BMC Bioinformatics

RESEARCH

Open Access

DCI-SiteDTA: drug-target affinity prediction based on binding sites detection and site-aware dual cross-interaction block

Jinyang Zhang[1], Xingyang Li[1], Bo Wei[1] and Yuni Zeng[1*]

*Correspondence:
Yuni Zeng
yunizeng@zstu.edu.cn
[1]School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China

Abstract

Background Predicting the binding affinity between drugs and proteins is crucial for accelerating drug discovery. However, traditional research methods typically treat binding site detection and affinity prediction as two separate tasks, lacking solutions that integrate both into a unified deep learning framework. Furthermore, existing approaches exhibit limitations in two aspects: (1) for binding site detection, fine-grained features within drug target binding pockets need to be encoded; and (2) for drug-target affinity (DTA) prediction, the fusion of drug and targets needs to consider a multidimensional fusion guided by binding sites, rather than via concatenation or simple attention mechanisms.

Results To address the aforementioned challenges, we propose a drug-target affinity prediction based on binding sites detection and site-aware dual cross-interaction block(DCI-SiteDTA). Specifically, a multi-scale feature fusion is employed to extract both local and global contextual information from protein residues to get the binding site vector. Then, with a binding site-guided strategy, we introduce a dual cross-interaction fusion block. Designed to model drug-target interactions, this module constructs multilevel representations based on drug, target, and binding site information to capture their fundamental biological patterns of interaction and integration. Finally, our proposed DCI-SiteDTA model is evaluated on Davis and KIBA benchmark datasets. The experimental results reveal that our proposed model yields better accuracy on both binding site detection and DTA prediction.

Conclusions Our proposed model with the binding sites-guided strategy and dual cross-interaction block improves the binding site detection and affinity prediction of drug-target pairs.

Keywords Drug-target affinity, Deep learning, Binding site detection, Multi-scale feature fusion

Introduction

The determination of drug-target binding affinity is a crucial step in modern drug discovery [1, 2]. However, traditional drug development pipelines are notoriously time-consuming and costly; developing a new drug typically requires over a decade and

billions of dollars in investment, accompanied by a high risk of clinical failure [3]. Consequently, there is an urgent need within the pharmaceutical field to leverage computational methods to accelerate virtual screening [4] and reduce Research and Development costs. Against this background, binding pocket detection and drug-target affinity (DTA) prediction have emerged as two critical downstream tasks, offering valuable insights for drug discovery.

Binding pocket detection plays an important role in the early stages of drug discovery. Traditional template-based and energy-based methods rely on high-quality templates or complex energy simulations, yet their effectiveness is often limited by data availability and a trade-off between accuracy and efficiency. With the advancement of computational power and the maturity of deep learning frameworks, machine learning algorithms based on geometric features have made marked progress in recent years [5, 6]. For instance, Fpocket [7] employs a pure geometry strategy, rapidly characterizing the morphological features of binding pockets by aggregating alpha-spheres on the protein surface. P2Rank [8] further introduces a supervised learning framework; by extracting local geometric features of Connolly points and applying a Random Forest classifier, it efficiently predicts druggable regions on the protein surface. However, it is worth noting that these methods are generally used as isolated preprocessing steps in the drug-development pipeline, with little synergistic interaction downstream affinity-prediction tasks. As a task of equal importance to binding pocket detection, drug-target affinity prediction initially relied on traditional machine learning models, such as Random Forest (RF) [9], Support Vector Machine (SVM)[10], SimBoost[11] and Kronecker Regularized Least Squares (KronRLS)[12]. Although these traditional methods achieved certain successes in early drug-target interaction studies, they face marked limitations. First, their performance heavily depends on hand-crafted features or predefined molecular fingerprints, a process that requires substantial domain expertise [13]. Second, their shallow architectures struggle to capture deep underlying patterns and long-range dependencies in large-scale biological data [14]. Therefore, deep learning methods have been widely used in DTA prediction according to their strong feature extraction ability. Approaches for predicting drug-target affinity can generally be categorized into sequence-based CNN models(e.g., DeepDTA [15], DeepCDA[16] and WideDTA[17]), molecular graph-based GNN models (e.g., GraphDTA[18], DGraphDTA[19]), and the recently emerging Transformer models, such as TransformerCPI[20], which leverage the powerful attention mechanism of Transformers to directly model global interaction relationships between protein sequences and drug molecule sequences. Specifically, TransformerCPI dynamically focuses on key substructures within drug and target sequences by introducing a sequence interaction module, thereby modeling intermolecular interactions with greater precision. TAG-DTA[21]) further explores the integration of molecular graph topology with protein sequence semantic information, enabling synergistic feature encoding for proteins and drugs to enhance prediction accuracy.

Despite these efforts, the existing methods have several areas for improvement:

- Although the importance of binding site detection for DTA prediction is clear, many studies still address these two tasks separately.
- The binding sites detection requires finer-grained features.

- The binding site-guided drug-target fusion should involve an emphasis, like the patterns in the binding area, rather than concatenation and simple attention-based interaction.

To overcome these limitations, we propose Drug-Target Affinity Prediction Based on Binding Sites Detection and Site-Aware Dual Cross-Interaction Block(DCI-SiteDTA), an end-to-end deep learning framework that unifies binding site detection and affinity prediction. Firstly, the model takes protein sequences and drug SMILES strings as input and employs parallel Transformer encoders to generate initial embeddings [22]. Secondly, a multi-scale feature fusion block is introduced to integrate the finer-grained features with the global context for binding site detection. Next, based on the binding site information, a binding site-guided strategy is proposed based on accurate binding site information and other top-K (with the highest probability of binding site parts) context. Finally, a dual cross-interaction fusion block is introduced to predict the drug-target affinity score. We evaluate our framework on several benchmark datasets and compare it with other recent methods. The results demonstrate that our proposed DCI-SiteDTA has superior predictive performance on both binding-site detection and affinity-prediction tasks. Furthermore, our DCI-SiteDTA differs technically from some recent models based on binding site prediction, such as TAG-DTA. First, unlike TAG-DTA, which directly integrates binding site prediction into the standard Transformer attention layer, our model employs a Top-K context strategy. Second, we introduce a Dual Cross-Interaction (DCI) module, distinguishing it from TAG-DTA, which relies on unidirectional cross-attention or simple concatenation followed by self-attention. Finally, while TAG-DTA uses average pooling or [CLS] tag reading, we devise a unique gated weighted pooling strategy.

The contributions of this work are threefold:

- We proposed an end-to-end deep model to predict the binding sites and affinity score of drug-protein pairs, named Drug-Target Affinity Prediction Based on Binding Sites Detection and Site-Aware Dual Cross-Interaction Block(DCI-SiteDTA).
- A multi-scale feature fusion is designed for not only finer-grained features but also the global context of drug-protein pairs for binding site detection.
- A binding-site aware dual cross-interaction block is proposed to fuse deep representations of drugs and proteins for DTA prediction.

Methods

We propose a novel end-to-end deep learning framework that unifies binding site detection and affinity prediction.

The model first takes as input the one-dimensional SMILES string of the drug and the sequence string of the target. These sequences are then tokenized via Token Embedding to capture physicochemical semantic representations. Subsequently, a Transformer encoder is employed to model long-range dependencies among the tokens [23]. Notably, for protein encoding, we introduce a multi-scale feature fusion with an adjustable sliding window (Fig. 1b) to enhance both local and global contextual information among protein residue segments. Following this, a one-dimensional binding site classifier–comprising a Transformer encoder and a Position-Aware Multi-Layer Perceptron (PWMLP)–learns to identify binding and non-binding residues from the outputs of the contextual conditioning and positional encoding modules.

