Expert Systems With Applications 306 (2026) 130918

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

MFCLDTA: Multi-scale feature contrastive learning for predicting drug-target binding affinity

Zhen Tian [a,b], Saisai Zhu [a,b], Zhixia Teng [c], Xiaoqiang Yan [a,*], Tao Wang [d,*]

[a] School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, 450001, China
[b] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, 324000, China
[c] College of Computer and Control Engineering, Northeast Forestry University, Harbin, 150001, China
[d] School of Computer Science, Northwestern Polytechnical University, Xi'an, 710072, China

ARTICLE INFO

Keywords:
Drug-target binding affinity
Drug discovery
Multi-scale features
Contrastive learning

ABSTRACT

Accurate prediction of drug-target binding affinity (DTA) is crucial for accelerating drug discovery and development. Although deep learning-based approaches have demonstrated remarkable success in DTA prediction, current methods usually suffer from two key limitations: (1) inadequate utilization of multi-scale feature information, leading to suboptimal representations of drugs and targets; and (2) challenges in effectively balancing features across scales for robust feature fusion. In this study, we propose MFCLDTA, a novel deep learning-based framework for DTA prediction through the multi-scale feature contrastive learning mechanism. The proposed approach captures drug and target features across three scales: molecular sequence, molecular structure, and affinity graph. Through the multi-scale contrastive learning mechanism, MFCLDTA maximizes mutual information between different scales while ensuring robust feature alignment. Integration of these multi-scale representations achieves superior DTA prediction performance. Comprehensive benchmarks on Davis and KIBA datasets demonstrate MFCLDTA consistently outperforms state-of-the-art baseline approaches. Ablation studies confirm the critical importance of both multi-scale feature integration and contrastive learning in enhancing prediction accuracy. The source code of MFCLDTA is available at https://github.com/ssaixiansheng/MFCLDTA.

1. Introduction

Drug discovery is a highly resource-intensive and time-consuming process with a low success rate, focused on identifying therapeutic agents that modulate specific protein-protein interactions for disease treatment (Takebe et al., 2018). Conventional approaches require substantial human and financial investment, with the development of a new drug costs $2.6 billion and exceeding 10 years (Prasad et al., 2017; Wouters et al., 2020). In this context, computer-aided drug discovery (CADD) has emerged as a pivotal approach in modern drug development, offering substantial reductions in both development cycles and costs (Öztürk et al., 2018). As a critical component of this process, DTA prediction has been extensively studied as an essential phase in drug discovery and development (Abbasi et al., 2021; Ezzat et al., 2019; Qian et al., 2022).

Recent decades have witnessed remarkable progress in DTA prediction methodologies, driven by rapid advancements in computational technologies. Traditional computational approaches, including molecular docking and molecular dynamics simulations, predict binding affinities by modeling intermolecular interactions (Hollingsworth & Dror, 2018; Kukol et al., 2008; Shoichet et al., 1999; Trott & Olson, 2010). However, these methods typically require extensive datasets of known ligand-protein complexes or complete 3D structural information of target proteins (Wang et al., 2021a). In addition, early machine learning techniques, such as support vector machines and random forests attempt to predict DTA through matrix-based computations, but remain constrained by computationally intensive processes and dependence on elaborate feature engineering (He et al., 2017; Li et al., 2015; Pahikkala et al., 2015).

The advent of artificial intelligence has revolutionized drug discovery through data-driven approaches, particularly deep learning methods that have demonstrated remarkable success in this domain (Yu et al., 2017). These approaches can be broadly classified into structure-based and non-structure-based paradigms (Thafar et al., 2019). Structure-based approaches utilize three-dimensional (3D) structural data of drug-target complexes. They usually employ advanced architectures, such as 3D convolutional neural networks (CNNs) (Wang et al., 2021b) and fully connected neural networks (FCNNs) (Zhu et al., 2020), to extract

* Corresponding authors.
E-mail addresses: ieztian@zzu.edu.cn (Z. Tian), zss_zzu@gs.zzu.edu.cn (S. Zhu), tengzhixia@nefu.edu.cn (Z. Teng), iexqyan@zzu.edu.cn (X. Yan), twang@nwpu.edu.cn (T. Wang).

https://doi.org/10.1016/j.eswa.2025.130918
Received 24 July 2025; Received in revised form 7 December 2025; Accepted 17 December 2025
Available online 21 December 2025

hierarchical features and model molecular interactions, thus improving prediction performance (Stepniewska-Dziubinska et al., 2018; Zhang et al., 2024). Nevertheless, the requirement for atomic-resolution 3D structures of both drug molecules and target proteins presents a substantial technical challenge. In contrast, non-structure-based deep learning methods primarily fall into two categories: sequence-based and graph-based approaches (Yu et al., 2024).

Sequence-based approaches always adopt the SMILES strings of drugs and target sequences as inputs, extracting biological sequence features through complex models. For example, DeepDTA employs CNNs to capture local sequence patterns from both SMILES strings and target sequences, streamlining the feature extraction pipeline (Öztürk et al., 2018). AttentionDTA enhances this approach by integrating CNNs with a bilateral multi-head attention mechanism, effectively fusing drug and target features to improve predictive performance (Zhao et al., 2019). Further advancing this paradigm, FusionDTA combines bidirectional long short-term memory (BiLSTM) networks with multi-head linear attention mechanisms, enabling knowledge distillation through attention-weighted global information aggregation (Graves & Graves, 2012; Yuan et al., 2022). The MFR-DTA framework introduces dual feature extractors to better characterize sequence specific attributes (Hua et al., 2023). Nevertheless, although these methods perform well in processing sequence data, their predictive accuracy remains constrained by the inability to incorporate molecular structural information of drugs and targets (Yang et al., 2022).

Graph-based methods typically convert sequence information into molecular graph representations to capture their structure-based features. As a pioneering framework, GraphDTA utilizes RDKit to convert drug SMILES strings into molecular graphs, subsequently encoding structural features through graph convolutional network (GCN) layers (Lovrić et al., 2019; Nguyen et al., 2021). DGraphDTA extends this approach by integrating both drug molecular graphs and target structure graphs, employing Pconsc4 for target graph construction to capture richer structural information (Jiang et al., 2020; Michel et al., 2019). The MSGNN-DTA model implements the end-to-end contact graph generation with the ESM model and further enhances feature representation through motif graph integration (Li et al., 2023; Wang et al., 2023). MgraphDTA combines multi-layer graph neural networks with convolutional neural networks for comprehensive structural feature extraction, achieving superior DTA prediction accuracy (Yang et al., 2022). Notably, GLGN-DTA incorporates a learnable soft adjacency matrix module within the graph architecture, facilitating more efficient refinement of molecular graph contextual structures (Qi et al., 2024). While graph-based methods outperform sequence-based approaches, their exclusive reliance on molecular structure confines them to inherently limited single-scale data. (Zhang et al., 2023b).

Compared to single-scale feature extraction approaches, multi-scale feature fusion methods have demonstrated superior performance by effectively capturing cross-scale correlations, thereby enhancing both model generalizability and prediction accuracy (Atrey et al., 2010). For instance, CrossAttentionDTI employs cross-modal attention mechanisms to enable feature interaction between drug molecular graphs and target protein sequences, while cross-attention modules align these two modalities (Zeng et al., 2022). HGRL-DTA combines coarse-grained representations derived from affinity graph with fine-grained representations extracted from molecular graphs, enhancing feature discriminability (Chu et al., 2022). AttentionMGT-DTA integrates protein sequences with 3D structural information via cross-attention, constructing a cross-scale drug-target interaction model that significantly improves performance (Wu et al., 2024). Additionally, PocketDTA utilizes pre-trained models combined with 3D binding pocket information to improve both prediction accuracy and model interpretability (Zhao et al., 2024). While integrating multi-scale features improves performance, these methods remain constrained by the deficient integration of richer-scale information, limiting robust feature representations and resulting in incomplete characterization of complex drug-target systems.

In recent years, contrastive learning has advanced the prediction of biological entity associations, demonstrating substantial progress (Wei et al., 2023). For example, FMDTA employs contrastive learning to balance feature information between string and graph modalities, demonstrating superior performance (Zhang et al., 2023a). CSCo-DTA implements a contrastive learning framework to maximize the mutual information between network-scale and molecular-scale representations, effectively uncovering their intrinsic relationships (Wang et al., 2024). Meanwhile, GraphCL-DTA advances the field by adopting contrastive learning for embedding space augmentation, surpassing traditional dropout-based strategies while better preserving molecular graph semantics (Yang et al., 2024). As a powerful self-supervised learning paradigm, contrastive learning operates by maximizing mutual information between positive pairs while distancing negative samples in the latent space (Liu et al., 2022). Despite these advances, contrastive learning remains underexplored for multi-scale feature fusion. To bridge this gap, we introduce a framework that systematically integrates drug-target features across three different scales, yielding their comprehensive and highly robust representations.

