Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Bidirectional Cross-Attention and Global Semantic Aggregation for Robust Drug–Target Interaction Prediction

Hao Pang
University of Nottingham Ningbo China
Fiseha Berhanu Tesema
Fiseha-Berhanu.Tesema@nottingham.edu.cn
University of Nottingham Ningbo China
Tianxiang Cui
University of Nottingham Ningbo China
Yuan Cheng
University of Nottingham Ningbo China

Research Article

Keywords: Drug–target interaction prediction, Transformer, Cross-attention, Deep learning, Bioinformatics, Molecular interaction modelling, Multi-scale pooling, Representation learning, Virtual screening, Computational drug discovery

Posted Date: July 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9809364/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Bioinformatics on September 21st, 2026. See the published version at https://doi.org/10.1186/s12859-026-06634-6.

Bidirectional Cross-Attention and Global Semantic Aggregation for Robust Drug–Target Interaction Prediction

Hao Pang[a], Fiseha Berhanu Tesema[a,*], Tianxiang Cui[a] and Yuan Cheng[a]

[a] School of Computer Science, Faculty of Science and Engineering, University of Nottingham Ningbo China, , Ningbo, 315100, Zhejiang, China

ARTICLE INFO

Keywords:
Drug–target interaction prediction
Transformer
Cross-attention
Deep learning
Bioinformatics
Molecular interaction modelling
Multi-scale pooling
Representation learning
Virtual screening
Computational drug discovery.

ABSTRACT

Drug–target interaction (DTI) prediction plays a critical role in virtual screening and computational drug discovery. Existing Transformer-based models such as MolTrans improve sequence representation learning but still rely on implicit interaction modelling, limited global semantic aggregation, and optimisation strategies that are sensitive to dataset imbalance and scarce-data settings. In this work, we propose a strengthened Transformer-based DTI framework integrating three key components: a bidirectional cross-attention mechanism for explicit modelling of drug–protein dependencies, a hybrid local–global representation scheme combining CNN-based interaction modelling with multi-scale global pooling, and a residual-enhanced classification head with stabilised optimisation using BCEWithLogitsLoss, AdamW, and cosine learning-rate scheduling. Comprehensive experiments on the BindingDB, BIOSNAP, and DAVIS benchmark datasets demonstrate that the proposed framework consistently outperforms MolTrans across multiple evaluation metrics, achieving up to 2.4 percentage-point improvements in AUROC and 11.8 percentage-point improvements in AUPRC. The largest gains are observed on the small and imbalanced DAVIS dataset, indicating improved robustness under scarce-data conditions. Ablation studies further confirm that cross-attention and global semantic aggregation contribute complementary performance benefits. Overall, the proposed framework provides a more expressive, interpretable, and robust approach for DTI prediction under realistic drug discovery settings.

1. Background

Drug discovery is a costly and time-intensive process, largely due to the need to experimentally screen vast chemical spaces of candidate compounds Hughes, Rees, Kalindjian and Philpott (2011); Broach, Thorner et al. (1996). Drug–target interaction (DTI) prediction offers an efficient computational alternative by estimating the likelihood that a small molecule binds to a specific protein target, thereby enabling virtual screening and drug repositioning Schenone, Dančík, Wagner and Clemons (2013). With the rapid growth of molecular and bioactivity data, deep learning has become a key approach for modelling complex biochemical interactions in silico Unterthiner, Mayr, Klambauer, Steijaert, Wegner, Ceulemans and Hochreiter (2014); Mayr, Klambauer, Unterthiner, Steijaert, Wegner, Ceulemans, Clevert and Hochreiter (2018).

Early deep learning models for DTI prediction typically encode drugs and proteins as sequences or hand-crafted descriptors, followed by convolutional or recurrent architectures Wen, Zhang, Niu, Sha, Yang, Yun and Lu (2017); Öztürk, Özgür and Ozkirimli (2018); Lee, Keum and Nam (2019). While effective at capturing local patterns, these methods often treat drug and protein representations independently and rely on simple fusion mechanisms such as concatenation. As a result, they fail to explicitly model the underlying interaction process, where specific substructures of a drug bind to particular regions of a protein Jia, Zhu, Ma, Cao, Li and Chen (2009).