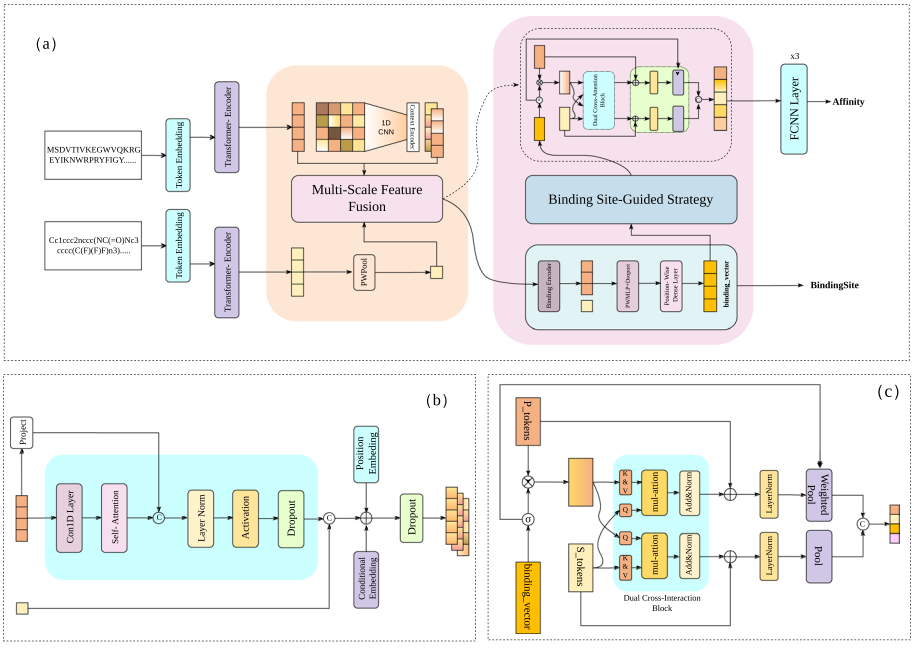

Fig. 1 **a** Overall architecture of Drug-Target Affinity Prediction Based on Binding Sites Detection and Site-Aware Dual Cross-Interaction Block(DCI-SiteDTA). The model takes protein sequences and drug SMILES strings as input. The parallel Transformer-Encoders and **b** the multi-scale feature fusion are used to extract and fuse features of drugs and proteins. Then, the binding site-guided strategy is designed to select the fused context of drug-protein pairs. Finally, **c** the dual cross-interaction block is proposed to predict affinity scores

Once the one-dimensional binding site vector is obtained, a binding sites-guided strategy with the other Top-K contexts is applied to dynamically filter high-confidence binding residues while preserving crucial contextual information, thereby laying the foundation for subsequent affinity prediction. The masked representations are then fed into a dual cross-interaction block (Fig. 1c). This module captures the deep connection between the drug and the target from multiple perspectives, enabling fine-grained and physiologically plausible feature interaction. Finally, the fused representations are concatenated and passed into a Fully Connected Neural Network (FCNN) to regress the binding affinity score.

Multi-scale feature fusion

Existing methods struggle to capture the critical spatial interactions between non-adjacent residues within drug-target binding pockets [24]. Therefore, this study aims to develop a feature encoding mechanism that simultaneously integrates local structure and long-range spatial dependencies to more comprehensively characterize binding sites. To overcome the limitations in feature extraction, we constructed a multi-scale feature fusion with an adjustable sliding window (Fig. 2). This encoder adopts a sequential hybrid architecture, with its core being a module that progressively fuses local perception with global interaction. Specifically, this branch first performs local perception through a one-dimensional convolutional layer (Conv1D)[25]: a convolutional kernel of size k acts as a sliding window, forcing the model to focus on contiguous adjacent segments of the amino acid sequence. This captures local contextual information represented by secondary structures (such as α-helix and β-sheet) or locally conserved motifs [26].

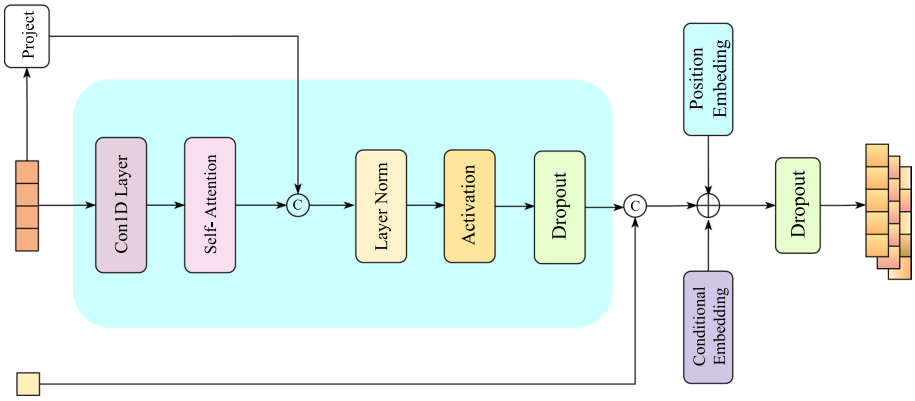

Fig. 2 Structure of the multi-scale feature fusion. This module operates in a single sequential branch to capture diverse feature granularities. The 1D Convolutional Layer extracts local sequential patterns, while the Self-Attention mechanism captures global long-range dependencies. The outputs are concatenated and normalized via Layer Normalization. Additionally, Conditional Embedding and Position Embedding are incorporated to preserve spatial and contextual information before the final Dropout layer

The input sequence is $\mathrm{X} \in \mathbb{R}^{L \times d}$, where L is the sequence length and d is the feature dimension.

$$\mathrm{H}_{\mathrm{local}} = \mathrm{Conv1D}(\mathrm{X}, \mathrm{W}_c, k) \tag{1}$$

Wc: Convolution kernel weights. k: Convolution kernel size (Patch Kernel Size). The paper mentions using $k \in \{9, 7, 5, 3\}$ in experiments. Hlocal: Feature patches capturing local contextual information. Subsequently, these local features $\mathrm{H}_{\mathrm{local}}$ are fed into a Transformer layer for global interaction. Through its self-attention mechanism, the Transformer can model long-range dependencies between different local patches, thereby inferring potential spatial interactions formed by distantly separated residues in the sequence [27].

$$\begin{aligned} \mathrm{H}_{\mathrm{attn}} &= \mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}\left(\frac{\mathrm{QK}^T}{\sqrt{d_k}}\right)\mathrm{V} \\ \mathrm{Q} &= \mathrm{H}_{\mathrm{local}}\mathrm{W}^Q, \quad \mathrm{K} = \mathrm{H}_{\mathrm{local}}\mathrm{W}^K, \quad \mathrm{V} = \mathrm{H}_{\mathrm{local}}\mathrm{W}^V \end{aligned} \tag{2}$$

• $\mathrm{W}^Q, \mathrm{W}^K, \mathrm{W}^V$: Projection matrices for queries, keys, and values, respectively. • d_k: Scaling factor. To capture latent associations between the target and the ligand, we fuse the contextual representation of the protein with the global representation of the drug. Simultaneously, we preserve the temporal information and specific constraints of the sequences by superimposing positional embeddings and conditional encoding, thereby optimizing the input distribution for subsequent modules.

$$\begin{aligned} \mathrm{H}_{\mathrm{attn}} &= \mathrm{SelfAttention}(\mathrm{H}_{\mathrm{local}}) \\ \mathrm{H}_{\mathrm{mix}} &= \mathrm{LayerNorm}(\mathrm{Concat}(\mathrm{H}_{\mathrm{attn}}, \mathrm{Project}(\mathrm{X}))) \\ \mathrm{Z} &= \mathrm{Dropout}(\mathrm{Activation}(\mathrm{H}_{\mathrm{mix}})) + \mathrm{E}_{\mathrm{pos}} + \mathrm{E}_{\mathrm{cond}} \end{aligned} \tag{3}$$

This design enables the model to understand protein sequences collaboratively at both the micro-scale (individual residues and their immediate neighbors) and the meso-scale (local structural fragments and their long-range associations). This enhances the expressiveness of the generated residue-level features, laying the foundation for the subsequent accurate prediction of drug-target binding sites. To investigate the relationship

between convolution kernel size and multi-feature extraction, we performed experiments with $k \in \{9, 7, 5, 3\}$. This study experimentally determined $k = 5$ to be the optimal local scale for binding site detection. This aligns with the physical scale of protein local structural units, effectively balancing the capture of local motifs with the suppression of background noise. This result strongly validates the effectiveness and necessity of the proposed local perception and global interaction mechanism for extracting the most relevant biophysical features.

Binding site-guided strategy

To address the issue of overly stringent thresholds in traditional hard masks that can lead to critical information loss, this study proposes a binding sites-guided strategy with the other Top-K contexts (Fig. 3). This approach not only preserves high-probability candidate regions but also incorporates contextual information to enhance the characterization of pocket physicochemical environments. Even in weak-signal scenarios, it adaptively focuses on key areas while effectively filtering noise.