To address these limitations, here we present MFCLDTA, a novel drug-target binding affinity prediction framework that leverages the contrastive learning mechanism to integrate multi-scale features: molecular sequences, molecular structures, and affinity graph. The proposed model mainly consists of three main components: (1) the multi-scale feature extraction module, (2) the multi-scale feature contrastive learning module, and (3) the prediction module. Specifically, we construct molecular graphs from raw sequence inputs, which convert drug SMILES strings via RDKit (Lovrić et al., 2019) and transform target protein sequences using Pconsc4 (Michel et al., 2019). These molecular entities are then represented as nodes in an affinity graph, where experimentally verified binding affinities serve as weighted edges. For feature extraction, graph convolutional networks (GCNs) extract structural features from molecular and affinity graph, while bidirectional LSTMs (BiLSTMs) model sequential patterns in drug SMILES strings and protein sequences. The proposed multi-scale contrastive learning mechanism integrates features across modalities of drugs and targets, enabling comprehensive representation learning for accurate affinity prediction. Comprehensive evaluations demonstrate that MFCLDTA outperforms current state-of-the-art methods across benchmark datasets (Davis et al., 2011; Tang et al., 2014). The main contributions of our work are summarized as follows:

- We propose the multi-scale feature fusion framework for Drug-Target Affinity (DTA) prediction, integrating molecular sequence, molecular structure, and affinity graph information simultaneously for both molecular entities.
- MFCLDTA incorporates the multi-scale feature contrastive learning mechanism that maximizes mutual information among features at different scales, facilitating robust alignment and fusion.
- Benchmark evaluations demonstrate that MFCLDTA establishes state-of-the-art performance across all evaluation metrics.

2. Methods

In this section, we will present the detailed architecture of MFCLDTA. As illustrated in Fig. 1, the architecture comprises three key components: (A) the multi-scale feature extraction module, (B) the multi-scale feature contrastive learning module, and (C) the DTA prediction module. The multi-scale feature extraction module integrates and aligns three distinct scales: molecular sequence, molecular structure, and affinity graph. Subsequently, the multi-scale feature contrastive learning module aligns these representations through the feature space optimization. Finally, MFCLDTA infers the potential DTA with the learned representations of drugs and targets.

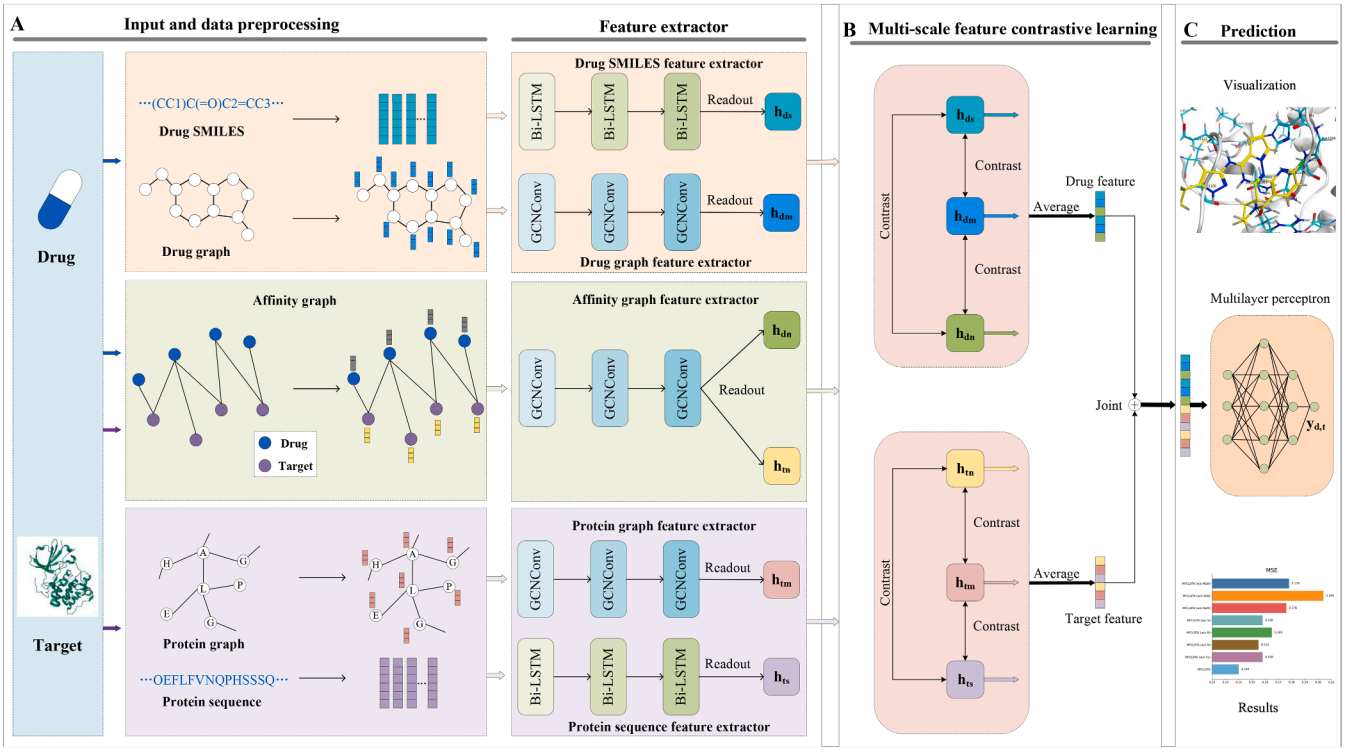

Fig. 1. The overall framework of MFCLDTA. The framework is divided into three main components: (A) the multi-scale feature extraction module, (B) the multi-scale feature contrastive learning module, and (C) the prediction module. The framework accepts drug molecules represented as SMILES strings and target proteins as amino acid sequences. During the initial preprocessing stage, raw inputs are transformed into three complementary representations: molecular sequence, molecular structure, and affinity graph. A three-layer bidirectional LSTM (BiLSTM) network processes sequence-based features, while GCNs with three layers extract structural patterns from both molecular and affinity graph. The multi-scale feature contrastive learning module then optimizes cross-scale feature alignment by maximizing mutual information between different feature modalities. Finally, the integrated representations are fed into a multilayer perceptron (MLP) to predict binding affinity values.

Table 1
The brief summary of experimental datasets.

Dataset	Drug	Protein	Affinity
Davis	68	442	30,056
KIBA	2,111	229	118,254

2.1. Datasets

To comprehensively evaluate the predictive performance of MFCLDTA, we perform comprehensive experiments on two widely used benchmark datasets: Davis (Davis et al., 2011) and KIBA (Tang et al., 2014). The Davis dataset comprises 68 drugs and 442 targets with 30,056 dissociation constant K_d measurements. Following standard practice, we convert K_d values to $pK_d = -log_{10}(K_d/10^9) - 5$ metrics ranging from 0.0 to 5.8, where higher values indicate stronger binding affinity. The KIBA dataset initially contains 52,498 drug entries and 467 target entries. Following standard preprocessing procedures (He et al., 2017), we obtain a filtered dataset containing 2,111 drugs, 229 targets, and 118,254 affinity scores. Table 1 summarizes the details of the two datasets. We preprocess drug structures using RDKit (Lovrić et al., 2019) and target sequences using Pconsc4 (Michel et al., 2019), converting them into molecular graph representations for model input.

2.2. Multi-scale feature extraction module

2.2.1. Sequence scale feature extraction

Prior to feature extraction, we first preprocess the drug SMILES strings and protein sequences. For drug SMILES strings, we utilize the molecular pre-trained model chemBERTa-2 to generate corresponding features for each drug string (Ahmad et al., 2022). These features are then averaged to obtain the overall drug embedding. For protein sequences, we employ the classic ProtT5-XL-UniRef50 pre-trained model to extract protein sequence features (Elnaggar et al., 2021). Subsequently, we compute the average features of all residues for each protein to generate the holistic protein sequence representation.

In this study, we adopt a novel sequence-based representation learning framework for multi-scale feature generation of drug molecules and target proteins. To overcome the temporal modeling limitations of conventional recurrent neural networks, the BiLSTM is utilized to capture cross-temporal dependencies via dual-path temporal encoding (Graves & Graves, 2012). The framework processes both molecular SMILES strings and protein sequences as input representations through its innovative bidirectional temporal correlation mechanism. By incorporating dynamic gating operations, our architecture simultaneously extracts local structural patterns and global contextual features at each sequence position. Furthermore, the architecture facilitates knowledge transfer among similar molecular substructures through category-specific feature sharing mechanisms.