Transformer-based approaches have recently improved DTI modelling by capturing long-range dependencies. Among them, MolTrans Huang, Xiao, Glass and Sun (2021) introduces a sub-structural tokenisation strategy based on frequent consecutive subsequences (FCS), enabling the model to learn representations from large-scale unlabeled data. It further constructs an interaction map using element-wise operations followed by convolution, achieving strong performance and improved interpretability. However, despite these advances, existing Transformer-based DTI models still rely on implicit and locally constrained interaction modelling. Specifically, their interaction mechanisms are limited to element-wise correlations and shallow convolutional processing, which restricts the ability to capture higher-order and long-range cross-modal dependencies between drug fragments and protein residues.

In addition, current architectures often lack explicit global semantic aggregation at the sequence level, making it difficult to integrate holistic contextual information beyond local interaction patterns. This limitation becomes particularly critical in small, noisy, or scarce-data settings commonly encountered in realistic drug discovery scenarios, where robust sequence-level representations are essential for stable generalisation. Furthermore, standard optimisation strategies based on Sigmoid and BCELoss can lead to unstable convergence and degraded performance under class imbalance, limiting the robustness of Transformer-based DTI models.In addition, current architectures often lack explicit global semantic aggregation at the sequence level, making it difficult to integrate holistic contextual information beyond local interaction patterns. This limitation becomes particularly critical in small, noisy, or scarce-data

*Corresponding author
Fiseha-Berhanu.Tesema@nottingham.edu.cn (F.B. Tesema)
ORCID(s):

settings commonly encountered in realistic drug discovery scenarios, where robust sequence-level representations are essential for stable generalisation. Furthermore, standard optimisation strategies based on Sigmoid and BCELoss can lead to unstable convergence and degraded performance under class imbalance, limiting the robustness of Transformer-based DTI models.

To address these limitations, we propose a strengthened Transformer-based DTI framework that explicitly models higher-order cross-modal dependencies through bidirectional alignment learning while incorporating hybrid local–global representation modelling and stabilised optimisation. Unlike prior approaches that rely on implicit interaction modelling, our method introduces a bidirectional cross-attention mechanism that allows drug tokens to attend to relevant protein residues and vice versa, enabling explicit learning of residue–substructure relationships. In parallel, we design a hybrid local–global representation scheme that combines convolutional interaction modelling with multi-scale global pooling to capture both fine-grained and holistic semantic information. Finally, we employ a residual-enhanced classification head together with stabilised optimisation strategies to improve convergence robustness, representation stability, and predictive performance under realistic and scarce-data drug discovery settings.

The main contributions of this work are as follows:

- We reformulate DTI prediction as a bidirectional cross-modal alignment problem and propose a cross-attention framework that explicitly models higher-order dependencies between drug substructures and protein residues.
- We introduce a hybrid local–global representation framework that combines CNN-based interaction modelling with multi-scale semantic aggregation to jointly capture fine-grained local interactions and global contextual dependencies.
- We employ a residual-enhanced classification head together with stabilised optimisation strategies, including BCEWithLogitsLoss, AdamW, and cosine learning-rate scheduling, to improve convergence stability and robustness under scarce-data settings.
- We conduct comprehensive evaluations under realistic drug discovery settings across BindingDB, BIOSNAP, and DAVIS benchmarks, demonstrating consistent improvements over MolTrans and strong robustness on small, imbalanced, and scarce-data datasets.

2. Related Work

Research on DTI prediction has progressed rapidly, from similarity-based approaches and feature-based machine learning models to deep neural networks and Transformer-based architectures Pahikkala, Airola, Pietilä, Shakyawar, Szwajda, Tang and Aittokallio (2015); Zheng, Ding, Mamitsuka and Zhu (2013); Mayr et al. (2018).

2.1. Similarity- and Feature-Based DTI Prediction

Classical computational DTI methods typically rely on chemical and biological similarity. Kernel-based regression and matrix factorisation approaches use drug and target similarity matrices to infer new interactions from known ones Pahikkala et al. (2015); Zheng et al. (2013). Feature-based methods instead construct numerical descriptors such as ECFP fingerprints Rogers and Hahn (2010) and PubChem features Bolton, Wang, Thiessen and Bryant (2008) for drugs and protein composition descriptors like CTD and PSC for proteins Dubchak, Muchnik, Holbrook and Kim (1995); Cao, Xu and Liang (2013). These features are then fed into shallow models such as logistic regression or gradient boosting He, Heidemeyer, Ban, Cherkasov and Ester (2017). Although such methods are efficient and interpretable, they depend heavily on hand-crafted features and often generalise poorly to unseen proteins or chemotypes Wen et al. (2017).

