleic acid (RNA) and disease.