For clarity, Equations (1) to (6) below precisely describe the forward computation process of the BiLSTM at each time step. These equations define the update mechanisms for the input gate, forget gate, output gate, and cell state. The core idea is that the BiLSTM employs a series of controllable gates to selectively retain or forget long-term and short-term dependencies within sequences, thereby generating high-quality context-aware feature representations for drug and target sequences. For drug SMILES string d, the forward temporal feature extraction is formally expressed as:

$$i_t = \sigma(W_{id} d_t + W_{ih} h_{(ds,t-1)} + b_i) \quad (1)$$

$$f_t = \sigma(W_{fd} d_t + W_{fh} h_{(ds,t-1)} + b_f) \quad (2)$$

$$o_t = \sigma(W_{od} d_t + W_{oh} h_{(ds,t-1)} + b_o) \quad (3)$$

$$g_t = tanh(W_{gd} d_t + W_{gh} h_{(ds,t-1)} + b_g) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (5)$$

$$h_{(ds,t)} = o_t \odot tanh(c_t) \quad (6)$$

where $h_{(ds,t)}$ denotes the hidden state output of the forward LSTM for drug d at time step t, and d_t represents the input token at time step t. The gate activations (i_t, f_t, o_t) correspond to the input, forget, and output gates, respectively, while g_t indicates the memory cell activation. The cell state c_t integrates information from the previous state c_{t-1} through gating mechanisms. At each time step t, the hidden state $h_{(ds,t-1)}$ interacts with input d_t to update both the cell state c_t and the hidden state $h_{(ds,t)}$. Here, $\sigma(\cdot)$ denotes the sigmoid activation function, $tanh(\cdot)$ denotes the hyperbolic tangent activation function, $\odot$ denotes element-wise multiplication (Hadamard product), and W, b are trainable weight matrices and bias terms. The final output of the BiLSTM is generated by concatenating the hidden state of the forward LSTM $h_{(ds,t)}$ and that of the backward LSTM $h_{(ds,t')}$. For each drug SMILES string d, the final sequence-based feature can be expressed as:

$$h_{ds} = h_{(ds,t)} || h_{(ds,t')} \quad (7)$$

here, || represents the connection operation.

We tokenize all benchmark target sequences using the same feature extraction method applied to drug SMILES, then process them with a BiLSTM. For each target sequence t, the model generates a final feature representation, denoted as h_{ts}:

$$h_{ts} = h_{(ts,t)} || h_{(ts,t')} \quad (8)$$

2.2.2. Molecule structure graph scale feature extraction

Drug molecule structure graph feature extraction. To comprehensively represent both chemical properties and topological structures of drug molecules, we employ RDKit, an open-source cheminformatics toolkit, to transform SMILES strings into molecular structure graphs (Lovrić et al., 2019). By mapping each chemical atom to a graph node and defining chemical bonds as inter-node edges, we construct an undirected molecular graph $G_d = (V_d, E_d)$, where V_d and E_d represent the sets of atomic nodes and chemical bond edges, respectively. Our GCN architecture hierarchically learns atom-wise feature vectors by aggregating local chemical environments and abstracting global topological patterns, thus encoding chemically meaningful fingerprints and preserving molecular connectivity and spatial topology. For each atom v in the molecular graph G_d, the GCN-based feature extraction is formally expressed as:

$$h_v^l = \sigma(\sum_{i \in N(v)} \frac{1}{\sqrt{m_v m_i}} h_i^{l-1} W^l) \quad (9)$$

where $\sigma(\cdot)$ represents a nonlinear activation function(Glorot et al., 2011), h_v^l is the hidden state of atom v at l-th layer, h_i^{l-1} is the feature vector of neighbor node i at layer $l-1$, node $v \in V_d$, $K(v)$ represents the neighborhood of v, $N(v) = K(v) \cup v$ defines the extended neighborhood, m_v is equal to $|N(v)|$, which indicates the degree of node v in graph G_d, and W^l represents the weight matrix of the l-th layer.

After applying the GCN operation, we obtain atomic-level feature representations for each atom in the molecular graph. These atom features are then aggregated into a molecule-level representation using a readout function. Formally, let h_v^L denote the output of the final GCN layer for atom v. For each drug molecule d, the readout operation is defined as follows:

$$h_{dm} = \phi(\{h_v^L | v \in V_d\}) \quad (10)$$

where $\phi(\cdot)$ is a differentiable readout function, h_{dm} represents the feature extraction of drug d based on the molecular scale.

Target molecular structure feature extraction. To acquire comprehensive target protein representations at the molecular scale, we adopt a graph-based approach inspired by DGraphDTA to construct 2D molecular graphs (Jiang et al., 2020). The target sequences are initially processed using Pconsc4 to predict residue contact maps (Michel et al., 2019). In these contact maps, amino acid residues serve as graph nodes with edges created between residue pairs having Euclidean distances below 0.5. This graph construction preserves spatial relationships while transforming the sequence into a molecular graph $G_t = (V_t, E_t)$, where V_t and E_t denote the vertex (residue) and edge sets, respectively. Following our drug molecular graph processing approach, we implement a GCN to perform convolution operations on each residue node r in the target molecular graph G_t, effectively aggregating its structural information:

$$h_r^l = \sigma(\sum_{j \in N(r)} \frac{1}{\sqrt{m_r m_j}} h_j^{l-1} W^l) \quad (11)$$

where h_r^l is the hidden state of atom r at l-th layer, h_j^{l-1} is the feature vector of neighbor node j at layer $l-1$, node $r \in V_t$, $K(r)$ represents the neighborhood of r, $N(r) = K(r) \cup r$ defines the extended neighborhood, m_r is equal to $|N(r)|$, which indicates the degree of node r in graph G_t, and W^l represents the weight matrix of the l-th layer.

The final topological structure features for each target molecule are obtained through a readout function operation:

$$h_{tm} = \phi(\{h_r^L | r \in V_t\}) \quad (12)$$

2.2.3. Affinity graph scale feature extraction

Known drug-target interactions provide critical prior knowledge for binding affinity prediction (Nogales et al., 2022). However, current approaches that incorporate sequence-based and molecular-based features often underutilize available affinity data during network construction. We address this by building a drug-target affinity graph $G_a = (M, N, \epsilon)$, where M denotes the number of drug molecules, N represents the number of target molecules, and ϵ encodes experimentally validated binding affinities. A zero value indicates unknown interactions, corresponding to absent edges in the affinity graph G_a. This construction yields an adjacency matrix $A \in \mathbb{R}^{(|M+N||M+N|)}$ representing the complete drug-target interaction network. The network takes an initial features $O = \begin{bmatrix} O^d \\ O^t \end{bmatrix}$ as input, where $O^d \in \mathbb{R}^{|M| \times b}$ and $O^t \in \mathbb{R}^{|N| \times b}$ represent initial features for drugs and targets, respectively, with b denoting the feature dimension. We employ GCNs to perform feature extraction and aggregation from this network:

$$H_n = \sigma(\hat{A} \cdot \sigma(\hat{A} O W_q^1) W_q^2) \quad (13)$$

where $\hat{A}$ is obtained by Laplace normalization of the adjacency matrix A, $\hat{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, D is a diagonal matrix, W_q^1 and W_q^2 are trainable weight parameters. The feature matrix H_n contains the learned representations of all entities, with each row vector $H_n[i,:] \in \mathbb{R}^b$ encoding the representation of either a drug or a target. Following these processing steps, we obtain the affinity graph representations h_{dn} and h_{tn} for each drug and target, respectively.

2.3. Multi-scale feature contrastive learning module

Our study derives multi-scale feature representations of drug-target systems across three distinct scales: molecular sequence, molecular structure, and affinity graph. Although each scale offers unique and valuable information, inherent correlations exist among them. To integrate features from these scales and enhance representation quality, we propose a multi-scale feature contrastive learning framework. This framework effectively maximizes mutual information across different scales while enabling robust feature alignment and fusion.

As illustrated in Fig. 1(C), MFCLDTA introduces a multi-scale feature contrastive learning approach where we systematically perform pairwise contrastive learning between features from different scales to

mitigate inter-scale discrepancies. A critical aspect of contrastive learning involves the definition of positive and negative samples.

In the proposed framework, each drug or target at a given scale serves as an anchor point, and its counterparts from other scales are designated as positive samples, while unrelated entities are considered negative samples. To enhance the effectiveness of positive sample selection, we introduce a novel method based on three similarity metrics: (i) molecular-based similarity (S_{dm}) computed using PubChem tools (Bolton et al., 2008), reflecting the principle that functionally similar drugs exhibit structural similarities; (ii) sequence-based similarity (S_{ts}) calculated via the Smith-Waterman algorithm (Smith et al., 1981); and (iii) network-based similarity is derived from metapath-based structural similarity following the drug-target-drug schema, which yields drug similarity measures (S_{dn}) and target similarity measures (S_{tn}) (Wang et al., 2024). Each similarity score is normalized, and the three similarity measures are equally weighted. The drug similarity scores S_{dm} and S_{dn} are summed, and the target similarity scores S_{ts} and S_{tn} are summed. After summation, the results are sorted in descending order. In a training batch, for a given anchor sample, its positive samples are selected as the top K samples from the aforementioned descending-order ranking, while other samples in the same batch are considered negative samples.