2.2. Deep Learning for DTI Prediction

Deep learning approaches aim to learn representations directly from raw SMILES and amino-acid sequences. DeepDTA Öztürk et al. (2018) applies CNNs to both drug and protein sequences to predict binding affinity. DeepConvDTI Lee et al. (2019) uses convolution and global max pooling on protein sequences combined with ECFP4 fingerprints for drugs. WideDTA Öztürk, Ozkirimli and Özgür (2019) extends this idea with multiple sequence representations, and several GNN-based methods such as GNN-CPI and GraphDTA incorporate molecular graphs for the ligand side Tsubaki, Tomii and Sese (2019); Nguyen, Le, Quinn, Nguyen, Le and Venkatesh (2021). These models improve performance over classical methods but still mainly rely on local receptive fields and often treat drug and protein encoders independently with only simple concatenation for fusion.

2.3. Transformer-Based DTI Models

Transformers Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser and Polosukhin (2017) have been adopted in chemoinformatics and bioinformatics due to their ability to capture long-range dependencies. MolTrans Huang et al. (2021) is a representative token-based Transformer model for DTI prediction. It decomposes drug and protein sequences into sub-structures using FCS mining, applies Transformer encoders to each modality and then constructs a pairwise interaction map through dot-product operations followed by a CNN and a shallow MLP decoder. MolTrans demonstrates strong performance and robustness across BIOSNAP, DAVIS and BindingDB, including scenarios with scarce labels and unseen drugs or targets. However, MolTrans still employs a relatively shallow interaction module and does not explicitly aggregate global sequence-level semantics; its training procedure is also not tailored for small, noisy datasets. These limitations motivate our design of a cross-attention-based interaction module with improved global pooling and stabilised optimisation.

2.4. Cross-Modal Attention in Bioinformatics

Cross-attention mechanisms have been widely used in multimodal tasks such as natural language inference, image–text retrieval and protein–ligand docking Gong, Luo and Zhang (2017). Recent works such as CAPLA and AttentionMGT-DTA apply cross-attention to model protein–ligand interactions and report improvements in binding-affinity prediction Jin, Wu, Chen, Pan, Wang, Xie, Quan and Lyu (2023); Li, Zhang, Guan and Zhou (2022). These results demonstrate that jointly attending over drug and protein representations can better capture inter-molecular dependencies. However, cross-attention has so far been less explored within token-based Transformer DTI frameworks like MolTrans, where the interaction module still largely relies on element-wise operations and CNNs.

2.5. Hybrid Local Convolution and Global Aggregation

Several studies have argued that purely local 2D convolution on interaction maps may lose global semantic information at the sequence level Li et al. (2022). To address this, hybrid architectures combine local convolutions with global pooling or attention mechanisms that summarise sequence-level context. Multi-scale pooling (average, max and attention) can provide complementary views of global semantics and help emphasise important regions. Our work follows this line by augmenting the interaction map with multi-scale global pooling over the cross-attention-enhanced token representations.

2.6. Deep Residual Classification Heads and Training Stability

Deep residual classification heads are widely used to increase discriminative capacity and improve gradient flow in vision and NLP models, and recent DTI models have started to adopt deeper MLP heads and shortcut connections to alleviate bottlenecks caused by shallow classifiers Mayr et al. (2018). For training stability, BCEWithLogitsLoss is preferred over a separate Sigmoid + BCELoss combination, and optimisers such as AdamW, together with learning-rate schedules and mixed-precision training, are recommended as best practice. We adopt these ideas to stabilise optimisation on small datasets (e.g. DAVIS) and to reduce variance across random splits.

2.7. Graph Transformers and Structured DTI Modelling

Graph-based Transformers such as Graphormer and its DTI variants integrate attention mechanisms with molecular graph structures, enabling simultaneous modelling of local chemical environments and global topology Gao, Zhang, Chen, Zhang, Wang, Wang, Li, Guo, Webb, Nguyen et al. (2024). Pocket-aware or subpocket-aware Transformers further incorporate protein structural information. These developments suggest that attention-based fusion and structure-aware encoders can substantially improve model expressiveness. Our framework can be seen as a step towards this direction within the MolTrans-style token-based setting, and it could be extended in future work to include graph or 3D structural encoders.