Specifically, the mechanism employs a conditionally dual-path masking strategy. Let $\mathrm{P} = \{p_1, p_2, ..., p_L\}$ denote the predicted binding site probability sequence, τ be the decision threshold, and K be the retention count selected based on sequence length L ($K = \lfloor \gamma \cdot L \rfloor$). The construction of the masking index set Ω_{select} follows this logic:

1. **Fidelity & correction mode:** When explicit binding signals are detected (i.e., $p_i \geq \tau$ exists), the model retains all predicted positive sites (fidelity) while additionally incorporating the Top-K single sites with the highest probabilities from non-binding regions. This operation aims to supplement structural context and reduce the risk of overlooking potential candidate sites.
2. **Denoising focus mode:** When no binding signals are detected, the model automatically switches modes, forcing retention only of the top K% of regions with the highest relative probabilities across the entire sequence. This prevents attention dispersion in the absence of precise localization and maximally suppresses background noise.

The selection process and final feature weighting formula are formalized as follows:

$$\begin{aligned} \Omega_{\text{pos}} &= \{i \mid p_i \geq \tau\} \\ \Omega_{\text{select}} &= \begin{cases} \Omega_{\text{pos}} \cup \text{TopK}\left(\{p_j \mid j \notin \Omega_{\text{pos}}\}, K\right), & \text{if } |\Omega_{\text{pos}}| > 0 \\ \text{TopK}\left(\mathrm{P}, K\right), & \text{otherwise} \end{cases} \\ \mathrm{H}'_i &= \mathrm{H}_i \cdot p_i \cdot \mathbb{I}(i \in \Omega_{\text{select}}) \end{aligned} \tag{4}$$

Here, H_i denotes the original feature, $\mathbb{I}(\cdot)$ is the indicator function, and H'_i represents the feature after soft masking filtering. To investigate parameter sensitivity, we conducted

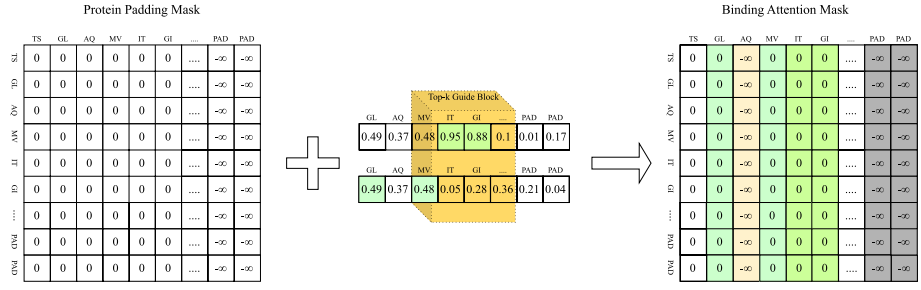

Fig. 3 Schematic diagram of the binding sites-guided strategy with the other Top-K contexts

comparative experiments on the Davis [28] and Kiba [29] datasets with selected K values: $K \in \{30\%, 50\%, 70\%, 100\%\}$ to determine the optimal balance between focused representation and denoising.

Building upon this, the model deeply integrates the extracted protein contextual features H' with the global features of the drug. This module effectively reduces the risk of overlooking potential candidate sites by dynamically filtering high-confidence residue combinations while preserving critical contextual information, thereby providing a reliable basis for subsequent drug-target affinity screening.

Dual cross-interaction block

To address the limitation of traditional models in feature fusion, where multi-scale interaction is often insufficient [30], this study proposes a dual cross-interaction block (Fig. 4). First, we employ the upstream predicted binding site vector V_{bind} as prior knowledge to perform a soft-weighting process on the protein features. Let $H_{prot} \in \mathbb{R}^{L_p \times d}$ denote the encoded protein sequence and $H_{smi} \in \mathbb{R}^{L_s \times d}$ denote the encoded drug molecule sequence. The predicted binding site logits are transformed via a sigmoid function to obtain the gating coefficients G, which are then applied to the protein features:

$$\begin{aligned} G &= \sigma(V_{bind}) \in \mathbb{R}^{L_p \times 1} \\ \tilde{H}_{prot} &= H_{prot} \odot G \end{aligned} \tag{5}$$

where $\odot$ denotes element-wise multiplication. The resulting $\tilde{H}_{prot}$ retains only features highly relevant to potential binding sites, effectively suppressing noise from non-binding regions and providing high signal-to-noise ratio inputs for subsequent feature

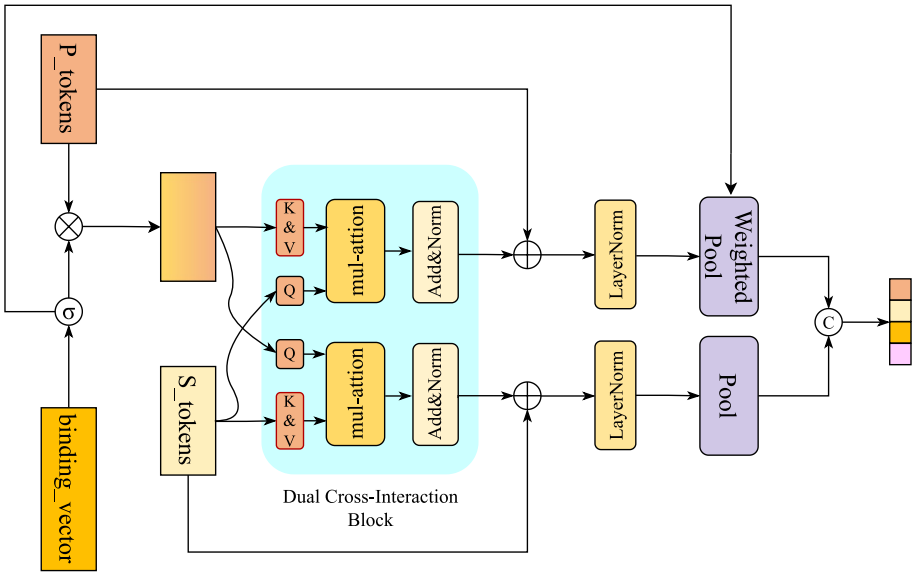

Fig. 4 The architecture of the Dual Cross-Interaction Block. This component facilitates information exchange between protein tokens ($P_{\mathbf{tokens}}$) and SMILES tokens ($S_{\mathbf{tokens}}$). The Dual Cross-Interaction Block employs a symmetric query-key-value (*Q*, *K*, *V*) structure to update the representation of each modality based on the other. The $\mathbf{binding\_vector}$ modulates the input $P_{\mathbf{tokens}}$ via a gating mechanism (σ). Finally, the Fusion Block integrates the features using both standard and Weighted Pooling followed by concatenation to form the final comprehensive representation

interaction. Second, we design a dual cross-interaction pathway consisting of two complementary views:

1. Ligand-view pathway (LVP): Using drug features H_{smi} as queries, and the gated protein features $\tilde{H}prot$ as both keys and values. This pathway simulates how a drug perceives and focuses on specific binding sites within the protein:

$$H_{smi}^{inter} = \text{LayerNorm}(H_{smi} + \text{MHA}(Q = H_{smi}, K = \tilde{H}_{prot}, V = \tilde{H}_{prot})) \tag{6}$$

2. Protein-view pathway (PVP): Using the gated protein features $\tilde{H}prot$ as queries, and drug features H_{smi} as both keys and values. This pathway models how binding sites adjust their feature representations according to the physicochemical properties of the drug:

$$H_{prot}^{inter} = \text{LayerNorm}(\tilde{H}_{prot} + \text{MHA}(Q = \tilde{H}_{prot}, K = H_{smi}, V = H_{smi})) \tag{7}$$

Finally, to generate a global representation for affinity prediction, sequence-level information needs to be aggregated. Since simple average pooling tends to dilute critical binding site information, we adopt a gated weighted pooling strategy for the interacted protein features H_{prot}^{inter}:

$$h_{prot}^{pool} = \frac{\sum_{i=1}^{L_p}(H_{prot,i}^{inter} \cdot G_i)}{\sum_{i=1}^{L_p} G_i + \epsilon} \tag{8}$$

This ensures that the final protein vector h_{prot}^{pool} remains highly focused on binding pocket regions. Ultimately, we concatenate the original global features ([CLS] tokens) with the interaction-enhanced local pooled features to form a fused vector V_{final}, which integrates multi-scale and multi-view information. This vector is then fed into a fully connected neural network (FCNN) for affinity regression:

$$V_{final} = \text{Concat}([h_{prot}^{cls}, h_{smi}^{cls}, h_{smi}^{inter[0]}, h_{prot}^{pool}]) \tag{9}$$

