

# HiF-DTA: Hierarchical Feature Learning Network for Drug–Target Affinity Prediction

1<sup>st</sup> Minghui Li

*School of Software Engineering  
Huazhong University of  
Science and Technology  
Wuhan, China  
minghuili@hust.edu.cn*

2<sup>nd</sup> Yuanhang Wang

*School of Software Engineering  
Huazhong University of  
Science and Technology  
Wuhan, China  
wangyuanhang@hust.edu.cn*

3<sup>rd</sup> Peijin Guo

*School of Cyber Science and Engineering  
Huazhong University of  
Science and Technology  
Wuhan, China  
gpj@hust.edu.cn*

4<sup>th</sup> Wei Wan

*Faculty of Data Science  
City University of Macau  
Macau, China  
weiwan@cityu.edu.mo*

5<sup>th</sup> Shengshan Hu

*School of Cyber Science and Engineering  
Huazhong University of  
Science and Technology  
Wuhan, China  
hushengshan@hust.edu.cn*

6<sup>th</sup> Shengqing Hu

*Union Hospital,  
Tongji Medical College  
Huazhong University of  
Science and Technology  
Wuhan, China  
hsqha@126.com*

**Abstract**—Accurate prediction of Drug–Target Affinity (DTA) is crucial for reducing experimental costs and accelerating early screening in computational drug discovery. While sequence-based deep learning methods avoid reliance on costly 3D structures, they still overlook simultaneous modeling of global sequence semantic features and local topological structural features within drugs and proteins, and represent drugs as flat sequences without atomic-level, substructural-level, and molecular-level multi-scale features. We propose HiF-DTA, a hierarchical network that adopts a dual-pathway strategy to extract both global sequence semantic and local topological features from drug and protein sequences, and models drugs multi-scale to learn atomic, substructural, and molecular representations fused via a multi-scale bilinear attention module. Experiments on Davis, KIBA, and Metz datasets show HiF-DTA outperforms state-of-the-art baselines, with ablations confirming the importance of global-local extraction and multi-scale fusion.

**Index Terms**—Drug-target affinity prediction, hierarchical feature learning, multi-scale feature fusion.

## I. INTRODUCTION

Accurate prediction of drug–target affinity (DTA) is essential for drug screening, immune modulation and precision medicine. Experimental assays are costly and slow, driving the adoption of deep-learning models that operate in an end-to-end manner. Existing deep-learning methods fall into two streams: structure- and sequence-based. The former hinges on 3-D coordinates that are scarce and expensive to obtain [1]–[4], whereas the latter relies solely on SMILES or amino-acid strings and is more scalable [5], [6].

Minghui’s work is supported in part by the National Natural Science Foundation of China (Grant No. 62572206). Shengshan’s work is supported in part by the National Natural Science Foundation of China (Grant No.62372196). The work is supported by the HPC Platform of Huazhong University of Science and Technology.

Shengqing is the corresponding author.

Nevertheless, current sequence-based models have three common drawbacks: (1) *Local topology neglect*: global semantics are emphasised, but binding-site or substructural cues are overlooked [7]. (2) *Insufficient multi-scale modelling*: atom- or residue-level features are used, yet intermediate substructures are rarely exploited [8], [9]. (3) *Modality isolation*: CNN or GNN encoders process sequences or graphs separately, leaving semantics and structures unpaired (e.g., DeepDTA [7], GraphDTA [10]).

We present HiF-DTA, a hierarchical feature-learning network that addresses the above issues. A dual-pathway encoder captures global sequence semantics (BiLSTM for drugs, Mamba for proteins) and local topology (PNA-based GNN on molecular/residue graphs). For drugs, atomic, substructural and molecular features are explicitly extracted and fused by multi-scale bilinear attention, enabling fine-grained interaction modelling.

The main contributions of this work are as follows:

- HiF-DTA integrates global and local cues via dual-pathway encoding for both drugs and proteins.
- Three-level drug features (atom, substructure, molecule) are jointly learned and bilinearly attended to protein clusters.
- State-of-the-art results on Davis, KIBA and Metz with ablations verifying the value of global-local and multi-scale designs.