Overall, existing DTI models still face limitations in explicitly modelling cross-modal dependencies, integrating global semantic context, and maintaining stable optimisation under realistic and scarce-data settings. These limitations motivate the proposed bidirectional cross-attention framework with hybrid local–global representation learning and stabilised optimisation.

3. Methods

Figure 1 presents an overview of the proposed bidirectional cross-attention framework for robust drug–target interaction prediction. Building upon the sub-structural encoding strategy of MolTrans, our framework focuses on three key aspects: (1) explicit cross-modal interaction modelling, (2) hybrid local–global representation learning, and (3) stabilised optimisation. These components are designed to address the limitations of implicit interaction modelling, insufficient global context aggregation, and unstable training observed in existing Transformer-based DTI methods.

3.1. Problem Formulation

Given a drug SMILES sequence d and a protein amino-acid sequence p, the goal of DTI prediction is to learn a mapping function:

$$f_\theta : \mathcal{D} \times \mathcal{P} \to [0, 1],$$

which estimates the probability $\hat{y} = f_\theta(d, p)$ that a drug interacts with a target protein. Each training sample is defined as a tuple (d, p, y), where $y \in \{0, 1\}$ denotes the interaction label. The model is trained using a binary classification objective and evaluated using AUROC, AUPRC, and F1-score.

3.2. Transformer-Based Sequence Encoding

Drug and protein sequences are first decomposed into sub-structural tokens using the FCS mining strategy introduced in MolTrans. Each token is embedded into a continuous vector space of dimension E, combined with positional encodings to preserve sequential order.

Independent Transformer encoders are then applied to each modality to capture contextual dependencies:

$$\boldsymbol{D}' \in \mathbb{R}^{B \times L_d \times E}, \qquad \boldsymbol{P}' \in \mathbb{R}^{B \times L_p \times E},$$

where $\boldsymbol{B}$ denotes the batch size, and L_d and L_p denote the sequence lengths of drug and protein tokens, respectively. Each encoder layer consists of multi-head self-attention, feed-forward networks, residual connections, and LayerNorm. This design enables the model to capture contextual and long-range intra-modal dependencies within drug and protein sequences, thereby improving sequence-level representation quality.

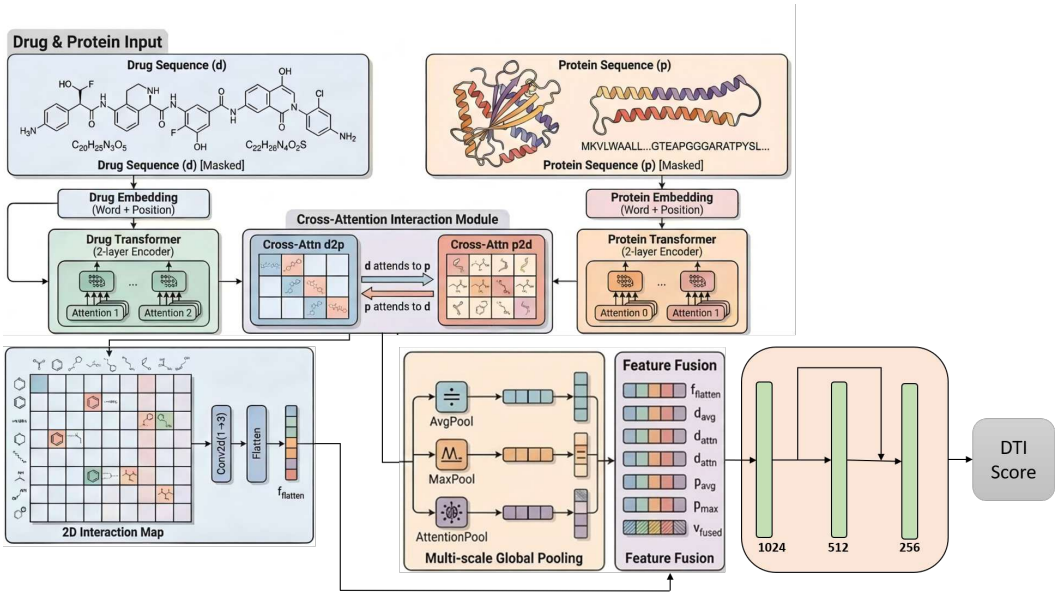


Figure 1: Overview of the proposed bidirectional cross-attention framework for drug–target interaction prediction. Drug and protein sequences are first encoded using independent Transformer encoders and aligned through bidirectional cross-attention to capture explicit cross-modal dependencies. A CNN-based interaction branch extracts higher-order local interaction patterns, while multi-scale global pooling aggregates sequence-level semantic information. The fused local–global representation is finally passed to a residual-enhanced classification head for interaction prediction.