The contrastive learning objective, implemented using InfoNCE loss (van den et al., 2018) enhances the consistency between anchor-positive pairs while increasing separation from negative samples. For drugs, this loss function is expressed as:

$$\begin{aligned}\mathcal{L}_d = &-\log \frac{\sum_{d_p\in P_{d_i}} \exp\left(s(h_{d_i s}, h_{d_p m})/\tau\right)}{\sum_{k=1}^{M} \exp\left(s(h_{d_i s}, h_{d_k m})/\tau\right)} \\ &-\log \frac{\sum_{d_p\in P_{d_i}} \exp\left(s(h_{d_i m}, h_{d_p n})/\tau\right)}{\sum_{k=1}^{M} \exp\left(s(h_{d_i m}, h_{d_k n})/\tau\right)} \\ &-\log \frac{\sum_{d_p\in P_{d_i}} \exp\left(s(h_{d_i n}, h_{d_p s})/\tau\right)}{\sum_{k=1}^{M} \exp\left(s(h_{d_i n}, h_{d_k s})/\tau\right)}\end{aligned} \tag{14}$$

where $h_{d_i s}$, $h_{d_i m}$, and $h_{d_i n}$ represent the sequence-scale, molecular structure-scale, and affinity graph-scale features of drug d_i, respectively. P_{d_i} denotes the positive sample set for drug d_i, M is the total number of drug molecules, $s(\cdot)$ represents the cosine similarity function, and τ is the temperature parameter. The target contrastive loss is defined as:

$$\begin{aligned}\mathcal{L}_t = &-\log \frac{\sum_{t_p\in P_{t_j}} \exp\left(s(h_{t_j s}, h_{t_p m})/\tau\right)}{\sum_{l=1}^{N} \exp\left(s(h_{t_j s}, h_{t_l m})/\tau\right)} \\ &-\log \frac{\sum_{t_p\in P_{t_j}} \exp\left(s(h_{t_j m}, h_{t_p n})/\tau\right)}{\sum_{l=1}^{N} \exp\left(s(h_{t_j m}, h_{t_l n})/\tau\right)} \\ &-\log \frac{\sum_{t_p\in P_{t_j}} \exp\left(s(h_{t_j n}, h_{t_p s})/\tau\right)}{\sum_{l=1}^{N} \exp\left(s(h_{t_j n}, h_{t_l s})/\tau\right)}\end{aligned} \tag{15}$$

where P_{t_j} represents the positive sample set of target t_j, and N is the number of all targets.

2.4. Prediction module

In MFCLDTA, multi-scale feature fusion is a critical component of the framework. After optimizing the representations of features from each scale (sequence, molecular structure, and affinity graph) through the multi-scale contrastive learning module, we directly perform an element-wise summation of the feature vectors from the three scales. This weighted fusion operation is an intermediate fusion at the feature level (Stahlschmidt et al., 2022). However, by leveraging contrastive learning, we effectively minimize inter-modal discrepancies. The multi-scale contrastive learning module ensures that features from different scales are projected into a shared, aligned semantic space, making the element-wise fusion operation effective. This approach retains the depth of modality-specific information mining while efficiently capturing cross-scale synergistic effects.

Specifically, the final representation of a drug d can be expressed as:

$$h_d = h_{ds} + h_{dm} + h_{dn} \tag{16}$$

Similarly, the final representation of a target t can be expressed as:

$$h_t = h_{ts} + h_{tm} + h_{tn} \tag{17}$$

The resulting drug-target pairs are fed into the MLP to predict binding affinity scores, which are subsequently used to evaluate model performance:

$$\hat{y}_{d,t} = MLP(h_d || h_t) \tag{18}$$

The mean square error (MSE) loss function is used to calculate the supervised loss in the prediction process:

$$\mathcal{L}_m = \sum_{(d,t)\in\mathbb{C}} \frac{1}{|\mathbb{C}|}(\hat{y}_{d,t} - y_{d,t})^2 \tag{19}$$

where $\mathbb{C}$ denotes the total number of drug-target pairs used for training, and $y_{d,t}$ represents the experimentally measured affinity value. The final total loss can be expressed as:

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_d + \mathcal{L}_t \tag{20}$$

3. Experiments

3.1. Evaluation and training setup

This study adopts three established evaluation metrics from prior research to comprehensively assess model performance: mean squared error (MSE), concordance index (CI), and regression mean (r_m^2) (Jiang et al., 2020; Öztürk et al., 2018). MSE quantifies the average squared deviation between predicted and actual values, serving as a standard regression benchmark. CI measures the ranking consistency of predictions, where values range from 0 to 1. The r_m^2 metric evaluates both predictive power and correlation strength, with higher values denoting greater reliability of the model.

To ensure a fair comparison, all baseline methods in this study are evaluated using identical training and test sets. We randomly split the Davis and KIBA datasets into training and test sets in a 5:1 ratio. We retrain and reevaluate all baseline models to eliminate potential biases arising from differences in data splitting, preprocessing, or evaluation criteria, thereby guaranteeing a fair comparison. For hyperparameter optimization, we employ 5-fold cross-validation (5-CV) on the training set.

Our proposed model is implemented using Python 3.10.15 with PyTorch 2.0.0 and PyTorch Geometric 2.3.1. All experimental evaluations are performed on a computing system equipped with an Intel Core i7-14700KF processor (3.40 GHz), 32 GB of RAM, and an NVIDIA GeForce RTX 4070 Ti SUPER graphics processing unit (GPU). The key hyperparameter configurations for our model are detailed in Table 2.

3.2. Comparison with baselines

We categorize prediction models into three types based on their information source: sequence-based, graph-based, or multi-scale. Our benchmarking framework compares: (1) sequence-based methods (Deep-DTA, AttentionDTA); (2) graph-based methods (GraphDTA, DGraphDTA, MgraphDTA); and (3) multi-scale methods (HGRL-DTA, CSCo-DTA, PocketDTA, AttentionMGT-DTA, MultiKD-DTA, MLC-DTA). All baselines are evaluated using optimal hyperparameters from their original publications

- DeepDTA (Öztürk et al., 2018): extracts feature representations from drug SMILES and target sequences using CNNs for DTA prediction.

Table 2
Hyperparameter settings of MFCLDTA.

Hyper-parameters	Setting
Learning rate	0.0002
Epoch	3000
Batch size	512
Dropedge rate	0.2
Optimizer	Adam
GCN layers	3
BILSTM layers	3
Temperature τ	0.2
Biased item λ	0.5
K	5

- AttentionDTA (Zhao et al., 2019): introduces an attention mechanism to enable cross-modal feature interaction after CNN-based feature extraction.
- GraphDTA (Nguyen et al., 2021): represents the first molecular graph-based approach that employs RDKit for SMILES-to-graph conversion, with structural features subsequently encoded through GCN layers.
- DGraphDTA (Jiang et al., 2020): extends GraphDTA by integrating drug molecular graphs with protein structure graphs, leveraging Pconsc4 to generate protein molecular graphs and incorporate additional structural information.
- MgraphDTA (Yang et al., 2022): combines multi-layer graph neural networks with convolutional neural networks to extract structural features for high-accuracy DTA predictions.
- HGRL-DTA (Chu et al., 2022): constructs a hierarchical learning framework by representing drug-target binding affinity as an affinity graph, integrating features learned from the affinity graph with those from molecular graphs to enhance molecular representation.
- CSCo-DTA (Wang et al., 2024): develops a method to learn drug and protein features from molecular and network scales, leveraging contrastive learning to capture information from both local and global perspectives.
- PocketDTA (Zhao et al., 2024): leverages pretrained models to enhance generalizability and integrates target binding pocket information with three-dimensional drug structural features through a bilinear attention network, thereby improving interpretability.
- AttentionMGT-DTA (Wu et al., 2024): introduces a model that represents drugs and proteins using molecular graphs and binding pocket graphs, integrating and interacting information from different modalities through two attention mechanisms.
- MultiKD-DTA (Hu et al., 2025): combines Graph Neural Networks, multi-scale convolutional networks, and the pre-trained protein language model ESM-2 to enhance the prediction performance of drug-target binding affinity.
- MLC-DTA (Zheng et al., 2025): integrates Equivariant Graph Neural Networks and multi-level contrastive learning to extract features from both molecular and network perspectives, thereby improving the accuracy of drug-target affinity prediction.

Table 3 presents the performance comparison between MFCLDTA and other baseline models on the Davis dataset. The boldface results indicate the best performance achieved among all models, while the underlined results denote the second best. As demonstrated, MFCLDTA achieves the best performance across all evaluation metrics on the Davis dataset. Specifically, MFCLDTA yields a 15.1% reduction in MSE, a 1% increase in CI, and a 2.6% improvement in r_m^2 compares to the best-performing baseline.

Moreover, the analysis reveals the following trends: sequence-based models exhibit inferior performance compared to graph-based methods due to their inability to incorporate molecular structural information. In contrast, graph-based methods perform better due to their ability

Table 3
Performance comparison of MFCLDTA and baseline models on the Davis dataset.

Dataset	Model	Category	MSE	CI	r_m^2
Davis	DeepDTA	Sequence-based method	0.261	0.878	0.630
	AttentionDTA		0.223	0.893	0.657
	GraphDTA	Graph-based method	0.229	0.893	0.649
	DGraphDTA		0.216	0.891	0.686
	MgraphDTA		0.207	0.900	0.710
	HGRL-DTA	Multi-scale based method	0.166	<u>0.909</u>	0.751
	CSCo-DTA		0.166	0.904	0.776
	PocketDTA		<u>0.165</u>	0.903	0.743
	AttentionMGT-DTA		0.193	0.891	0.699
	MultiKD-DTA		0.201	0.903	0.764
	MLC-DTA		0.169	0.907	<u>0.785</u>
	MFCLDTA(ours)		**0.140**	**0.918**	**0.808**

Table 4
Performance comparison of MFCLDTA and baseline models on the KIBA dataset.