## II. MATERIALS AND METHODOLOGY

### A. Datasets

We evaluated the performance of our model on three widely used benchmark datasets: Davis [11], Metz [12], and KIBA [13]. See Table I for details. In these datasets, a smaller  $K_d$  value indicates a stronger binding affinity between the drug and the target. To reduce the variance, the  $K_d$  values in the

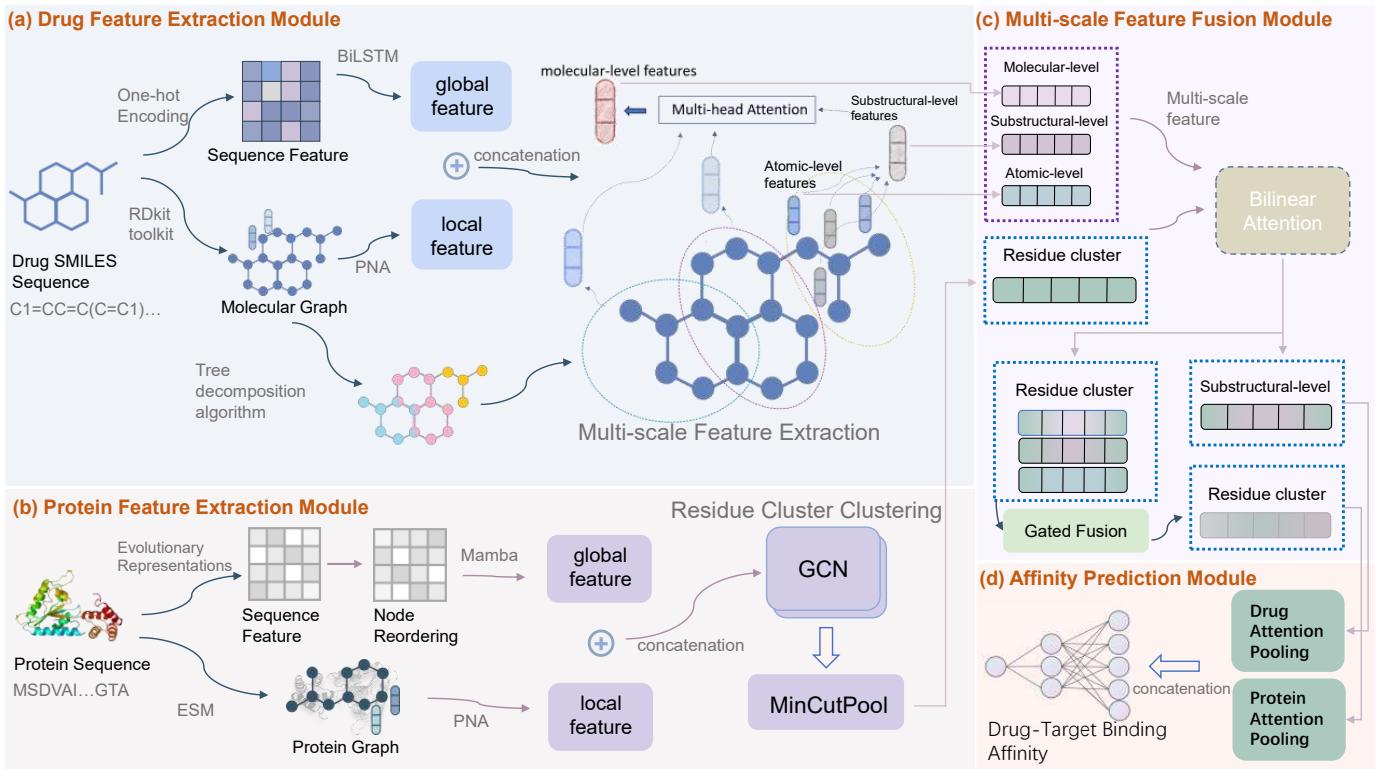


Fig. 1. Pipeline of the proposed HiF-DTA scheme.

Davis dataset are commonly transformed into a logarithmic scale using the following formula:

$$pK_d = -\lg \left( \frac{K_d}{10^9} \right) \quad (1)$$

### B. Overview

HiF-DTA comprises four modules: drug extractor, protein extractor, fusion unit, and affinity predictor (Figure 1).

The drug extractor encodes SMILES into one-hot sequences and RDKit graphs, extracts global and local features via BiLSTM and PNA, applies tree decomposition for atomic, substructural, and molecular levels, and integrates them through multi-head attention into a multi-scale drug tensor.

The protein extractor derives ESM2 embeddings, enhances global semantics using Mamba, constructs a residue graph processed by PNA for local details, and refines the combined features via a two-layer GCN and MinCutPool to obtain residue-cluster representations.

The fusion unit interacts multi-scale drug and residue-cluster features through bilinear attention, updates substructural descriptors, and integrates multi-scale cluster representations via gated fusion into a unified embedding.

The affinity predictor performs attentive pooling on unified and substructural features, concatenates them, and predicts binding affinity via a fully connected network.