3.3. Bidirectional Cross-Attention

While self-attention captures intra-sequence dependencies, it does not explicitly model interactions between drug and protein modalities. To address this limitation, we introduce a bidirectional cross-attention mechanism that explicitly aligns drug substructures with protein residues.

Specifically, we compute:

$$\hat{D} = \mathrm{MHA}(D', P', P'), \qquad \hat{P} = \mathrm{MHA}(P', D', D'),$$

where MHA denotes multi-head attention.In this formulation, drug tokens attend to protein representations and vice versa, enabling the model to explicitly capture cross-modal dependencies and biologically meaningful residue–substructure interactions that are difficult to model using implicit element-wise correlation mechanisms alone.

Residual connections and LayerNorm are applied to stabilise training:

$$D'' = \mathrm{LayerNorm}(D' + \hat{D}), \qquad P'' = \mathrm{LayerNorm}(P' + \hat{P}).$$

This design reformulates DTI prediction as a cross-modal alignment problem, allowing the model to move beyond local correlation-based interaction modelling towards explicit modelling of higher-order cross-modal dependencies and semantic alignment between drug substructures and protein residues.

3.4. Interaction Map Construction

Based on the cross-attention-enhanced representations D'' and P'', we construct an interaction map to model fine-grained relationships between drug substructures and protein residues.

We first compute an element-wise interaction tensor via broadcasting:

$$I = D'' \odot P'',$$

where $I \in \mathbb{R}^{B \times L_d \times L_p \times E}$. This operation captures pairwise interactions across all token combinations. To obtain a compact 2D interaction representation, we aggregate along the embedding dimension:

$$M \in \mathbb{R}^{B \times L_d \times L_p}.$$

The resulting interaction map M encodes fine-grained local interaction patterns between drug and protein tokens, enriched with explicit cross-modal dependency information obtained through bidirectional attention. A lightweight 2D CNN is then applied to M to extract local interaction features:

$$f_{\mathrm{conv}} = \mathrm{CNN}(M).$$

This branch focuses on capturing higher-order local interaction patterns among drug substructures and protein residues, similar to MolTrans, but benefits from the enhanced cross-modal representations introduced by bidirectional attention.

3.5. Hybrid Local–Global Representation

While convolutional processing effectively captures local interaction patterns, it lacks the ability to model global sequence-level semantics. To address this, we introduce a hybrid representation framework that combines local interaction features with global contextual information.

Global Feature Extraction. We apply three complementary pooling operations—average pooling, max pooling, and attention pooling—to both drug and protein representations:

$$d_{\text{avg}}, d_{\max}, d_{\text{attn}}, \quad p_{\text{avg}}, p_{\max}, p_{\text{attn}}.$$

The attention pooling is defined as:

$$\alpha_i = \frac{\exp(w^\top h_i)}{\sum_j \exp(w^\top h_j)}, \qquad h_{\text{attn}} = \sum_i \alpha_i h_i,$$

where h_i denotes token features and w is a learnable parameter. These pooling operations provide complementary semantic views of the interaction space: average pooling captures holistic contextual information, max pooling emphasises highly activated discriminative patterns, and attention pooling adaptively focuses on biologically informative interaction regions.

Global Feature Fusion. All pooled features are fused via concatenation to form a unified global representation:

$$G = [d_{\text{avg}}, d_{\max}, d_{\text{attn}}, p_{\text{avg}}, p_{\max}, p_{\text{attn}}].$$

Local–Global Fusion. The final representation is obtained by combining global semantic features with local interaction features extracted from the CNN:

$$z = [f_{\text{conv}}, G].$$

This hybrid representation enables the model to jointly capture fine-grained local interaction patterns and holistic sequence-level semantic dependencies, thereby improving robustness, interpretability, and generalisation, particularly under small, imbalanced, and scarce-data settings.