Dataset	Model	Category	MSE	CI	r_m^2
KIBA	DeepDTA	Sequence-based method	0.194	0.863	0.673
	AttentionDTA		0.214	0.852	0.633
	GraphDTA	Graph-based method	0.167	0.890	0.699
	DGraphDTA		0.141	0.895	0.767
	MgraphDTA		0.132	0.900	0.800
	HGRL-DTA	Multi-scale based method	0.129	<u>0.904</u>	0.789
	CSCo-DTA		<u>0.127</u>	0.901	0.804
	PocketDTA		0.136	0.895	0.782
	AttentionMGT-DTA		0.140	0.893	0.786
	MultiKD-DTA		0.141	0.899	0.793
	MLC-DTA		0.132	0.896	<u>0.805</u>
	MFCLDTA(ours)		**0.120**	**0.909**	**0.810**

to capture the structural characteristics of drugs and targets. However, graph-based methods often experience over-smoothing artifacts during the embedding process. Finally, multi-scale based methods outperform graph-based methods, as integrating features across multiple scales not only mitigates over-smoothing but also enables the model to capture more comprehensive representations of drug-target features.

To further validate the performance of MFCLDTA, we conduct additional comparative experiments on the KIBA dataset. As shown in Table 4, MFCLDTA consistently achieves the best performance across all evaluation metrics, demonstrating its high accuracy and excellent generalization ability.

To visually evaluate the predictive performance of our model, we present scatter plots in Fig. 2 that illustrate the relationship between predicted values and experimental measurements for MFCLDTA on both the Davis and KIBA datasets. The x-axis corresponds to the experimental measurements, while the y-axis corresponds to the model's predicted values. Data points closer to the red dashed line indicate higher predictive accuracy. Notably, most samples are symmetrically distributed around the central line, demonstrating strong agreement between the predicted and experimental affinity values. Furthermore, the predicted points for the KIBA dataset are more densely clustered along the central axis, suggesting that MFCLDTA achieves superior predictive accuracy on this dataset.

3.3. The performance under different cold-start scenarios

In previous studies, we validate the performance of our model on two benchmark datasets using 5-fold cross-validation. However, practical applications frequently involve cold-start scenarios where both the drug and the target are unknown. Hence, the generalization ability of the model is particularly important in DTA prediction tasks. Building upon prior work (Wang et al., 2022), we evaluate the generalization

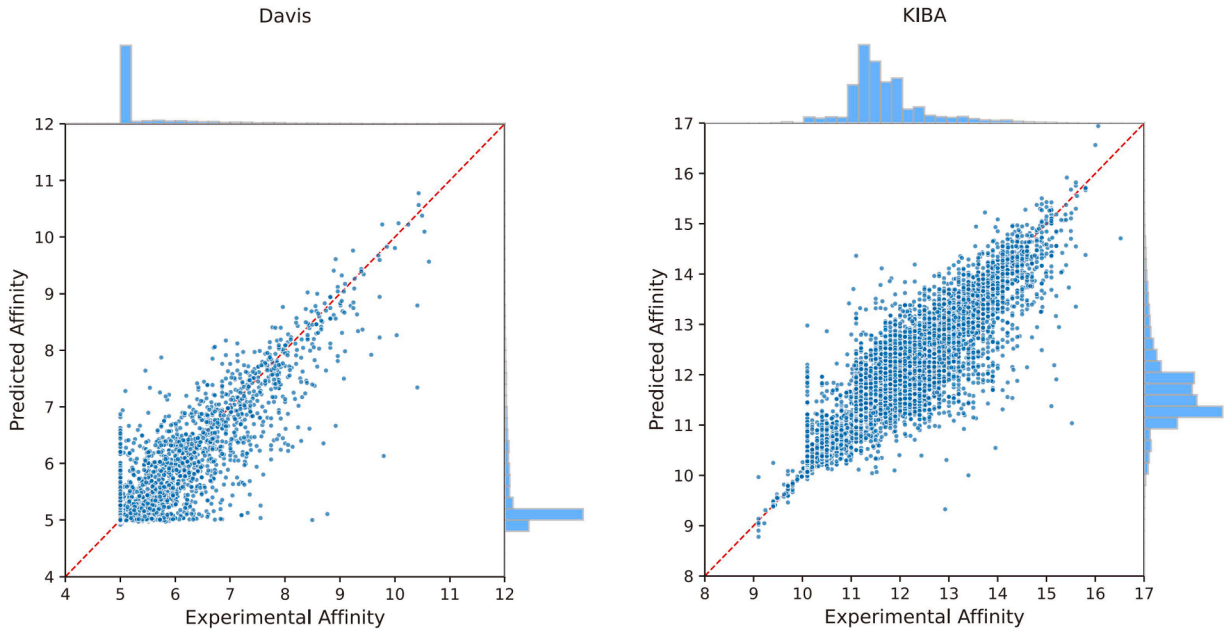

Fig. 2. Scatter plot of the predicted affinity values against the experimental measurements for MFCLDTA on Davis and KIBA datasets.

Table 5
Statistics of drug and target quantities under different cold-start scenarios.

Scenarios	Drug quantity		Target quantity	
	Training set	Test set	Training set	Test set
Drug Cold-start	57	11	442	442
Target Cold-start	68	68	301	60
All cold-start	57	11	301	60

performance of our model using three different cold-start scenarios. Table 5 details the specific counts of drugs and targets in the training and test sets under these different cold-start scenarios:

- Drug cold-start: each drug in the test set does not appear in the training set.
- Target cold-start: each target in the test set does not appear in the training set.
- All cold-start: neither drugs nor targets in the test set appear in the training set.

In more realistic and complex cold-start scenarios, we comprehensively evaluate MFCLDTA alongside nine baseline models on Davis dataset and the results are presented in Table 6. As expected, the performance of all models shows a significant decrease, demonstrating the substantial challenge that models face in unknown circumstances and further emphasizing the importance of model generalization. Compared to the other baseline models, the MFCLDTA model exhibits superior performance across all three cold-start scenarios. Specifically, in the Drug cold-start scenario, the MFCLDTA model achieves the best performance in terms of MSE and r_m^2, with a 1.7% decrease in MSE and a 3% increase in r_m^2. In the Target cold-start scenario, the MFCLDTA model improves by 0.6% and 1.9% in terms of CI and r_m^2, respectively. In the All cold-start scenario, the MFCLDTA model demonstrates the best performance across all three metrics, with a 2.3% decrease in MSE, and 2.6% and 5.4% increases in CI and r_m^2, respectively. These findings indicate that MFCLDTA, by leveraging known information to integrate features across multiple scales, demonstrates robust generalization and resilience under complex experimental conditions.

3.4. Ablation study

In this section, we perform ablation studies to investigate the contributions of individual components within our model and validate the underlying design assumptions. Under consistent experimental settings, we evaluate MFCLDTA and its variants on the Davis dataset.

MFCLDTA (w/o M&N): only learns sequence-based features of drugs and targets, without using molecular structure and affinity graph features.

MFCLDTA (w/o S&N): only learns molecular structure features of drugs and targets, without using sequence-based and affinity graph features.

MFCLDTA (w/o S&M): only learns affinity graph features of drugs and targets, without using sequence-based and molecular structure features.

MFCLDTA (w/o N): learns sequence-based and molecular structure features of drugs and targets, but does not use affinity graph features.

MFCLDTA (w/o M): learns sequence-based and affinity graph features of drugs and targets, but does not use molecular structure features.

MFCLDTA (w/o S): learns molecular structure and affinity graph features of drugs and targets, but does not use sequence-based features.

MFCLDTA (w/o CL): learns features from three scales: sequence-based, molecular structure, and affinity graph, but does not employ contrastive learning, instead relying solely on simple feature concatenation.