### C. Drug Feature Extraction Module

*1) Sequence Feature and Molecular Graphs Representation:* The module takes SMILES strings as input. Each atom is one-

TABLE I  
Statistical Analysis of Benchmark Datasets

Dataset	Drugs	Proteins	Affinities
Davis	68	442	30056
Metz	240	121	13669
KIBA	2111	229	118254

hot encoded into a 43-dimensional vector capturing connectivity, hydrogen count, hybridization and aromaticity, constituting the sequence feature  $\mathbf{h}_{\text{sequence}}$  [14], [15]. We also parse the SMILES strings using the RDKit toolkit into a molecular graph  $G = (V, E)$ .

*2) Global and Local Features Extraction:* For the global sequence feature, we map the sequence feature vector  $\mathbf{h}_{\text{sequence}}$  through two fully connected layers with two activation functions and a final normalization to obtain the initial feature  $\hat{\mathbf{h}}_{\text{Atom}}$  of the drug sequence.

$$\hat{\mathbf{h}}_{\text{Atom}} = \text{Norm}(\sigma(\sigma(\mathbf{h}_{\text{sequence}} W_1 + b_1) W_2 + b_2)) \quad (2)$$

where  $W_1$  and  $W_2$  are the weight matrices of the first and second fully connected layers, respectively, and  $b_1$ ,  $b_2$  are the corresponding bias vectors. The activation function  $\sigma(\cdot)$  denotes the ReLU nonlinearity, and  $\text{Norm}(\cdot)$  represents layer normalization applied to the final output. Then, we extract contextual feature  $\mathbf{h}_{\text{BiLSTM}}$  of the global sequence using a BiLSTM [16].

$$\mathbf{h}_{\text{BiLSTM}} = \text{BiLSTM}(\hat{\mathbf{h}}_{\text{Atom}}) \quad (3)$$

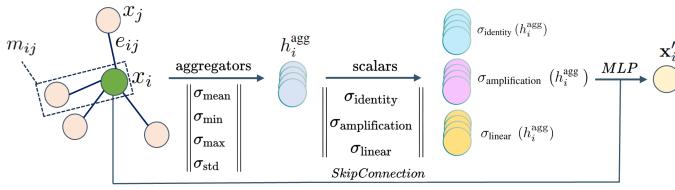


Fig. 2. Architecture of the Principal Neighbourhood Aggregation (PNA).

For local structural modeling, we use the Principal Neighbourhood Aggregation (PNA) [17] within a Message Passing Neural Network (MPNN) to capture atom interactions in molecular graphs (Fig. 2). For atom  $i$  and neighbor  $j$  with features  $x_i$ ,  $x_j$ , and bond feature  $e_{ij}$ , the message  $m_{ij}$  is computed by concatenating these features and projecting them into hidden dimension  $d$  through  $\text{MLP}_{\text{pre}}$ .

$$m_{ij} = \text{MLP}_{\text{pre}}([x_i \| x_j \| \phi(e_{ij})]) \quad (4)$$

where  $\|$  denotes feature concatenation operation, and  $\phi(\cdot)$  is an edge encoder mapping edge features to the hidden dimension. The local feature of node  $i$  aggregates messages  $m_{ij}$  from neighbors  $j$ , followed by scaling with aggregators  $A = \{\sigma_{\text{mean}}, \sigma_{\text{min}}, \sigma_{\text{max}}, \sigma_{\text{std}}\}$  and scalers  $S = \{\sigma_{\text{identity}}, \sigma_{\text{amplification}}, \sigma_{\text{linear}}\}$ , as:

$$h_i = \bigoplus_{s \in S} s \left( \bigoplus_{a \in A} a(\{m_{ij} \mid j \in \mathcal{N}(i)\}) \right) \quad (5)$$

where  $\bigoplus$  is the concatenation operation, and  $\mathcal{N}(i)$  denotes the aggregation of the set of neighbor nodes of node  $i$ .

The original features  $x_i$  are concatenated with the local features  $h_i$ , followed by a nonlinear transformation through the  $\text{MLP}_{\text{post}}$ , which reduces the dimensionality to  $d$ . A subsequent linear projection is then applied to generate the final node representation  $x'_i$ :

$$x'_i = W \cdot \text{MLP}_{\text{post}}([x_i \| h_i]) \quad (6)$$

Through message passing, each node representation is updated to yield the local atom feature  $\text{hMPNN}$ . This feature is concatenated with the global atom feature  $\text{hBiLSTM}$  and refined via linear mapping, activation, and layer normalization to form the final atom feature  $\text{h}_{\text{Atom}}$ .

$$\text{h}_{\text{Atom}} = \text{Norm}(\sigma(W[\text{h}_{\text{MPNN}} \| \text{h}_{\text{BiLSTM}}] + b)) \quad (7)$$

where,  $\text{Norm}(\cdot)$  represents layer normalization applied to the final output,  $\sigma(\cdot)$  denotes the ReLU activation function,  $W$  is the weight matrix that projects the concatenated features from dimension  $2d$  down to  $d$ , and  $b$  represents the bias vector.