Experiments and results

Dataset

Consistent with current mainstream competitive approaches [31], we utilized two widely used benchmark datasets Davis and KIBA to train, validate, and test our model. The Davis dataset contains selectivity analysis data specific to the human catalytic protein kinase group. This dataset covers interactions between 442 kinases and 72 kinase inhibitors, originally comprising 31,824 drug-target interaction (DTI) pairs. Affinity values are measured by dissociation constants (Kd). To adapt to regression tasks and reduce variance in the data distribution, we transformed Kd values into logarithmic space (pK_d) using the formula $pK_d = -\log_{10}(K_d/10^9)$. To standardize feature counts and avoid loss of relevant sequence information, we performed length preprocessing on the data following the TAG-DTA strategy. We fixed the length of target protein sequences between 264 and 1400 amino acid residues and restricted SMILES string lengths to 38–72 characters. After preprocessing and deduplication, the final Davis dataset used for experiments comprised 29,187 samples. To further evaluate model generalization, we incorporated

the KIBA dataset. Derived from Tang et al.'s research, this dataset addresses inconsistencies in bioactivity metrics (K_i, K_d, IC_{50}) across diverse sources. KIBA scores integrate these varied bioactivity types into a comprehensive drug-target bioactivity matrix. The original KIBA dataset comprises 467 target proteins and 52,498 ligands. Following similar data cleaning criteria, we treated the KIBA score as the prediction target and excluded samples with excessively short or long sequences (target sequences capped at 1000 characters, SMILES strings capped at 100 characters). The final constructed experimental KIBA dataset comprises 229 target proteins and 2,111 drug molecules, totaling 118,254 validated interaction pairs. For both datasets, we employed either random splitting or chemogenomic representative splitting to partition data into training and testing sets, ensuring robustness in experimental evaluation. To train the 1D binding site classifier, ground-truth binding site labels were required. In this study, we adopted the comprehensive 1D binding pocket dataset curated by the TAG-DTA baseline. These labels were originally derived by extracting authentic 3D protein-ligand complexes from PDB-based databases (scPDB, PDBBind, and BioLiP) and mapping the spatial interacting residues onto 1D UniProt sequences to form binary vectors. By cross-referencing our affinity datasets with this binding pocket database via protein sequence alignment, we determined the exact annotation coverage. Specifically, in our constructed KIBA dataset, 163 out of the 228 unique target proteins possess explicit ground-truth binding site labels, corresponding to a coverage rate of 71.49.The relevant information is shown in Table 1.

Evaluation metrics

To evaluate the performance of various models, our study employed regression metrics for drug-target affinity (DTA) prediction, namely the Root Mean Squared Error (RMSE) loss, the Concordance Index (CI), and the coefficient of determination (r^{2m}). RMSE measures the error between measured and predicted values, while CI calculates the probability of concordance between pairs of predictions and measurements. For binding site prediction, which is a binary classification task, performance was assessed using five key metrics: Accuracy (ACC), Matthews Correlation Coefficient (MCC), Precision, Recall, and the F1-Score.

Accuracy represents the proportion of correctly predicted samples among the total samples. A higher accuracy indicates a greater overall correct prediction rate by the model. However, in datasets with class imbalance, relying solely on accuracy may not fully reflect the model's true performance. The Matthews Correlation Coefficient (MCC) is a comprehensive measure for binary classification models. MCC values range from -1 to +1, where +1 indicates perfect prediction, -1 indicates total disagreement, and 0 signifies prediction no better than random. A value closer to +1 denotes stronger overall discriminative power of the model on both positive and negative samples, and it is insensitive to class imbalance.

Precision reflects the proportion of true positive predictions among all positive predictions made by the model. In the context of binding site detection, a higher precision

Table 1 Statistics of the proteins and binding site label coverage in the benchmark datasets

Dataset	Total unique proteins	Proteins with 1D binding labels	Coverage rate
KIBA	228	163	71.49%
Davis	423	153	36.17%

indicates a lower rate of false positives, meaning the model is highly reliable when it predicts a residue as a binding site. The F1-Score is the harmonic mean of Precision and Recall, serving as a balanced composite metric. A higher F1-Score indicates a more robust model.The calculation formulas for the aforementioned evaluation metrics are provided below

1. Root mean squared error (RMSE)

$$\mathrm{RMSE}(t,p) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - p_i)^2} \tag{10}$$

where t_i is the measured binding affinity, and where p_i is the predicted binding affinity.

2. Concordance index (CI)

$$\mathrm{CI} = \frac{1}{Z}\sum_{i,j:\, t_i \neq t_j} \sigma(t_i, t_j) \cdot h(p_i - p_j) \tag{11}$$

where t_i and p_i denote the measured binding affinity and model prediction of the i -th sample, respectively, and Z is a normalization constant. $\sigma(t_i, t_j)$ is an indicator function that equals 1 if $t_i > t_j$, and 0 otherwise. $h(x)$ is the Heaviside step function defined as:

$$h(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0.5, & \text{if } x = 0 \\ 0, & \text{if } x < 0 \end{cases} \tag{12}$$

3. Coefficient of determination ($r^{2\,m}$)

$$r^{2m} = 1 - \frac{\sum_{i=1}^{n}(t_i - p_i)^2}{\sum_{i=1}^{n}(t_i - \bar{t})^2} \tag{13}$$

where $\bar{t}$ is the mean of the measured binding affinities.

4. Accuracy (ACC)

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

where TP, TN, FP, and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively.

5. Matthews correlation coefficient (MCC)

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{15}$$

where TP, TN, FP, and FN are defined as above.

6. Precision

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{16}$$

where TP and FP are the numbers of true positives and false positives, respectively.

7. Recall (Sensitivity)

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

where TP and FN are the numbers of true positives and false negatives, respectively.

8. F1-Score

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

Settings of hyperparameters and experimental environment

All experiments in this study were conducted on a hardware platform equipped with four NVIDIA RTX 3090 (24GB) GPUs. The implementation was carried out using Python 3.9.6 and TensorFlow 2.8.0, with the training process facilitated by the Adam optimizer. The specific parameter configurations are detailed in Table 2.

Experiment 1: Multi-scale feature fusion for binding site detection and DTA prediction

To investigate the influence of local contextual encoding on both binding site detection and affinity prediction, we conducted a series of experiments with varying patch sizes (k-mer windows) within the multi-scale feature fusion module. The baseline model employs a standard sequential encoder without explicit local feature aggregation. We compared its performance against four improved variants, which integrate local convolutional blocks with kernel sizes of 9, 7, 5, and 3 residues, respectively.

As summarized in Table 3, the incorporation of local receptive fields consistently enhances performance across both binding prediction and affinity regression tasks on the Davis and KIBA datasets. Specifically, all k-mer improved models outperform the baseline in terms of binding vector metrics (ACC, Recall, Precision, F1, MCC) and affinity metrics (RMSE, CI). Among them, the 5-mer configuration achieves the best overall performance, attaining the highest scores in ACC, Recall, Precision, MCC, and CI on both datasets, while also yielding the lowest RMSE.We present the visualized performance results on the Davis and KIBA datasets in Figs. 5 and 6.

For binding site detection, the 5-mer model achieves ACC values of 0.8877 and 0.8878 on Davis and KIBA, respectively, along with balanced improvements in Recall and Precision. This suggests that a local window of five residues effectively captures meaningful structural motifs relevant to pocket formation, without introducing excessive noise from distal regions. Smaller windows (e.g., 3-mer) exhibit a slight decline in precision, likely

Table 2 DCI-SiteDTA Hyperparameter Settings

Parameter	Value
Batch size	32
1D binding pocket pre-train epochs	30
Binding affinity train epochs	3
1D binding pocket train epochs	1
Epochs[a]	500
Dropout rate	0.1

[a] Initial number of epochs to allow convergence of the model, where early stopping and model checkpoint over the training cycles were applied to avoid overfitting

Table 3 Performance comparison of different patch sizes on Davis and Kiba datasets (Best results are in bold).