Fig. 3 depicts the performance comparison of MFCLDTA and its seven variants on the Davis dataset. It is evident that MFCLDTA outperforms all other variants, thereby validating the correctness of the research methodology. Notably, variants such as MFCLDTA (w/o M&N), MFCLDTA (w/o S&N), and MFCLDTA (w/o S&M), which utilize only single-scale feature information, demonstrate inferior performance compared to variants like MFCLDTA (w/o N), MFCLDTA (w/o M), and MFCLDTA (w/o S), which incorporate feature fusion from multiple scales. This result highlights the advantage of multi-scale information integration, as it enables the model to capture a more comprehensive set of drug-target features and compensates for the limitations of single-scale representations. Consequently, it demonstrates that the predictive performance of the model improves as information from more scales is integrated, with the combination of all three scales outperforming those that utilize only two. Additionally, compared to the variant MFCLDTA (w/o CL) that excludes the contrastive learning module, MFCLDTA achieves

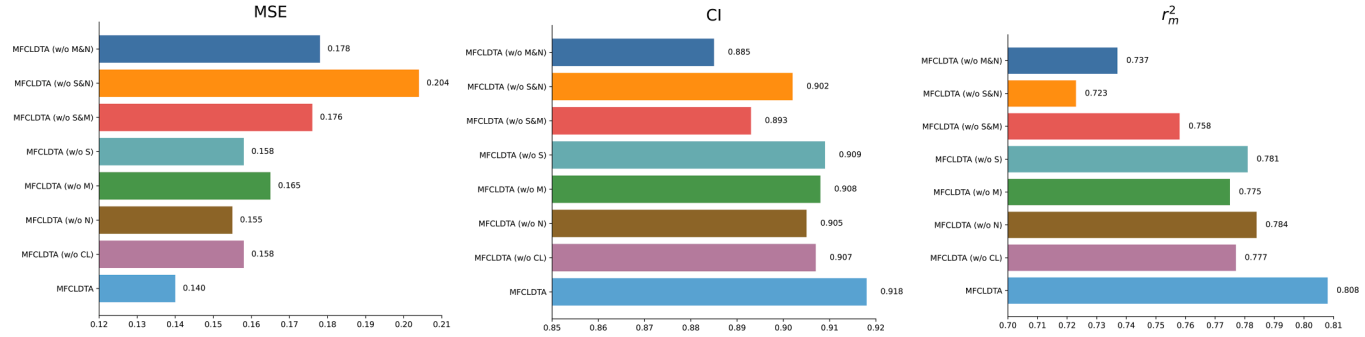

Fig. 3. The ablation study results of MFCLDTA with MSE, CI, and r_m^2 metrics on the Davis dataset.

Table 6
Performance of MFCLDTA and the baselines on the Davis dataset under different cold start scenarios.

Scenarios	Model	MSE	CI	r_m^2
Drug cold-start	DeepDTA	0.871	0.593	0.047
	AttentionDTA	0.816	0.670	0.101
	GraphDTA	0.784	0.673	0.129
	DGraphDTA	0.806	0.652	0.119
	MgraphDTA	0.834	0.621	0.104
	HGRL-DTA	0.757	0.684	0.163
	CSCo-DTA	0.765	0.678	0.151
	PocketDTA	0.753	**0.704**	0.158
	AttentionMGT-DTA	0.749	0.676	0.162
	MultiKD-DTA	0.762	0.668	0.160
	MLC-DTA	0.756	0.673	0.155
	MFCLDTA(ours)	**0.736**	0.693	**0.168**
Target cold-start	DeepDTA	0.512	0.735	0.278
	AttentionDTA	0.431	0.787	0.304
	GraphDTA	0.763	0.689	0.158
	DGraphDTA	0.426	0.795	0.310
	MgraphDTA	**0.377**	0.813	0.415
	HGRL-DTA	0.393	0.813	0.394
	CSCo-DTA	0.395	0.809	0.364
	PocketDTA	0.389	0.771	0.424
	AttentionMGT-DTA	0.406	0.824	0.327
	MultiKD-DTA	0.409	0.807	0.378
	MLC-DTA	0.392	0.813	0.427
	MFCLDTA(ours)	0.383	**0.829**	**0.435**
All cold-start	DeepDTA	0.697	0.508	0.012
	AttentionDTA	0.679	0.554	0.009
	GraphDTA	0.796	0.569	0.016
	DGraphDTA	0.658	0.572	0.026
	MgraphDTA	0.744	0.528	0.002
	HGRL-DTA	0.632	0.611	0.038
	CSCo-DTA	0.629	0.616	0.044
	PocketDTA	0.598	0.637	0.053
	AttentionMGT-DTA	0.601	0.631	0.055
	MultiKD-DTA	0.617	0.598	0.035
	MLC-DTA	0.605	0.622	0.038
	MFCLDTA(ours)	**0.585**	**0.654**	**0.058**

superior results across various predictive metrics, further confirming the effectiveness of the contrastive learning component.

Furthermore, to comprehensively evaluate the role of contrastive loss functions within the MFCLDTA framework and validate the robustness of our method, we introduce two additional classic contrastive learning loss functions: Max-Margin Loss (Shah et al., 2022) and Triplet Loss (Ge, 2018), and conduct detailed comparative experiments on the Davis dataset.

In our MFCLDTA framework, we replace the original InfoNCE loss with these two loss functions respectively, while keeping all other model components and hyperparameters, such as learning rate, batch size, and temperature parameter, entirely unchanged to ensure a fair comparison. The evaluation results on the Davis dataset are presented in Table 7.

Table 7
Results of running different loss functions on the Davis dataset.

Loss function	MSE	CI	r_m^2
Max-Margin Loss	0.152	0.905	0.781
Triplet Loss	0.147	0.911	0.792
InfoNCE (ours)	0.140	0.918	0.808

As shown in Table 7, the InfoNCE loss adopted in our original model achieves the best performance across all three evaluation metrics. This indicates that InfoNCE more effectively utilizes all negative samples within a batch, guiding the model to learn more discriminative multi-scale feature representations. Both Max-Margin and Triplet losses also yield competitive results but exhibit a slight performance gap compared to InfoNCE. We speculate that this may be because these two loss functions focus more on learning the relative ordering of sample pairs, whereas InfoNCE more precisely models the distribution of the entire feature space, thereby enabling the learning of more robust features.

3.5. Parameter optimization and analysis

This section examines two critical hyperparameters governing model performance: (1) the number of positive samples K per anchor in the multi-scale contrastive learning framework, and (2) the temperature parameter τ controlling contrastive loss computation.

In the contrastive learning framework, the selection of positive samples directly influences the model's ability to capture intrinsic data features, serving as a crucial hyperparameter affecting the effectiveness of contrastive learning. While increasing the number of positive samples generally enhances feature robustness and reduces sampling bias, it may simultaneously diminish discriminative contrast. The temperature parameter τ plays a pivotal role in controlling the sensitivity of the loss function, directly affecting the model's distinction between positive and negative samples.

We also conduct experiments on the Davis dataset, systematically varying K from 1 to 10 and testing the temperature parameter τ in increments of 0.2 over the range 0.2 to 1.0, to evaluate their impact on model performance. As illustrated in Fig. 4, the axes represent the number of positive samples K and the temperature parameter τ. It can be observed that our model achieves optimal performance when K is set to 5 and τ is set to 0.2.

3.6. Interpretability and visualisation analysis

To elucidate the mechanisms of interaction between drug molecules and target proteins, we conduct visualization experiments and related analyses. These studies help analyze and predict potential protein binding sites while enhancing the interpretability of our model. Specifically, we select two well-characterized target proteins from the Davis dataset

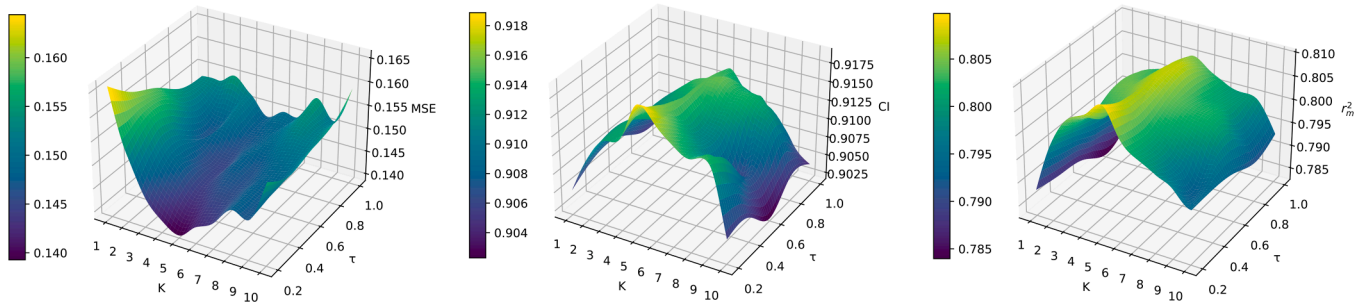

Fig. 4. Parameter sensitivity analysis of parameter K and τ with the Davis dataset.