3) *Multi-scale Feature Extraction*: The tree decomposition algorithm [18] partitions the molecular graph into substructures, then based on the atom-to-substructure mapping  $\text{map}_{v \rightarrow c}$ , the substructural features are updated by performing mean pooling over the atomic features  $\tilde{H}_c$ :

$$\tilde{H}_c = \text{MeanPool}(\text{h}_{\text{Atom}}, \text{map}_{v \rightarrow c}) \in R^{C \times d} \quad (8)$$

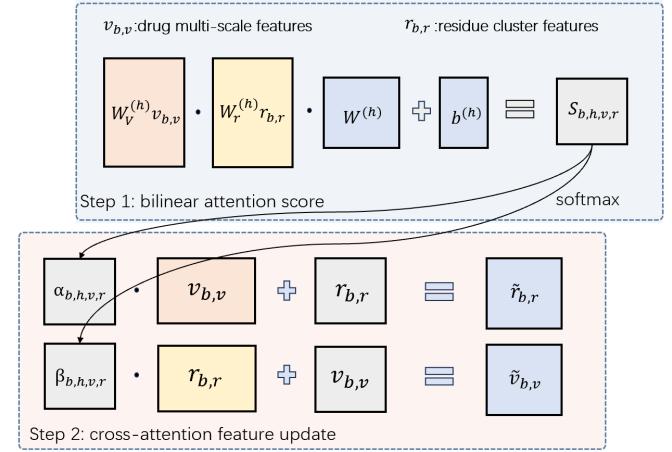


Fig. 3. Illustration of the Multi-scale Feature Fusion Module.

MeanPool averages features over  $C$  substructures. Type-based label embeddings first yield initial substructural features  $\mathbf{h}_{\text{Group}}$ ;  $\tilde{H}_c$  then passes through a linear layer plus activation and joins  $\mathbf{h}_{\text{Group}}$  via a residual link to produce updated substructural features  $H_c$ .

$$H_c = \mathbf{h}_{\text{Group}} + \text{ReLU}(W \cdot \tilde{H}_c) \in R^{C \times d} \quad (9)$$

$H_c$  is reshaped into  $h$  attention heads of dimension  $d/h$ , and each head computes an attention score  $\alpha_i$  via an independent multi-layer perceptron:

$$\alpha_i = \text{MLP}_i(H_c^{(i)}) \in \mathbb{R}^{C \times 1}, \quad i = 1, \dots, h \quad (10)$$

The attention scores are regularized with dropout and normalized per-molecule via batch-wise  $\text{batch}_c$  and Softmax to obtain  $\hat{\alpha}_i$ :

$$\hat{\alpha}_i = \text{Softmax}(\text{Dropout}(\alpha_i), \text{batch}_c) \quad (11)$$

The attention weights from all heads are concatenated to form a unified attention vector  $\hat{\alpha}$ :

$$\hat{\alpha} = [\hat{\alpha}_1 \| \hat{\alpha}_2 \| \cdots \| \hat{\alpha}_h] \quad (12)$$

This attention vector is applied to the substructural features via element-wise multiplication. A batch-wise weighted sum pooling is then performed to generate the fused molecular representation:

$$\text{h}_{\text{mol}} = \text{ScatterSum}(\hat{\alpha} \odot H_c, \text{batch}_c) \quad (13)$$

where  $\odot$  denotes element-wise multiplication and ScatterSum aggregates the weighted features within each batch by summing them according to the batch indices.

TABLE II  
Hyperparameter Settings

Hyper-parameters	Setting
Learning rate	{1e-3, 5e-4}
Batch size	64
Epoch	400
Optimizer	Adam
total_layer	3
hidden_channels	200
number of residue clusters per layer	[5, 10, 20]
Heads number of the Bilinear cross-attention	4
Heads of Drug Multi-scale feature updating	4
Dropout of Drug Multi-scale feature updating	0.2

TABLE III  
Comparison Results on the Davis Benchmark Dataset

Method	CI $\uparrow$	$r_m^2 \uparrow$	PCC $\uparrow$	MSE $\downarrow$
GraphDTA	0.888	0.699	0.8475	0.232
AttentionDTA	0.8947	0.7404	0.8721	0.1912
ColdDTA	0.8938	0.7606	0.8861	0.1695
AttentionMGT-DTA	0.891	0.699	-	0.193
GOaidDTA	0.891	0.654	0.85	0.229
TF-DTA	0.8856	0.6703	-	0.2312
TEFDTA	0.8925	0.7403	0.8617	0.21
DGDTA	0.899	0.702	-	0.225
PSICHIC	0.884	0.7262	0.8802	0.1783
HiF-DTA	<b>0.9026</b>	<b>0.762</b>	<b>0.8921</b>	<b>0.1654</b>