3.6. Enhanced Classification Head

The prediction head is a three-layer MLP with hidden sizes $1024 \to 512 \to 256 \to 1$. GELU activation is used between layers, LayerNorm replaces BatchNorm and dropout is applied for regularisation. Residual connections between hidden layers improve gradient flow and mitigate optimisation difficulties. The final output logit is passed to `BCEWithLogitsLoss` during training.

3.7. Training Stabilisation

Training stability is a critical factor in DTI prediction, particularly for small and imbalanced datasets such as DAVIS. To improve convergence behaviour and reduce optimisation variance, we adopt a set of stabilised training strategies.

Loss Function. We employ `BCEWithLogitsLoss` as the training objective, which combines a sigmoid activation with binary cross-entropy in a numerically stable formulation. Compared to a separate Sigmoid + BCELoss pipeline, this approach mitigates numerical instability and improves gradient behaviour.

Optimisation. The model is optimised using AdamW, which decouples weight decay from gradient updates and provides better generalisation compared to standard Adam. A cosine learning-rate scheduler with warm-up is applied to ensure smooth convergence and avoid abrupt learning-rate changes during early training.

Regularisation and Efficiency. To further stabilise optimisation and reduce overfitting, dropout is applied within the enhanced classification head during training. In addition, automatic mixed-precision training is used to improve computational efficiency and reduce memory consumption, while also helping to stabilise numerical computations.

Dataset-Specific Adjustments. Hyperparameters such as learning rate, batch size, and training epochs are adapted for each dataset. In particular, for smaller datasets, we employ slightly lower learning rates and longer warm-up phases to prevent overfitting and ensure stable optimisation.

4. Experiments

4.1. Datasets

We evaluate the proposed framework under realistic drug discovery settings using three widely adopted benchmark datasets for drug–target interaction prediction: BindingDB, BIOSNAP, and DAVIS Liu, Lin, Wen, Jorissen and Gilson (2007); Zitnik, Sosic and Leskovec (2018); Davis, Hunt, Herrgard, Ciceri, Wodicka, Pallares, Hocker, Treiber and Zarrinkar (2011). These datasets vary in scale, data distribution, and class imbalance, enabling a comprehensive assessment of model performance.

BindingDB is a large-scale biochemical dataset containing experimentally measured binding affinities between drug compounds and protein targets. BIOSNAP is a biomedical interaction dataset derived from multiple sources, including DrugBank and STITCH, and is commonly used for binary interaction prediction. DAVIS is a kinase inhibitor dataset with relatively small size and high imbalance, making it particularly challenging for robust model training.

Following prior work Huang et al. (2021), continuous binding affinity values in DAVIS are binarised using a threshold ($K_d < 30$) to obtain positive interactions. Negative samples are generated by randomly pairing drugs and proteins that are not known to interact.

Table 1 summarises the key statistics of the datasets, including the number of drugs, proteins, and interaction pairs.

Table 1
Dataset Comparison Summary

Metric	BindingDB	BIOSNAP	DAVIS
Number of Drugs	7,165	4,510	68
Number of Proteins	1,254	2,181	379
Number of Interactions	32,601	27,483	11,103
Positive Interactions	9,166	13,836	1,506
Negative Interactions	23,435	13,647	9,597
Positive Ratio (%)	28.12%	50.34%	13.56%
Negative Ratio (%)	71.88%	49.66%	86.44%

4.2. Data Splitting and Evaluation Metrics

For each dataset, we adopt an 80/10/10 split into training, validation, and test sets under a warm-start setting, consistent with the evaluation protocol of MolTrans. To ensure robustness, experiments are conducted across multiple random seeds, and average performance is reported.

We evaluate model performance using AUROC, AUPRC, and F1-score. AUROC measures the overall ranking capability, while AUPRC is particularly informative for imbalanced datasets by emphasising precision–recall trade-offs. F1-score provides a balanced measure of classification accuracy.

In addition, we report the epoch corresponding to the best validation performance and the overall training time to analyse convergence behaviour and computational efficiency.

4.3. Implementation Details

The proposed framework is implemented in PyTorch with GPU acceleration and automatic mixed-precision training. The embedding dimension is set to 128, with two Transformer encoder layers and eight attention heads. The intermediate feature dimension is 512, while the maximum sequence lengths are fixed to 50 for drug sequences and 545 for protein sequences. The drug and protein vocabularies contain 2,586 and 16,693 tokens, respectively.