Dataset	Method	Binding vector metrics					Affinity metrics	
		ACC	Recall	Precision	F1	MCC	RMSE	CI
Davis	Baseline	0.8628	0.7957	0.6868	0.7364	0.6836	0.4443	0.8990
	+ 9mer	0.8807	0.8152	0.7444	0.7764	0.7328	0.4271	0.9081
	+ 7mer	0.8840	0.8196	0.7557	0.7852	0.7407	0.4225	0.9062
	+ 5mer	**0.8877**	**0.8309**	**0.7562**	**0.7840**	**0.7412**	**0.4211**	**0.9096**
	+ 3mer	0.8774	0.8132	0.7322	0.7689	0.7231	0.4235	0.9087
Kiba	Baseline	0.8650	0.8016	0.6839	0.7366	0.6846	0.5179	0.8300
	+ 9mer	0.8841	0.8247	0.7381	0.7776	0.7339	0.5157	0.8327
	+ 7mer	0.8852	0.8301	0.7284	0.7745	0.7303	0.5155	0.8320
	+ 5mer	**0.8878**	**0.8391**	**0.7572**	**0.7855**	**0.7430**	**0.5148**	**0.8329**
	+ 3mer	0.8836	0.8251	0.7350	0.7763	0.7319	0.5193	0.8319

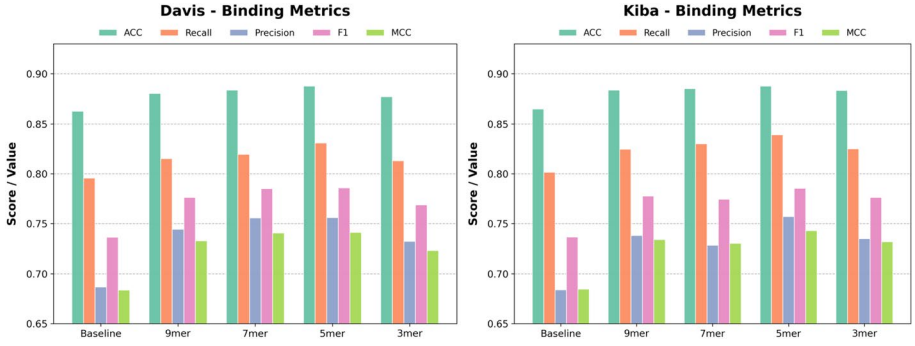

Fig. 5 Performance with binding site information on Davis and KIBA datasets

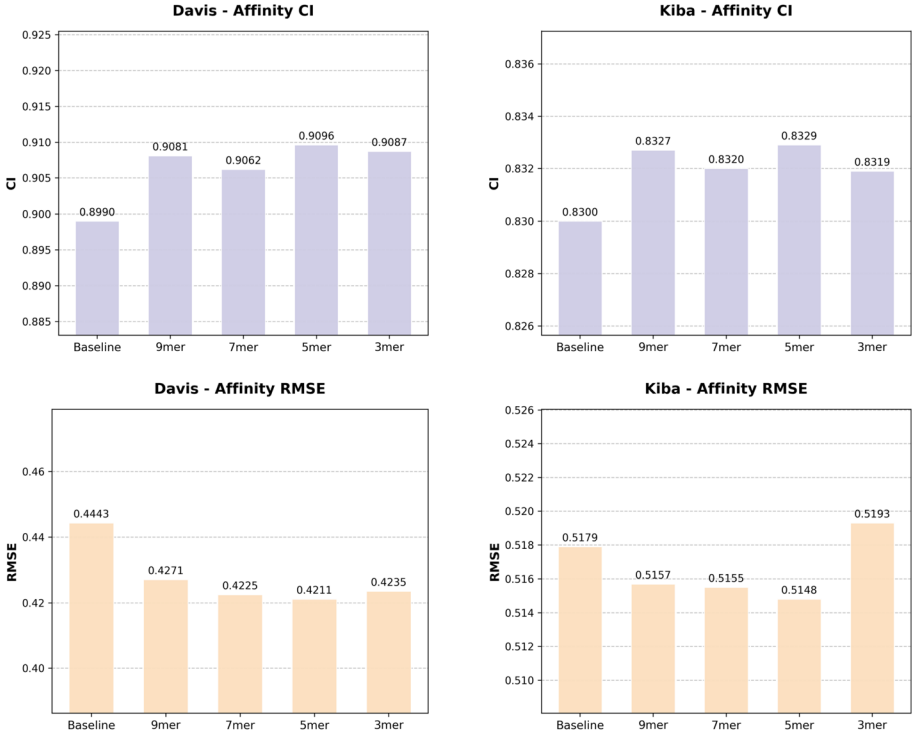

Fig. 6 Affinity prediction performance on Davis and KIBA datasets

due to insufficient context, while larger windows (9-mer) show marginally lower recall, possibly attributable to over-smoothing or inclusion of irrelevant residues.we find that a single turn of α-helix spans approximately 3.6 residues [32] and β-turns is typically 4 residues [33]. Therefore, a 3-mer may be too short to capture the structural semantics, and a 7-mer could risk on introducing background noise from adjacent non-interacting regions, thereby diluting the highly localized signals of binding pockets.

In affinity prediction, the 5-mer model yields RMSE values of 0.4211 (Davis) and 0.5148 (KIBA), alongside CI scores of 0.9096 and 0.8329, respectively. The consistent superiority of the 5-mer setting indicates that local structural patterns centered on binding residues provide discriminative features that enhance the model's ability to estimate binding strength. This observation aligns with biophysical intuition, as many functional protein–ligand interactions are mediated by local clusters of 4–6 residues forming key contact surfaces.

These results underscore the importance of local sequence context in joint binding–affinity modeling and demonstrate that an intermediate receptive field (5-mer) offers an optimal trade-off between locality and contextual breadth for the tasks studied. Future extensions could explore adaptive or multi-kernel architectures to dynamically adjust the receptive field based on sequence characteristics.

Experiment 2: Effects of the binding site-guided strategy

To balance the need for focused feature representation against the risk of losing critical binding context, we introduced a binding site-guided strategy with other Top-*K* contexts. This strategy dynamically selects subsets of residue positions based on predicted binding probabilities, integrating contextual regions while preserving high-confidence binding sites to mitigate the risk of overlooking potential candidate sites. We evaluated four Top-*K* retention ratios $K \in \{0.3, 0.5, 0.7, 1.0\}$ under the optimal 5-mer local encoding setting.

As shown in Table 4, the choice of K markedly influences affinity prediction performance on both the Davis and KIBA datasets. On Davis, setting $K = 0.5$ achieves the best trade-off, yielding the lowest MSE (0.171) and RMSE (0.414), along with the highest CI (0.919). This represents a clear improvement over the baseline 5-mer model ($\mathrm{RMSE} = 0.421,\ \mathrm{CI} = 0.910$) and demonstrates the effectiveness of the masking

Table 4 Impact of different Top-K ratios on model performance under 5mer patch size setting (Best results are in bold).

Dataset	Method	Affinity metrics		
		MSE	RMSE	CI
Davis	Baseline	0.186	0.431	0.899
	5mer Improved	0.177	0.421	0.910
	+ $K = 0.3$	0.174	0.417	0.916
	+ $K = 0.5$ (Ours)	**0.171**	**0.414**	**0.919**
	+ $K = 0.7$	0.172	0.414	0.917
	+ $K = 1.0$	0.173	0.416	0.913
Kiba	Baseline	0.268	0.518	0.830
	5mer Improved	0.265	0.515	0.833
	+ $K = 0.3$	0.276	0.526	0.831
	+ $K = 0.5$ (Ours)	**0.263**	**0.513**	**0.833**
	+ $K = 0.7$	0.266	0.516	0.832
	+ $K = 1.0$	0.273	0.522	0.830

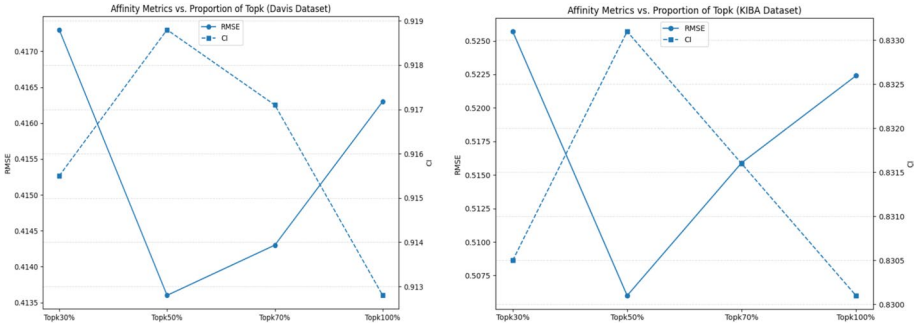

Fig. 7 Comparative visualization of different top-K ratios on model performance across the Davis and KIBA Datasets (with 5mer Patches)

Table 5 Performance comparison with and without the 1D binding site classifier (Best results are in bold).