#### D. Protein Feature Extraction Module

1) *Sequence Feature and protein Graphs Representation:* ESM2 33rd-layer embeddings of size  $R \times 1280$  are taken as evolutionary features. Thresholded ESM2 contact probabilities define the protein graphs  $G = (R, L)$  with residues as nodes and contacts as edges. Sequence features concatenate these embeddings with a residue matrix that combines 21-dimensional one-hot encoding and twelve physicochemical descriptors. The contact matrix is converted to edge features through radial basis function mapping.

2) *Global and Local Features Extraction:* Each protein node is associated with a batch index denoted as  $\text{batch}_R$ . First, the nodes are reordered based on their node degree  $d_i$  and batch index  $\text{batch}_R$ , and the resulting permutation is recorded as  $\pi$ :

$$\pi = \text{argsort}(\text{array}[d_i, \text{batch}_R]) \quad (14)$$

The reordered node features are obtained, and the sparse node features are converted into dense matrices  $H$  according to batches:

$$H \in \mathbb{R}^{B \times V_m \times d} \quad (15)$$

where  $V_m$  is the maximum number of nodes within a batch, and missing positions are filled with masks. Based on the Mamba structured state space model, the global interaction features  $H'$  of nodes are computed as:

$$H' = \text{Mamba}(H) \quad (16)$$

Then, the sparse valid nodes are extracted, and the inverse permutation is applied to restore the original node order:

$$\mathbf{R}_{\text{Mamba}} = H'[\text{mask}]_{\pi^{-1}} \quad (17)$$

TABLE IV  
Comparison Results on the Metz Benchmark Dataset

Method	CI $\uparrow$	$r_m^2 \uparrow$	PCC $\uparrow$	MSE $\downarrow$
GraphDTA	0.8621	0.7079	0.8548	0.1714
ArKDTA	0.843	-	-	0.1703
AttentionDTA	0.8755	0.7048	0.8565	0.1612
ColdDTA	0.8738	0.7116	0.8622	0.1553
TEFDTA	0.8445	0.6171	0.8376	0.1873
PSICHIC	0.8751	0.7234	0.8688	0.1469
HiF-DTA	<b>0.8831</b>	<b>0.7486</b>	<b>0.8774</b>	<b>0.1369</b>

TABLE V  
Comparison Results on the KIBA Benchmark Dataset

Method	CI $\uparrow$	$r_m^2 \uparrow$	PCC $\uparrow$	MSE $\downarrow$
GOaidDTA	0.876	0.706	0.868	0.179
AttentionDTA	0.8799	0.735	0.8739	0.1668
ColdDTA	0.8689	0.7054	0.8671	0.1762
TF-DTA	0.8768	0.7344	-	0.1771
TEFDTA	0.8675	0.7065	0.8546	0.1864
AttentionMGT-DTA	0.893	0.786	-	0.14
PSICHIC	0.8816	0.7571	0.8783	0.1556
HiF-DTA	<b>0.8948</b>	<b>0.7869</b>	<b>0.8947</b>	<b>0.1387</b>

For local structure we reuse the atomic-graph MPNN to yield residue features, concatenate them with Mamba outputs, and fuse the pair through linear projection, ReLU and layer normalization to obtain the unified residue feature:

$$\mathbf{R}_{\text{residue}} = \text{Norm}(\sigma(W[\mathbf{R}_{\text{MPNN}} || \mathbf{R}_{\text{Mamba}}] + b)) \quad (18)$$

where  $\text{Norm}(\cdot)$  represents layer normalization applied to the final output,  $\sigma(\cdot)$  denotes the ReLU activation function,  $W$  is the weight matrix that projects the concatenated features from dimension  $2d$  down to  $d$ , and  $b$  represents the bias vector.

3) *Residue Cluster Clustering:* A two-layer GCN refines  $\mathbf{R}_{\text{residue}}$ , softmax with batch padding outputs assignment matrix  $M \in \mathbb{R}^{B \times R_m \times cl}$ , and dense mincut pooling algorithm pools [19]  $\mathbf{R}_{\text{residue}}$  via  $M$  to yield the residue cluster feature matrix  $\mathbf{R}_{\text{cluster}} \in \mathbb{R}^{B \times cl \times d}$ .