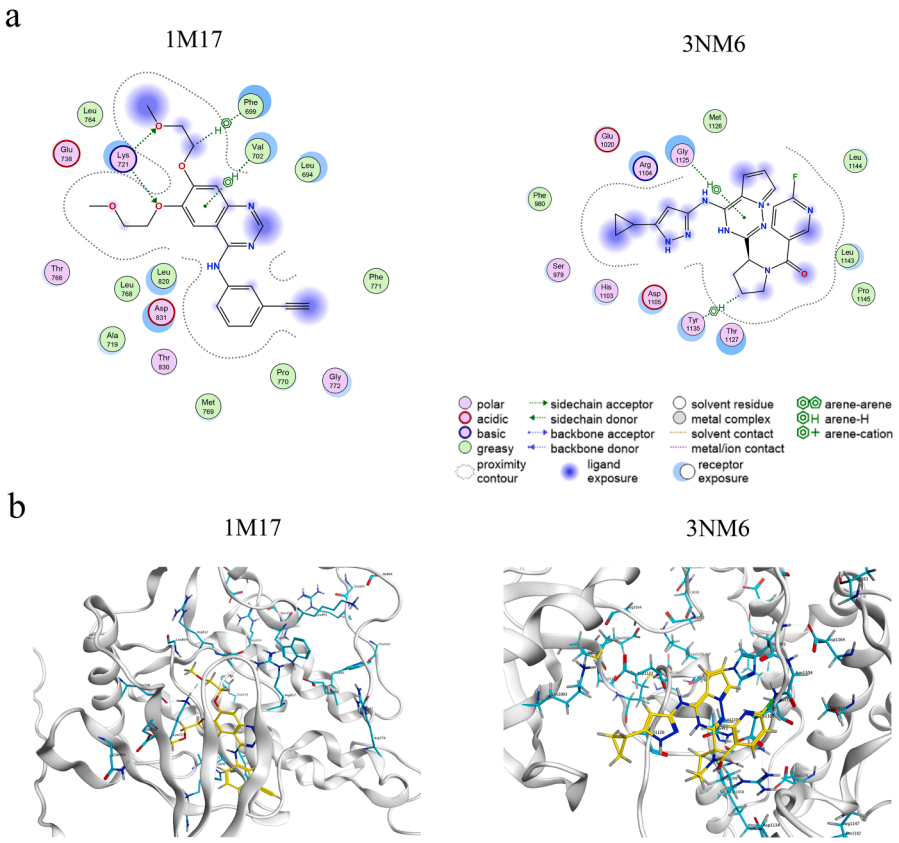

Fig. 5. Visualization of Ligand-Protein Residue Interactions. (a) The ligand-protein interaction diagram displays the two-dimensional structure of the ligand at the center, surrounded by various protein residues. Different colors indicate residues with distinct properties, and blue regions highlight ligand areas that closely interact with protein residues. (b) The 3D binding visualization depicts the ligand in yellow and protein residues in cyan. Dashed lines illustrate the interactions between the ligand and protein residues.

(PDB ID: 1M17 and 3NW6) and visualize ligand-protein interactions using the Molecular Operating Environment (MOE) software (Vilar et al., 2008). As shown in Fig. 5, key atoms and structural features are specially highlighted in the visualization results.

For the PDB structure 1M17 (receptor tyrosine kinase complexed with the 4-anilinoquinazoline inhibitor erlotinib), our model accurately identifies and predicts the essential role of the amide carbonyl functional group in ligand-protein binding. This functional group, serving as a carboxamide linker between the phenyl and pyridine rings, provides a carbonyl oxygen atom that acts as a hydrogen bond acceptor to the backbone amine of Phe699. Moreover, the terminal amide carbonyl forms a hydrogen bond with the side chain amine donor of Lys721, further underscoring the interpretability of our model. Additionally, the nitrogen atom on the pyridine ring serves as a hydrogen bond acceptor, interacting with the hydrogen atom donor of the backbone amine of Val702. Notably, important ligand regions that are closely packed with protein residues are highlighted in blue. These regions do not participate in observable chemical reactions such as hydrogen bonding but instead make significant hydrophobic contacts with the protein's binding pocket. Through the hydrophobic effect, these regions facilitate the displacement of water molecules, thereby promoting a more stable ligand binding conformation. Furthermore, Glu738 and Asp831, highlighted in red, do not directly form hydrogen bonds or salt bridges with the ligand but may play key roles in ligand recognition and stabilizing the binding site environment.

In the PDB structure 3NW6 (a crystal structure of the insulin-like growth factor receptor complexed with a carbon-linked proline isostere inhibitor), the interpretability of MFCLDTA once again elucidates the specific sites and modes of ligand-protein interactions in detail. The main skeleton of the ligand targets the imino group adjacent to the hydroxyl group, which acts as a hydrogen bond acceptor and forms a hydrogen bond with the phenolic hydroxyl group of the Tyr1135 side

Table 8
Performance on the Davis dataset with different amounts of affinity used.

Affinity	MSE	CI	r_m^2
MFCLDTA (w/o N)	0.155	0.905	0.784
20% affinity	0.154	0.905	0.787
50% affinity	0.149	0.911	0.796
80% affinity	0.143	0.916	0.804
MFCLDTA	0.140	0.918	0.808

chain. Additionally, the π-system of the ligand's aromatic six-membered ring engages in a π-H interaction with the Cα-hydrogen of the Gly1125 backbone, further helping to define the binding position of the ligand.

Based on the above visualization analyses, the proposed model exhibits strong interpretability, successfully elucidating receptor-protein interaction mechanisms and predicting potential binding sites. This capability may facilitate the identification of cryptic interactions and expedite drug discovery.

3.7. Performance at different amounts of affinity

To more intuitively demonstrate the effect of varying the number of target pairs on model performance, we conduct experiments on the Davis dataset using only 20%, 50%, and 80% of the known affinity values, respectively. Here, MFCLDTA (w/o N) refers to the variant from our ablation study that completely omits the affinity graph. The experimental results are shown in Table 8.

The results indicate that as the number of drug-target pairs used to construct the affinity graph increases, the model's performance improves progressively. This demonstrates that the affinity graph provides complementary information and enhances the model's performance. In other words, a greater number of known interactions helps in learning richer node representations, thereby improving the model's generalization capability for unseen drug-target pairs.

4. Conclusion

Based on a systematic review and analysis of existing research, this study introduces MFCLDTA. This novel deep learning model integrates multi-scale feature information and a contrastive learning mechanism to predict drug-target binding affinity more accurately. Our model simultaneously employs BiLSTM to capture contextual information from drug SMILES and target sequences at the sequence scale, utilizes GCN to extract topological relationships in drug molecular structure graphs and target protein interaction networks at the molecular structure scale, and further applies GCN to capture interactions between drugs and targets at the affinity graph scale, thereby achieving a comprehensive representation of drugs and targets. Through a multi-scale contrastive learning module, MFCLDTA maximizes the mutual information between features at different scales, ensuring robust feature alignment and effective fusion. Systematic evaluations on the Davis and KIBA benchmark datasets demonstrate that MFCLDTA, incorporating three different scales of information, outperforms current state-of-the-art methods across all key evaluation metrics. Furthermore, ablation experiments validate the critical role of both the multi-scale feature fusion strategy and the contrastive learning mechanism in enhancing prediction performance. Additionally, cold-start experiment results indicate that MFCLDTA exhibits excellent generalization ability and predictive robustness for novel drugs and targets, highlighting its significant potential in the field of drug discovery. Notably, visualization analysis indicates that MFCLDTA not only makes accurate predictions but also identifies atoms and residues that contribute critically to binding affinity, providing valuable insights into understanding drug-target interaction mechanisms.

In summary, MFCLDTA, through its innovative multi-scale feature integration and contrastive learning architecture, provides a powerful and versatile solution for accurate DTA prediction. We believe that this framework holds significant application potential in advancing computer-aided drug discovery. Future work will explore incorporating information from more scales (such as 3D structures) into this framework and further optimizing contrastive learning strategies.

CRediT authorship contribution statement

Zhen Tian: Conceptualization, Methodology, Software, Writing – original draft; **Saisai Zhu:** Investigation, Supervision; **Zhixia Teng:** Software, Investigation, Validation; **Xiaoqiang Yan:** Writing – review & editing, Conceptualization; **Tao Wang:** Conceptualization, Writing – review & editing, Supervision.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 62371423, 62271132), the Municipal Government of Quzhou (No. 2024D033) and the Natural Science Foundation of Henan (Nos. 252300421226 and 252300421504), Natural Science Foundation of Heilongjiang Province (No. LH2024F001).

References

Abbasi, K., Razzaghi, P., Poso, A., Ghanbari-Ara, S., & Masoudi-Nejad, A. (2021). Deep learning in drug target interaction prediction: current and future perspectives. *Current Medicinal Chemistry*, *28*(11), 2100–2113.

Ahmad, W., Simon, E., Chithrananda, S., Grand, G., & Ramsundar, B. (2022). Chemberta-2: Towards chemical foundation models. arXiv:2209.01712.

Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, *16*, 345–379.

Bolton, E. E., Wang, Y., Thiessen, P. A., & Bryant, S. H. (2008). Pubchem: integrated platform of small molecules and biological activities. In *Annual reports in computational chemistry* (pp. 217–241). Elsevier (*vol. 4*).

Chu, Z., Huang, F., Fu, H., Quan, Y., Zhou, X., Liu, S., & Zhang, W. (2022). Hierarchical graph representation learning for the prediction of drug-target binding affinity. *Information Sciences*, *613*, 507–523.

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., Hocker, M., Treiber, D. K., & Zarrinkar, P. P. (2011). Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, *29*(11), 1046–1051.

van den, O. A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv:1807.03748.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M. et al. (2021). Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine intelligence*, *44*(10), 7112–7127.

Ezzat, A., Wu, M., Li, X.-L., & Kwoh, C.-K. (2019). Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings in Bioinformatics*, *20*(4), 1337–1357.

Ge, W. (2018). Deep metric learning with hierarchical triplet loss. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 269–285).

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 315–323). JMLR Workshop and Conference Proceedings.

Graves, A., & Graves, A. (2012). Long short-term memory. *Supervised Sequence Labelling with Recurrent Neural Networks*, (pp. 37–45).

He, T., Heidemeyer, M., Ban, F., Cherkasov, A., & Ester, M. (2017). Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal of Cheminformatics*, *9*, 1–14.