Dataset	1D binding classifier	Affinity prediction		
		MSE (↓)	RMSE (↓)	CI (↑)
Davis	On (ours)	**0.177**	**0.4211**	**0.9096**
	Off	0.202	0.4490	0.9010
Kiba	On (ours)	**0.265**	**0.5148**	**0.8329**
	Off	0.265	0.5150	0.8310

"On" indicates the complete model with the 1D binding site classification block enabled, while "Off" represents the model where this guidance module is disabled, forcing the model to rely solely on global features without targeted pocket priors

mechanism. Higher retention ratios ($K = 0.7, 1.0$) lead to slight performance degradation, likely due to increased noise from non-binding regions. Conversely, a stricter selection ($K = 0.3$) also underperforms, suggesting that over-aggressive filtering may discard structurally relevant context.

On the KIBA dataset, a similar trend is observed: $K = 0.5$ again delivers the strongest results, with $\mathrm{RMSE} = 0.513$ and $\mathrm{CI} = 0.833$, outperforming both more restrictive ($K = 0.3$) and more inclusive ($K = 1.0$) settings. The visualization of this pattern across both datasets is presented in Fig. 7.

The consistent superiority of $K = 0.5$ across both datasets indicates that retaining approximately half of the most relevant residue positions optimally balances signal preservation against noise suppression. This ratio allows the model to maintain high-confidence binding sites while incorporating sufficient flanking residues to capture local physicochemical and structural context—an important factor for accurate affinity estimation.

In summary, the binding sites-guided strategy with the other Top-K contexts provides a robust and tunable strategy to guide the model's attention toward binding-relevant regions. The empirical results confirm that an intermediate retention ratio ($K = 0.5$) generalizes well across different datasets and contributes to improved prediction stability and accuracy.To evaluate the impact of cascade errors caused by upstream site detection bias, we simulated extreme errors by completely disabling the one-dimensional classifier. Results on the Davis dataset show that completely removing site guidance increases the RMSE from 0.421 to 0.449 (Table 5.) These findings indicate that even in the worst-case scenario of severe detection bias or complete failure, the baseline global interaction path in the DCI module effectively prevents catastrophic downstream degradation. Our architecture has a certain degree of tolerance to upstream errors.

Experiment 3: The dual cross-interaction for DTA prediction

To effectively integrate protein and ligand representations for affinity prediction, we compared four feature-fusion strategies: (1) simple concatenation, (2) gated fusion, (3) single-directional cross-attention [34], and (4) the proposed dual cross-interaction block. The latter employs a binding-guided gating module followed by dual cross-interaction pathways (protein-to-ligand and ligand-to-protein), simulating the induced-fit process between the target pocket and the drug molecule.

As summarized in Table 6, the dual cross-interaction strategy consistently achieves the best performance on both the Davis and KIBA datasets. On Davis, it attains the lowest MSE (0.170) and RMSE (0.412) while maintaining a high CI (0.919), outperforming all other fusion approaches. Similarly, on KIBA, it yields the lowest MSE (0.262) and RMSE (0.512) along with the highest CI (0.835). In contrast, single-directional cross-attention shows a noticeable drop in performance (e.g., RMSE increases to 0.437 on Davis and 0.525 on KIBA), indicating that unidirectional feature alignment is insufficient to capture the mutual adaptation between the two binding partners.

The superiority of the dual cross-interaction block can be attributed to its ability to model bidirectional context-aware interactions. By using the predicted binding vector as a soft gating signal, the model first suppresses noise from non-binding regions of the protein sequence. It then performs two complementary attention passes: (1) Ligand-to-protein attention allows the drug representation to focus on the most relevant binding residues. (2) Protein-to-ligand attention enables the protein to adjust its feature expression based on the chemical properties of the drug.

This dual cross flow mimics the real-world induced-fit dynamics, where both the ligand and the binding pocket undergo conformational adjustments. The final pooled representation, obtained through binding-gated weighted pooling, retains high signal-to-noise ratio and multi-scale information, which is then fed into a fully connected network for affinity regression.To rigorously isolate the intrinsic contribution of the interaction mechanisms, independent of our proposed upstream modules, we conducted a granular ablation study on a "pure baseline." In this setting, the multi-scale feature fusion (5-mer) and the binding site-guided strategy (K=0.5) were completely removed. As shown in Table 7, even when operating on basic sequential features, the Dual Cross-Interaction block consistently outperforms simple Concatenation, Gated Fusion, and standard Cross-Attention. This demonstrates that the bidirectional cross-attention pathways (LVP and PVP) are fundamentally superior at capturing the mutual structural adaptations between drugs and targets, confirming that the DCI module

Table 6 Performance comparison with different fusion strategies on Davis and Kiba datasets (Best results are in bold).

Dataset	Fusion strategy	Affinity prediction		
		MSE (↓)	RMSE (↓)	CI (↑)
Davis	Concatenation	0.171	0.414	0.919
	Gated fusion	0.195	0.442	0.913
	Cross-attention	0.191	0.437	0.911
	Dual cross-interaction	**0.170**	**0.412**	**0.919**
Kiba	Concatenation	0.263	0.513	0.833
	Gated fusion	0.267	0.517	0.828
	Cross-attention	0.276	0.525	0.832
	Dual cross-interaction	**0.262**	**0.512**	**0.835**

"Concatenation" represents the base model using simple feature concatenation (5-mer + $K = 0.5$)

Table 7 Performance comparison of fusion strategies under the pure baseline setting (Best results are in bold).

Dataset	Fusion strategy	Affinity prediction		
		MSE (↓)	RMSE (↓)	CI (↑)
Davis	Concatenation	0.197	0.444	0.899
	Gated fusion	0.203	0.451	0.892
	Cross-attention	0.193	0.439	0.910
	Dual cross-interaction	**0.179**	**0.423**	**0.912**
Kiba	Concatenation	0.268	0.5179	0.830
	Gated fusion	0.271	0.521	0.827
	Cross-attention	0.278	0.527	0.829
	Dual cross-interaction	**0.263**	**0.513**	**0.835**

All fusion strategies in this table are evaluated under the pure baseline setting (without 5-mer multi-scale feature fusion and $K = 0.5$ binding site guidance)

provides substantial performance gains completely independent of the other architectural enhancements.

These results confirm that explicitly modeling mutual attention between drug and target is crucial for accurate affinity prediction. The dual cross-interaction block not only improves quantitative metrics but also enhances the interpretability of the model, as the attention weights can be visualized to identify key interacting residues and atoms–a valuable feature for guiding further drug optimization.

Furthermore, it is crucial to highlight the technical distinctions between our DCI block and the attention mechanisms employed in recent binding-guided models like TAG-DTA. While previous methods predominantly rely on unidirectional cross-attention or simple concatenation followed by self-attention, they often fail to capture the bidirectional structural adaptations during drug-target binding. Our DCI block fundamentally overcomes this by introducing two symmetrical pathways (LVP and PVP) to computationally mimic the biophysical "induced-fit" paradigm, allowing localized protein features and drug features to iteratively refine each other. Additionally, unlike standard models that use average pooling or a [CLS] token readout–which can severely dilute localized binding signals–we engineered a unique gated weighted pooling strategy (Eq. 8). This ensures that the final aggregated vector maintains a high signal-to-noise ratio strictly focused on the binding interface. As evidenced in Table 7,this specific dual-attention and weighted pooling architecture substantially outperforms conventional cross-attention strategies.

Experiment 4: Performance comparison to recent works

To comprehensively evaluate the predictive capability of our proposed framework, we compared it with ten state-of-the-art methods on the Davis dataset, a widely recognized benchmark for drug–target affinity prediction. The competing methods span a range of architectures, including convolutional networks (DeepDTA, Sim-CNN-DTA[35]), graph neural networks (GraphDTA-GCNNet[18]), Transformers (TransformerCPI, DTITR[36], TAG-DTA), and hybrid designs (DeepCDA, HyperAttentionDTI[37], MFR-DTA[38], PMMR[39]).

As shown in Table 8, our model achieves the best performance across all four evaluation metrics: MSE (0.180), RMSE (0.424), CI (0.916), and R^2 (0.790). These results represent a consistent improvement over the strongest recent baselines. For instance, compared to the recent TAG-DTA (2024) and PMMR (2025) models, our method

Table 8 Performance comparison of different models on the Davis dataset (Best results are in bold).