#### E. Multi-scale Feature Fusion Module

Thereafter, residue cluster features  $\mathbf{R}_{\text{cluster}}$  and drug multi-scale features enter a multi-head bilinear cross-attention fusion network shown in Fig. 3.

For each attention head  $h$ , residue cluster features  $r_{b,r}$  and drug multi-scale features  $v_{b,v}$  from the  $b$ -th sample are first projected into a shared latent space using learnable projection matrices  $W_r^{(h)}$  and  $W_v^{(h)}$ , respectively.

$$\hat{r}_{b,r} = W_r^{(h)} r_{b,r} \quad \hat{v}_{b,v} = W_v^{(h)} v_{b,v} \quad (19)$$

For each attention head  $h$ , the projected drug multi-scale features  $\hat{v}_{b,v}$  and residue cluster features  $\hat{r}_{b,r}$  are multiplied element-wise. These are further weighted by a learnable per-channel scalar vector  $W^{(h)} \in \mathbb{R}^{d \times k}$  to obtain the bilinear attention score,  $k$  denotes the compression factor of the low-rank pooling:

$$S_{b,h,v,r} = \hat{v}_{b,v} \cdot \hat{r}_{b,r} \cdot W^{(h)} + b^{(h)} \quad (20)$$

TABLE VI  
Ablation Results on Different Multi-scale Features and Different Feature Fusion Strategies

Multi-scale Features			Fusion Strategy			Evaluation Metrics			
Molecular	substructural	Atomic	Bilinear Attention	Concatenation	Addition	CI ↑	$r_m^2 \uparrow$	PCC ↑	MSE ↓
✓	✓	✓	✗	✓	✗	0.8733	0.7205	0.8551	0.1691
✓	✓	✓	✗	✗	✓	0.8544	0.6867	0.8371	0.1753
✓	✗	✗	✓	✗	✗	0.8754	0.7363	0.871	0.1483
✗	✓	✗	✓	✗	✗	0.8793	0.738	0.8721	0.1433
✗	✗	✓	✓	✗	✗	0.869	0.7374	0.8591	0.1481
✓	✓	✓	✓	✗	✗	<b>0.8831</b>	<b>0.7486</b>	<b>0.8774</b>	<b>0.1369</b>

TABLE VII  
Ablation Results on Different Drug Feature Representation Network

Drug Representation	CI ↑	$r_m^2 \uparrow$	PCC ↑	MSE ↓
Global Features Only	0.8545	0.6867	0.837	0.175
Local Features Only	0.878	0.725	0.866	0.149
Both Global & Local Features	<b>0.8831</b>	<b>0.7486</b>	<b>0.8774</b>	<b>0.1369</b>

TABLE VIII  
Ablation Results on Different Protein Feature Representation Network

Protein Representation	CI ↑	$r_m^2 \uparrow$	PCC ↑	MSE ↓
Global Features Only	0.8754	0.6854	0.8593	0.1563
Local Features Only	0.8809	0.7447	0.8681	0.1492
Both Global & Local Features	<b>0.8831</b>	<b>0.7486</b>	<b>0.8774</b>	<b>0.1369</b>

where  $b^{(h)}$  is a learnable bias scalar. The attention weights  $\alpha_{b,h,v,r}$  are then obtained by applying a softmax function across all residue clusters.

$$\alpha_{b,h,v,r} = \text{softmax}_r(S_{b,h,v,r}) = \frac{\exp(S_{b,h,v,r})}{\sum_{r'} \exp(S_{b,h,v,r'})} \quad (21)$$

$$\tilde{r}_{b,r} = r_{b,r} + \sum_v \alpha_{b,h,v,r} \cdot v_{b,v} \quad (22)$$

The attention weights  $\beta_{b,h,v,r}$  are then obtained by applying a softmax function across all drug multi-scale features. Using these attention weights, the multi-scale feature of each drug is updated by aggregating information from all residue clusters, followed by a residual connection to retain its original features:

$$\beta_{b,h,v,r} = \text{softmax}_v(S_{b,h,v,r}) = \frac{\exp(S_{b,h,v,r})}{\sum_{v'} \exp(S_{b,h,v',r})} \quad (23)$$

$$\tilde{v}_{b,v} = v_{b,v} + \sum_r \beta_{b,h,v,r} \cdot r_{b,r} \quad (24)$$

The interaction is repeated independently at the atomic, substructural and molecular levels of the drug to produce three updated residue-cluster features  $r_{b,r}^{(\text{atom})}$ ,  $r_{b,r}^{(\text{substructure})}$  and  $r_{b,r}^{(\text{drug})}$ , which are then fused into the final residue-cluster feature  $\tilde{R}_{b,r}$  through softmax-weighted averaging.

$$\tilde{R}_{b,r} = a_1 \tilde{r}_{b,r}^{(\text{atom})} + a_2 \tilde{r}_{b,r}^{(\text{substructure})} + a_3 \tilde{r}_{b,r}^{(\text{drug})} \quad (25)$$

#### F. Affinity Prediction Module

The affinity prediction module integrates multi-scale protein and drug features using attention-based pooling and an MLP.