The model is trained using the AdamW optimiser with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-5}. A warm-up cosine learning-rate scheduler is applied over 60 epochs with 3 warm-up epochs. The batch size is set to 16. Dropout is applied with a rate of 0.1 for both attention and hidden layers to improve regularisation.

For the enhanced classification head, the hidden dimensions are configured as $1024 \rightarrow 512 \rightarrow 256 \rightarrow 1$, using GELU activation, LayerNorm, and residual connections between hidden layers.

To improve stability on smaller datasets such as DAVIS, we employ a slightly lower learning rate, increased dropout, and longer warm-up scheduling.

4.4. Baselines

We compare the proposed model with the original MolTrans framework, which serves as the primary baseline due to its strong performance and similar token-based representation design. In addition, we include representative deep learning approaches for DTI prediction, including DeepDTA, DeepConv-DTI, and GraphDTA, to provide a broader comparison across different modelling paradigms.

To ensure fair comparison, we follow the same data preprocessing and evaluation protocol as MolTrans wherever applicable. For methods where re-implementation is not feasible, we report results from the literature.

4.5. Ablation Study

To evaluate the contribution of each component, we conduct ablation experiments by enabling cross-attention, multi-scale global pooling, and the enhanced classification head individually and in combination.

This analysis allows us to isolate the effect of each module and understand their complementary roles. In particular, cross-attention aims to improve cross-modal alignment, global pooling enhances semantic representation, and the residual classification head increases discriminative capacity.

5. Results

5.1. Overall Performance

Table 2 presents the quantitative comparison between the proposed framework and baseline models across the BindingDB, BIOSNAP, and DAVIS datasets.

The proposed framework consistently outperforms the original MolTrans baseline across all datasets and evaluation metrics, demonstrating the effectiveness of explicit cross-modal interaction modelling and hybrid local–global representation learning. On BindingDB, the model achieves an AUROC improvement from 0.9073 to 0.9202 and an AUPRC increase from 0.5899 to 0.6344. Similar gains are observed on BIOSNAP, where AUROC improves from 0.8884 to 0.8927 and AUPRC from 0.8968 to 0.9023.

Notably, the most significant improvements are obtained on the DAVIS dataset, where AUROC increases from 0.9022 to 0.9262 and AUPRC from 0.3721 to 0.4898. Given the small size and strong class imbalance of DAVIS, these results indicate that the proposed framework is particularly effective in handling challenging data distributions and learning robust interaction patterns under scarce-data conditions.

Overall, the consistent performance gains across datasets of varying scale demonstrate the robustness, interpretability, and generalisation capability of the proposed framework.

All results are averaged over multiple runs, demonstrating consistent performance improvements across different random initialisations.

5.2. Training Dynamics and Stability

Figure 2 illustrates the training and validation curves for the proposed model and its variants.

On BindingDB, the introduction of cross-attention leads to faster convergence, reaching optimal performance in significantly fewer epochs compared to the baseline. On BIOSNAP, improvements are more moderate but consistently stable, with smoother validation curves indicating better generalisation.