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular dynamics simulation for all. *Neuron*, *99*(6), 1129–1143.

Hu, R., Ge, R., Deng, G., Fan, J., Tang, B., & Wang, C. (2025). MultiKD-DTA: Enhancing drug-target affinity prediction through multiscale feature extraction. *Interdisciplinary Sciences: Computational Life Sciences*, (pp. 1–11).

Hua, Y., Song, X., Feng, Z., & Wu, X. (2023). Mfr-dta: a multi-functional and robust model for predicting drug–target binding affinity and region. *Bioinformatics*, *39*(2), btad056.

Jiang, M., Li, Z., Zhang, S., Wang, S., Wang, X., Yuan, Q., & Wei, Z. (2020). Drug–target affinity prediction using graph neural network and contact maps. *RSC Advances*, *10*(35), 20701–20712.

Kukol, A. et al. (2008). Molecular modeling of proteins (vol. 443). Springer.

Li, H., Leung, K.-S., Wong, M.-H., & Ballester, P. J. (2015). Low-quality structural and interaction data improves binding affinity prediction via random forest. *Molecules*, *20*(6), 10947–10962.

Li, W., Wang, C.-h., Cheng, G., & Song, Q. (2023). International conference on machine learning. *Transactions on Machine Learning Research*, .

Liu, X., Song, C., Huang, F., Fu, H., Xiao, W., & Zhang, W. (2022). GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction. *Briefings in Bioinformatics*, *23*(1), bbab457.

Lovrić, M., Molero, J. M., & Kern, R. (2019). Pyspark and RDKit: moving towards big data in cheminformatics. *Molecular Informatics*, *38*(6), 1800082.

Michel, M., Menéndez Hurtado, D., & Elofsson, A. (2019). Pconsc4: fast, accurate and hassle-free contact predictions. *Bioinformatics*, *35*(15), 2677–2679.

Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., & Venkatesh, S. (2021). GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, *37*(8), 1140–1147.

Nogales, C., Mamdouh, Z. M., List, M., Kiel, C., Casas, A. I., & Schmidt, H. H. (2022). Network pharmacology: curing causal mechanisms instead of treating symptoms. *Trends in Pharmacological Sciences*, *43*(2), 136–150.

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

Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szwajda, A., Tang, J., & Aittokallio, T. (2015). Toward more realistic drug–target interaction predictions. *Briefings in Bioinformatics*, *16*(2), 325–337.

Prasad, V., De Jesús, K., Mailankody, S. (2017). The high price of anticancer drugs: origins, implications, barriers, solutions. In *Nature reviews Clinical oncology*, *14*(6),(pp. 381–390). Nature Publishing Group UK London

Qi, H., Yu, T., Yu, W., & Liu, C. (2024). Drug–target affinity prediction with extended graph learning-convolutional networks. *BMC Bioinformatics*, *25*(1), 75.

Qian, Y., Ding, Y., Zou, Q., & Guo, F. (2022). Identification of drug-side effect association via restricted boltzmann machines with penalized term. *Briefings in Bioinformatics*, *23*(6), bbac458.

Shah, A., Sra, S., Chellappa, R., & Cherian, A. (2022). Max-margin contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8220–8230). (*vol. 36*).

Shoichet, B. K., Leach, A. R., & Kuntz, I. D. (1999). Ligand solvation in molecular docking. *Proteins: Structure, Function, and Bioinformatics*, *34*(1), 4–16.

Smith, T. F., Waterman, M. S. et al. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, *147*(1), 195–197.

Stahlschmidt, S. R., Ulfenborg, B., & Synnergren, J. (2022). Multimodal deep learning for biomedical data fusion: a review. *Briefings in Boinformatics*, *23*(2), bbab569.

Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., & Siedlecki, P. (2018). Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, *34*(21), 3666–3674.

Takebe, T., Imai, R., & Ono, S. (2018). The current status of drug discovery and development as originated in united states academia: the influence of industrial and academic collaboration on drug discovery and development. *Clinical and Translational Science*, *11*(6), 597–606.

Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., & Aittokallio, T. (2014). 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.

Thafar, M., Raies, A. B., Albaradei, S., Essack, M., & Bajic, V. B. (2019). Comparison study of computational prediction tools for drug-target binding affinities. *Frontiers in Chemistry*, *7*, 782.

Trott, O., & Olson, A. J. (2010). Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, *31*(2), 455–461.

Vilar, S., Cozza, G., & Moro, S. (2008). Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Current Topics in Medicinal Chemistry*, *8*(18), 1555–1572.

Wang, J., Wen, N., Wang, C., Zhao, L., & Cheng, L. (2022). Electra-dta: a new compound-protein binding affinity prediction model based on the contextualized sequence encoding. *Journal of Cheminformatics*, *14*(1), 14.

Wang, J., Xiao, Y., Shang, X., & Peng, J. (2024). Predicting drug–target binding affinity with cross-scale graph contrastive learning. *Briefings in Bioinformatics*, *25*(1), bbad516.

Wang, K., Zhou, R., Li, Y., & Li, M. (2021a). DeepDTAF: a deep learning method to predict protein–ligand binding affinity. *Briefings in Bioinformatics*, *22*(5), bbab072.

Wang, S., Song, X., Zhang, Y., Zhang, K., Liu, Y., Ren, C., & Pang, S. (2023). Msgnn-dta: Multi-scale topological feature fusion based on graph neural networks for drug–target binding affinity prediction. *International Journal of Molecular Sciences*, *24*(9), 8326.

Wang, Z., Zheng, L., Liu, Y., Qu, Y., Li, Y.-Q., Zhao, M., Mu, Y., & Li, W. (2021b). Onionnet-2: a convolutional neural network model for predicting protein-ligand binding affinity based on residue-atom contacting shells. *Frontiers in Chemistry*, *9*, 753002.

Wei, J., Zhuo, L., Zhou, Z., Lian, X., Fu, X., & Yao, X. (2023). Gcfmcl: predicting mirna-drug sensitivity using graph collaborative filtering and multi-view contrastive learning. *Briefings in Bioinformatics*, *24*(4), bbad247.

Wouters, O. J., McKee, M., & Luyten, J. (2020). Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *JAMA*, *323*(9), 844–853.

Wu, H., Liu, J., Jiang, T., Zou, Q., Qi, S., Cui, Z., Tiwari, P., & Ding, Y. (2024). AttentionMGT-DTA: A multi-modal drug-target affinity prediction using graph transformer and attention mechanism. *Neural Networks*, *169*, 623–636.

Yang, X., Yang, G., & Chu, J. (2024). GraphCL-DTA: a graph contrastive learning with molecular semantics for drug-target binding affinity prediction. *IEEE Journal of Biomedical and Health Informatics*, *28*(8), 4544–4552.

Yang, Z., Zhong, W., Zhao, L., & Chen, C. Y.-C. (2022). Mgraphdta: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical Science*, *13*(3), 816–833.

Yu, H., Xu, W.-X., Tan, T., Liu, Z., & Shi, J.-Y. (2024). Prediction of drug–target binding affinity based on multi-scale feature fusion. *Computers in Biology and Medicine*, *178*, 108699.

Yu, Z., Yu, J., Fan, J., & Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 1821–1830).

Yuan, W., Chen, G., & Chen, C. Y.-C. (2022). FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. *Briefings in Bioinformatics*, *23*(1), bbab506.

Zeng, Y., Chen, X., Peng, D., Zhang, L., & Huang, H. (2022). Multi-scaled self-attention for drug–target interaction prediction based on multi-granularity representation. *BMC Bioinformatics*, *23*(1), 314.

Zhang, L., Ouyang, C., Liu, Y., Liao, Y., & Gao, Z. (2023a). Multimodal contrastive representation learning for drug-target binding affinity prediction. *Methods*, *220*, 126–133.

Zhang, X., Li, Y., Wang, J., Xu, G., & Gu, Y. (2023b). A multi-perspective model for protein–ligand-binding affinity prediction. *Interdisciplinary Sciences: Computational Life Sciences*, *15*(4), 696–709.

Zhang, Y., Li, S., Meng, K., & Sun, S. (2024). Machine learning for sequence and structure-based protein–ligand interaction prediction. *Journal of Chemical Information and Modeling*, *64*(5), 1456–1472.

Zhao, L., Wang, H., & Shi, S. (2024). PocketDTA: an advanced multimodal architecture for enhanced prediction of drug- target affinity from 3d structural data of target binding pockets. *Bioinformatics*, *40*(10), btae594.

Zhao, Q., Xiao, F., Yang, M., Li, Y., & Wang, J. (2019). AttentionDTA: prediction of drug–target binding affinity using attention model. In *2019 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 64–69). IEEE.

Zheng, M., Sun, G., & Fan, Y. (2025). Mlc-dta: Drug-target affinity prediction based on multi-level contrastive learning and equivariant graph neural networks. *Neurocomputing*, *637*, 130052.

Zhu, F., Zhang, X., Allen, J. E., Jones, D., & Lightstone, F. C. (2020). Binding affinity prediction by pairwise function based on neural network. *Journal of Chemical Information and Modeling*, *60*(6), 2766–2772.