Method	MSE (↓)	RMSE (↓)	CI (↑)	r^2 (↑)
DeepDTA (2018)	0.235	0.485	0.871	0.707
GraphDTA-GCNNet (2020)	0.322	0.566	0.850	0.598
TransformerCPI (2020)	0.285	0.533	0.839	0.493
DeepCDA (2020)	0.232	0.482	0.879	0.710
Sim-CNN-DTA (2021)	0.276	0.525	0.869	0.656
HyperAttentionDTI (2022)	0.241	0.491	0.879	0.699
DTITR (2022)	0.216	0.465	0.880	0.730
MFR-DTA (2023)	0.221	0.470	0.905	0.705
TAG-DTA (2024)	0.199	0.446	0.898	0.752
PMMR (2025)	0.186	0.431	0.910	0.781
Ours	**0.180**	**0.424**	**0.916**	**0.790**

Table 9 Performance comparison under the **Cold-Start** setting on the Kiba dataset. The best results are highlighted in bold (Best results are in bold).

Method	MSE (↓)	RMSE (↓)	CI (↑)
DeepDTA (2018)	0.350	0.592	0.771
GraphDTA (2021)	0.360	0.600	0.755
MolTrans (2021)	0.330	0.574	0.780
TAG-DTA (2024)	0.278	0.527	0.831
GramSeq-DTA (2025)	0.277	0.526	0.823
Ours	**0.276**	**0.525**	**0.833**

reduces RMSE by approximately 5% and 1.6%, respectively, while also raising CI and R^2 to new highs. Notably, the progressive improvement from earlier models (e.g., DeepDTA, $\mathrm{RMSE} = 0.485$) to our approach illustrates the cumulative impact of architectural advances—especially the integration of multi-scale feature fusion, a binding sites-guided strategy with the other Top-K contexts, and dual cross-interaction block.Furthermore, to statistically evaluate the predictive performance differences between the proposed model (DCI-SiteDTA) and key benchmark models, we conducted pairwise comparisons of CI scores using paired *t*-tests (95% confidence intervals). The results indicate that the proposed DCI-SiteDTA framework outperforms all benchmark models in terms of CI scores, with all pairwise comparison *p*-values falling below 0.05 and 0.01. Notably, the two-tailed *p*-value from the paired comparison between DCI-SiteDTA and TAG-DTA was 6.956×10^{-4}.

Furthermore, to assess the generalizability of the model on unseen data, we conducted experiments on the KIBA data set under the **cold** start setting (Table 9). In this challenging scenario, our model still maintains robust performance, performing competitively against existing methods with comparatively lower error rates ($\mathrm{MSE} = 0.276$, $\mathrm{RMSE} = 0.525$) and a higher concordance index ($\mathrm{CI} = 0.833$).

While GramSeq-DTA (2025)[40] and TAG-DTA (2024) achieve competitive results with MSE of 0.277 and 0.278, respectively, our model shows further improvement in accuracy. Specifically, the CI of our model reaches 0.833, which is moderately better than the 0.831 of TAG-DTA and 0.823 of GramSeq-DTA. These results suggest that the learned representations may generally generalize to novel compounds and proteins–a potentially useful property in practical drug-discovery applications. Recognizing the critical importance of model generalizability in practical drug discovery, we further evaluated DCI-SiteDTA on the Davis dataset under a stringent cold-start setting (evaluating

on unseen targets). As shown in Table 10, this scenario is considerably more challenging than random splitting. However, DCI-SiteDTA continues to exhibit superior robustness compared to generalization-focused baselines such as TAG-DTA and LKE-DTA[41]. By achieving an RMSE of 0.506 and a CI of 0.835, our framework demonstrates that the localized structural priors provided by the binding site detection module, combined with the bidirectional feature adjustment in the DCI block, effectively prevent overfitting and capture universally applicable drug-target interaction patterns.

Case study: Attention visualization and induced-fit mechanism

A core motivation behind the Dual Cross-Interaction (DCI) block is to computationally simulate the biophysical adaptive fitting process, where the binding pocket dynamically adjusts its conformation and physicochemical properties to accommodate a specific ligand. To validate whether our model successfully captures this phenomenon, we extracted and visualized the cross-attention weights from the Protein-View Pathway (PVP) of the DCI block. In the PVP, the gated protein features act as queries, while the drug SMILES features serve as keys and values. Thus, a high attention weight mathematically indicates that a specific protein residue is significantly altering its feature representation in response to a specific functional group of the drug. Figure 8 illustrates the PVP cross-attention heatmap for the well-known complex of the ABL1(E255K) mutant kinase and the inhibitor SKI-606 (Bosutinib). Rather than uniformly attending to the entire drug molecule, the protein residues exhibit highly specific, localized attention "hotspots" corresponding to established biophysical interactions: Halogen Group Recognition: The critical gatekeeper residue, T315, assigns an exceptionally high attention weight to the C l (chlorine atom) token of the drug. This mathematically mirrors the well-documented induced-fit pocket adaptation required to accommodate the bulky, hydrophobic halogen group of Bosutinib. Core Scaffold Anchoring: Residues in the hinge region, specifically F317 and M318, heavily concentrate their attention on the n c c c substructure, which corresponds to the quinoline core of the drug. This aligns perfectly with the biological formation of essential hydrogen bonds at the kinase hinge region. Mutant Site Interaction: The mutated residue K255 exhibits a strong attention focus on the $C\#N$ (nitrile group), capturing the specific steric and electrostatic interactions introduced by the $E255K$ mutation. These localized attention patterns provide concrete visual evidence that the PVP successfully models the adaptive fitting mechanism, allowing the target protein's representations to be structurally modulated by specific moieties of the drug molecule.

Table 10 Performance comparison under the **Cold-Start** setting on the Davis dataset (Best results are in bold).

Method	MSE (↓)	RMSE (↓)	CI (↑)
DeepDTA (2018)	0.424	0.651	0.783
Transformer CPI (2020)	0.343	0.586	0.706
DTITR (2022)	0.292	0.540	0.752
MFR-DTA (2023)	0.277	0.526	0.771
TAG-DTA (2024)	0.261	0.511	0.785
LKE-DTA (2025)	0.324	0.569	0.833
Ours	**0.256**	**0.506**	**0.835**

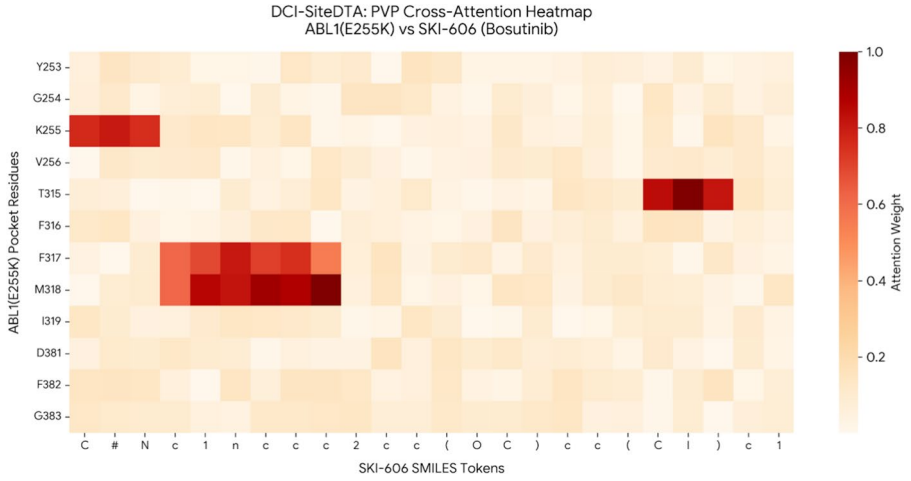

Fig. 8 PVP cross-attention heatmap for the ABL1(E255K) and SKI-606 (Bosutinib) complex. The y-axis represents the key pocket residues of the protein (Queries), and the x-axis represents the SMILES tokens of the drug (Keys/Values). Darker colors indicate higher attention weights, demonstrating that specific protein residues dynamically update their features in response to precise functional groups of the drug (e.g., T315 attending to the Chlorine atom), thereby computationally simulating the induced-fit process

Conclusions

In this study, we propose DCI-SiteDTA, a novel end-to-end deep learning framework that unifies binding site detection and affinity prediction. By integrating dual Transformer encoders for protein and ligand representation, a multi-scale feature fusion, a binding sites-guided strategy with the other Top-K contexts and the dual cross-interaction block, our model provides a unified approach to jointly infer binding sites and affinity values. The binding sites-guided strategy with the other Top-K contexts reduces the risk of overlooking potential candidate sites while preserving contextual information, whereas the dual cross-interaction fusion module simulates the induced-fit process to capture the fine-grained interaction mechanisms between drugs and targets.