Protein attention pooling computes residue-level attention via residue-cluster weights and the assignment matrix, aggregates residue features into a global representation, and refines it through a two-layer MLP.

For drugs, multi-scale attention across molecular, substructural, and atomic levels generates atomic attention scores, which are aggregated and transformed by a two-layer MLP into the final drug vector.

The resulting protein and drug vectors are concatenated and passed to an MLP to predict the binding affinity score.

## III. EXPERIMENT AND RESULT

### A. Experiment Setting

To ensure reliable results, five-fold cross-validation was applied. The learning rate started at 0.001 and decayed to 0.0005 after 100 epochs. All datasets used a batch size of 64 and were trained for up to 400 epochs, with early stopping if no loss reduction occurred within 100 epochs. Hyperparameter details are summarized in Table II.

### B. Evaluation Metrics

Since DTA prediction is a regression task, we adopt the Concordance Index (CI), Modified Correlation Coefficient ( $r_m^2$ ), Pearson Correlation Coefficient (PCC), and Mean Squared Error (MSE) as evaluation metrics. These metrics quantify both ranking consistency and absolute deviation.

### C. Experimental Results

We compared our method with GraphDTA [10], AttentionDTA [20], ColdDTA [21], AttentionMGT-DTA [22], GOaidDTA [23], TF-DTA [24], TEFDTA [25], DGDTA [26], PSICHIC [14], and ArKDTA [27].

As shown in Tables III, IV, and V, HiF-DTA achieves the best performance across all three benchmark datasets. It consistently ranks highest across all four evaluation metrics: CI,  $r_m^2$ , PCC, and MSE.

On Davis, HiF-DTA achieves a CI of 0.9026—the first reported result to surpass the 0.9 threshold. On Metz, it lowers the MSE to 0.1369, outperforming the previous best benchmark by 1%. On the large-scale and highly heterogeneous KIBA benchmark, HiF-DTA further establishes state-of-the-art performance with a PCC of 0.8947.

#### D. Ablation Experiments

We assessed the effect of different drug feature extraction networks on DTA prediction using the Metz dataset under three settings: global only, local only, and combined global-local features. Results in Table VII show that integrating both feature types yields superior performance. Similarly, Table VIII compares protein extraction architectures under the same settings, confirming that joint modeling of global semantic and local structural features consistently improves prediction accuracy.

We further examined the contribution of multi-scale features—atomic, substructural, and molecular levels—and compared fusion strategies including bilinear attention, concatenation, and addition. As shown in Table VI, our multi-scale fusion module achieves the best results across all metrics, with the substructural level performing best among single-scale variants.

## IV. CONCLUSION

This paper proposed HiF-DTA, a hierarchical feature learning network for DTA prediction. By adopting a dual-pathway encoding strategy, HiF-DTA integrates global sequence semantics and local structural features of drugs and proteins. It further captures multi-scale drug–target interactions (atomic, substructural, and molecular levels) and fuses them via a bilinear attention mechanism to enhance interaction representation. Experiments on three benchmark datasets show that HiF-DTA consistently surpasses state-of-the-art baselines, highlighting its potential in computational drug discovery.

## REFERENCES

- [1] Jonathan Greer, John W Erickson, John J Baldwin, and Michael D Varney. Application of the three-dimensional structures of protein target molecules in structure-based drug design. *Journal of medicinal chemistry*, 37(8):1035–1054, 1994.
- [2] Márcio Dorn, Mariel Barbachan e Silva, Luciana S Buriol, and Luis C Lamb. Three-dimensional protein structure prediction: Methods and computational strategies. *Computational biology and chemistry*, 53:251–276, 2014.
- [3] DB Wetlauffer and S Ristow. Acquisition of three-dimensional structure of proteins. *Annual review of biochemistry*, 42(1):135–158, 1973.
- [4] Peijin Guo, Minghui Li, Hewen Pan, Ruixiang Huang, Lulu Xue, Shengqing Hu, Zikang Guo, Wei Wan, and Shengshan Hu. Multi-modality representation learning for antibody-antigen interactions prediction. *arXiv preprint arXiv:2503.17666*, 2025.
- [5] Minghui Li, Yao Shi, Shengqing Hu, Shengshan Hu, Peijin Guo, Wei Wan, Leo Yu Zhang, Shirui Pan, Jizhou Li, Lichao Sun, et al. Mvsfab: accurate antibody-antigen binding affinity prediction via multi-view sequence feature learning. *Bioinformatics*, 41(5):btae579, 2025.
- [6] Minghui Li, Zikang Guo, Yang Wu, Peijin Guo, Yao Shi, Shengshan Hu, Wei Wan, and Shengqing Hu. Vidta: Enhanced drug-target affinity prediction via virtual graph nodes and attention-based feature fusion. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 42–47. IEEE, 2024.
- [7] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.
- [8] Zhaohan Meng, Zaiqiao Meng, Ke Yuan, and Iadh Ounis. Fusiondta: Fine-grained binding discovery with token-level fusion for drug-target interaction. *arXiv preprint arXiv:2406.01651*, 2024.
- [9] Jiannuo Li and Lan Yao. Hcaf-dta: drug-target binding affinity prediction with cross-attention fused hypergraph neural networks. *arXiv preprint arXiv:2504.02014*, 2025.
- [10] Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.
- [11] Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.
- [12] James T Metz, Eric F Johnson, Niru B Soni, Philip J Merta, Lemma Kifle, and Philip J Hajduk. Navigating the kinome. *Nature chemical biology*, 7(4):200–202, 2011.
- [13] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of chemical information and modeling*, 54(3):735–743, 2014.
- [14] Huan Yee Koh, Anh TN Nguyen, Shirui Pan, Lauren T May, and Geoffrey I Webb. Psichic: physicochemical graph neural network for learning protein-ligand interaction fingerprints from sequence data. *BioRxiv*, pages 2023–09, 2023.
- [15] Peijin Guo, Minghui Li, Hewen Pan, Bowen Chen, Yang Wu, Zikang Guo, Leo Yu Zhang, Shengshan Hu, and Shengqing Hu. Uncertainty-aware metabolic stability prediction with dual-view contrastive learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 300–316. Springer, 2025.
- [16] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. The performance of lstm and bilstm in forecasting time series. In *2019 IEEE International conference on big data (Big Data)*, pages 3285–3292. IEEE, 2019.
- [17] Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. *Advances in neural information processing systems*, 33:13260–13271, 2020.
- [18] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.
- [19] Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral clustering with graph neural networks for graph pooling. In *International conference on machine learning*, pages 874–883. PMLR, 2020.
- [20] Qichang Zhao, Guihua Duan, Mengyun Yang, Zhongjian Cheng, Yao-hang Li, and Jianxin Wang. Attentiondta: Drug–target binding affinity prediction by sequence-based deep learning with attention mechanism. *IEEE/ACM transactions on computational biology and bioinformatics*, 20(2):852–863, 2022.
- [21] Kejie Fang, Yiming Zhang, Shiyu Du, and Jian He. Colddta: utilizing data augmentation and attention-based feature fusion for drug–target binding affinity prediction. *Computers in Biology and Medicine*, 164:107372, 2023.
- [22] Hongjie Wu, Junkai Liu, Tengsheng Jiang, Quan Zou, Shujie Qi, Zhiming Cui, Prayag Tiwari, and Yijie Ding. Attentionmgt-dta: A multi-modal drug–target affinity prediction using graph transformer and attention mechanism. *Neural Networks*, 169:623–636, 2024.
- [23] Lingling Zhao, Peijin Xie, Lingfeng Hao, Tiantian Li, and Chunyu Wang. Gene ontology aided compound protein binding affinity prediction using bert encoding. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1231–1236. IEEE, 2020.
- [24] Wenjun Li, Yiqiang Zhou, and Xiwei Tang. Tf-dta: A deep learning approach using transformer encoder to predict drug–target binding affinity. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 418–421. IEEE, 2023.
- [25] Zongquan Li, Pengxuan Ren, Hao Yang, Jie Zheng, and Fang Bai. Tefdta: a transformer encoder and fingerprint representation combined prediction method for bonded and non-bonded drug–target affinities. *Bioinformatics*, 40(1):btad778, 2024.
- [26] Haixia Zhai, Hongli Hou, Junwei Luo, Xiaoyan Liu, Zhengjiang Wu, and Junfeng Wang. Dgdt: dynamic graph attention network for predicting drug–target binding affinity. *BMC bioinformatics*, 24(1):367, 2023.
- [27] Mogan Gim, Junseok Choe, Seungheun Baek, Jueon Park, Chaeun Lee, Minjae Ju, Sumin Lee, and Jaewoo Kang. Arkdta: attention regularization guided by non-covalent interactions for explainable drug–target binding affinity prediction. *Bioinformatics*, 39(Supplement\_1):i448–i457, 2023.