Experimental evaluation on the well-established Davis and KIBA benchmarks demonstrates that our framework achieves competitive performance in both binding site classification and affinity prediction tasks. Notably, the model shows robustness under cold-start settings, indicating its practical potential in early-stage drug discovery where labeled data are scarce. Compared with various state-of-the-art methods, our approach improves key evaluation metrics, including mean squared error, root mean squared error, and concordance index.

Despite these encouraging results, the framework has certain limitations: (1) The current architecture primarily relies on sequential inputs. Although it leverages predicted binding pockets to guide affinity learning, it does not explicitly incorporate three-dimensional structural information. Future work will focus on extending the model to exploit geometric and surface features of proteins. (2) Recognizing the impact of surface pockets on DTA–particularly drug molecular profiles and the tertiary structure of targets–we note that leveraging advanced structural data, for instance, using graph neural networks (GNNs) to extract high-level structural features, represents a prominent and promising research direction for future work. (3) Our evaluation is based on the Davis and KIBA datasets, which are mostly kinases. From a biological perspective, data from non-kinase families may be further considered.

In summary, the integration of binding site detection and affinity prediction within an end-to-end framework offers a promising direction for improving the accuracy and interpretability of drug–target interaction modeling. With the increasing availability of protein structural data and advances in geometric deep learning, we anticipate that deeper fusion of sequence, structure, and dynamic information will substantially enhance the utility of such models in accelerating rational drug design.

Abbreviations

DTA	Drug-target affinity
DTI	Drug-target interaction
Top-K	The K highest-ranked results / items
SMILES	Simplified molecular input line entry system
CNN	Convolutional neural network
FCNN	Fully connected neural network
PWMLP	Position-aware multi-layer perceptron
MLP	Multilayer perceptron
LVP	Ligand-view representation fusion
PVP	Protein-view representation fusion
GNN	Graph neural network
MCC	Matthews correlation coefficient
MSE	Mean squared error
RMSE	Root mean square error
CI	Concordance index
F1	F1 score

Acknowledgements
Not applicable.

Author contributions
J.Z. contributed to the design of the work, the acquisition, analysis, interpretation of data, and drafted the manuscript; Y.Z. contributed to the conception, design and substantively revised the work; X.L. and B.W. contributed to analysis and provided substantive revisions. All authors reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (Grant No. 62302456).

Data availability
All datasets used in this study are publicly available. The Davis dataset and bindingsite set are referenced from the following link: https://github.com/larngroup/TAG-DTA. The original KIBA dataset was constructed by Tang et al. (Pharmacol. Res., 2014?, which can be accessed via the following link: https://github.com/hkmztrk/DeepDTA/tree/master/data. The dataset provided in this study can be accessed via the following link: https://github.com/zjy1007/DCI-SiteDTA.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no conflict of interest.

Consent for publication
Not applicable.

Received: 15 January 2026 / Accepted: 8 April 2026
Published online: 24 April 2026

References

1. Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. Br J Pharmacol. 2011;162:1239–49.
2. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of r&d costs. J Health Econ. 2016;47:20–33.
3. Gashaw I, Ellinghaus P, Sommer A, Asadullah K. What makes a good drug target? Drug Disc Today. 2012;17:24–30.
4. Shoichet BK. Virtual screening of chemical libraries. Nature. 2004;432:862–5.
5. Vamathevan J, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Disc. 2019;18:463–77.
6. Jumper J, et al. Highly accurate protein structure prediction with alphafold. Nature. 2021;596:583–9.
7. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinform. 2009;10:168.

8. Krivák R, Hoksza D. P2rank: Machine learning based tool for prediction of ligand binding sites from protein structure. J Cheminform. 2018;10:39.
9. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
10. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
11. He T, Heidemeyer M, Ban F, Cherkasov A, Ester M. Simboost: A read-across approach for predicting drug-target binding affinities using gradient boosting machines. J Cheminform. 2017;9:24.
12. Pahikkala T, et al. Toward more realistic drug-target interaction predictions. Brief Bioinform. 2015;16:325–37.
13. Öztürk H, Özgür A, Ozkirimli E. Deepdta: Deep drug-target binding affinity prediction. Bioinformatics. 2018;34:i821–9.
14. Pahikkala T, et al. Toward more realistic drug-target interaction predictions. Brief Bioinform. 2015;16:325–37.
15. Öztürk H, Özgür A, Ozkirimli E. Deepdta: Deep drug-target binding affinity prediction. Bioinformatics. 2018;34:i821–9.
16. Abbasi K, et al. Deepcda: Deep cross-domain adaptation for drug-target affinity prediction. Bioinformatics. 2020;36:4633–41.
17. Öztürk H, Ozkirimli E, Özgür A. Widedta: Prediction of drug-target binding affinity. Bioinformatics. 2019;35:2669–76.
18. Nguyen T, et al. Graphdta: Predicting drug–target binding affinity with graph neural networks. Bioinformatics. 2021;37:1140–7.
19. Jiang H, Wang Y, Meng C, Yang X, Sun Y. Dgraphdta: Prediction of drug-target binding affinity using graph neural networks. Bioinformatics. 2020;36:i801–9.
20. Chen L, et al. Transformercpi: Improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics. 2020;36:4406–14.
21. Monteiro NRC, Oliveira JL, Arrais JP. Tag-dta: Binding-region-guided strategy to predict drug-target affinity using transformers. Expert Syst Appl. 2024;238:122334.
22. Vaswani A, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30.
23. Devlin J, Chang M-W, Lee K Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019:4171–4186.
24. Jimenez J, Doerr S, Martinez-Rosell G, Rose AS, De Fabritiis G. Deepsite: Protein-binding site predictor using 3d-convolutional neural networks. Bioinformatics. 2017;33:3036–42.
25. Lee I, Keum J, Nam H. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput Biol. 2019;15:e1007129.
26. Schenone M, Dančík V, Wagner BK, Clemons PA. Target identification and mechanism of action in chemical biology and drug discovery. Nat Chem Biol. 2013;9:232–40.
27. Elnaggar A, et al. Prottrans: Toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell. 2021;44:7112–27.
28. Davis MI, et al. Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol. 2011;29:1046–51.
29. Tang J, et al. Making sense of large-scale kinase inhibitor bioactivity data. J Chem Inf Model. 2014;54:735–43.
30. Huang K, Xiao C, Glass LM, Zitnik M. Moltrans: molecular interaction transformer for drug-target interaction prediction. Bioinformatics. 2021;37:830–6.
31. Huang K, et al. Deeppurpose: a deep learning library for drug–target interaction prediction. Bioinformatics. 2020;36:5545–7.
32. Pauling L, Corey RB, Branson HR. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci. 1951;37:205–11.
33. Richardson JS. The anatomy and taxonomy of protein structure. Adv Protein Chem. 1981;34:167–339.
34. Bai P, Miljković F, John B, Lu H. Drugban: Interpretable bilinear attention network with domain adaptation for drug-target interaction prediction. Nat Mach Intell. 2023;5:322–32.
35. Zhao L, Wang J, Pang L, Liu Y, Zhang J. Sim-cnn-dta: similarity-based convolutional neural network for drug-target affinity prediction. Front Genet. 2021;12:747684.
36. Monteiro NRC, Oliveira JL, Arrais JP. Dtitr: end-to-end drug-target binding affinity prediction with transformers. Comput Biol Med. 2022;147:105772.
37. Zhao Q, Zhao H, Zheng K, Wang J. Hyperattentiondti: improving drug-protein interaction prediction by sequence-based deep learning with attention mechanism. Bioinformatics. 2022;38:655–62.
38. Hua Y, Song X, Feng Z, Wu X. Mfr-dta: a multi-functional and robust model for predicting drug-target binding affinity and region. Bioinformatics. 2023;39:btad056.
39. Ouyang X, et al. Improving generalizability of drug-target binding prediction by pre-trained multi-view molecular representations. Bioinformatics. 2025;41:btaf002.
40. Debnath K, Rana P, Ghosh P. Gramseq-dta: a grammar-based drug-target affinity prediction approach fusing gene expression information. Biomolecules. 2025;15:405.
41. Mou J, et al. Lke-dta: predicting drug-target binding affinity with large language model representations and knowledge graph embeddings. Mol Diversity. 2025:1–14.

Publisher's Note