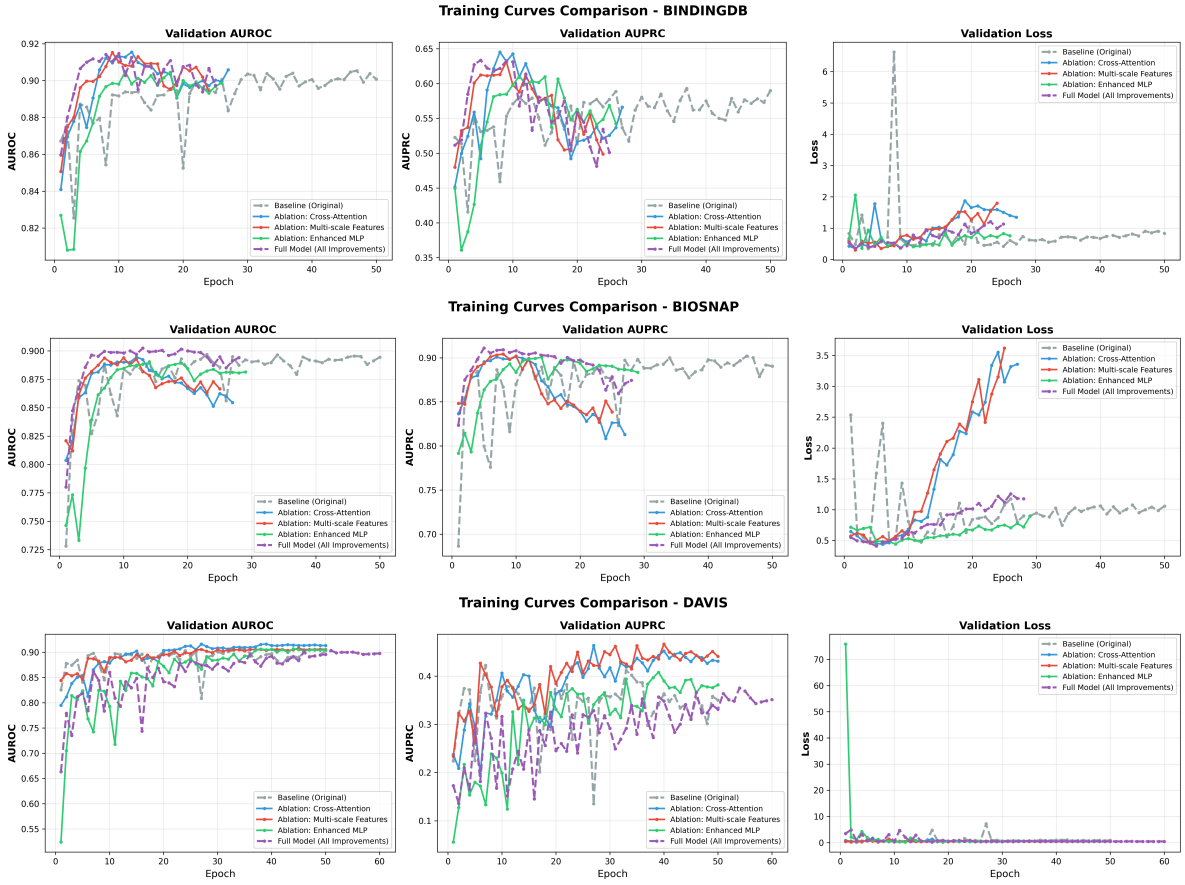


Figure 2: Training and validation curves for the original MolTrans model and the proposed variants on BindingDB, BIOSNAP and DAVIS. On BindingDB, cross-attention notably accelerates convergence and improves final performance. On BIOSNAP, gains from individual components are modest but the combined model generalises best. On DAVIS, the full model yields the most stable optimisation behaviour.

The most notable effect is observed on the DAVIS dataset, where the proposed framework substantially reduces fluctuations in validation performance. Compared to the baseline, the training curves exhibit lower variance and fewer oscillations, suggesting improved optimisation stability.

These results demonstrate that the combination of explicit cross-modal interaction modelling, hybrid local–global representation learning, and stabilised optimisation contributes not only to improved predictive accuracy, but also to more reliable, interpretable, and stable training behaviour under challenging and scarce-data settings.

5.3. Ablation Study

To analyse the contribution of each component, we evaluate three ablation variants: (i) cross-attention only, (ii) multi-scale global pooling only, and (iii) enhanced classification head only.

The results reveal that each component contributes in a complementary manner. The cross-attention module consistently improves AUROC across all datasets, indicating its effectiveness in modelling cross-modal dependencies. In particular, it provides a +1.01% AUROC gain on BindingDB and +0.65% on DAVIS.

The multi-scale global pooling module yields substantial improvements in AUPRC, with gains of +5.98% on BindingDB and +4.88% on DAVIS. This suggests that global semantic aggregation enhances the model's ability to correctly identify positive interactions, especially under class imbalance.

The enhanced classification head alone provides limited improvements and can slightly reduce AUROC in some cases, indicating that increased model capacity without improved representation may not be sufficient.

Importantly, when combined, these components produce the best overall performance, demonstrating that explicit cross-modal interaction modelling, global semantic aggregation, and stabilised optimisation contribute in a complementary and mutually reinforcing manner. Specifically, cross-attention improves semantic alignment between drug and protein representations, global pooling enhances representation robustness, and the enhanced classifier refines decision boundaries.

Table 2 summarises validation and test performance for all variants, together with best epoch and training time.

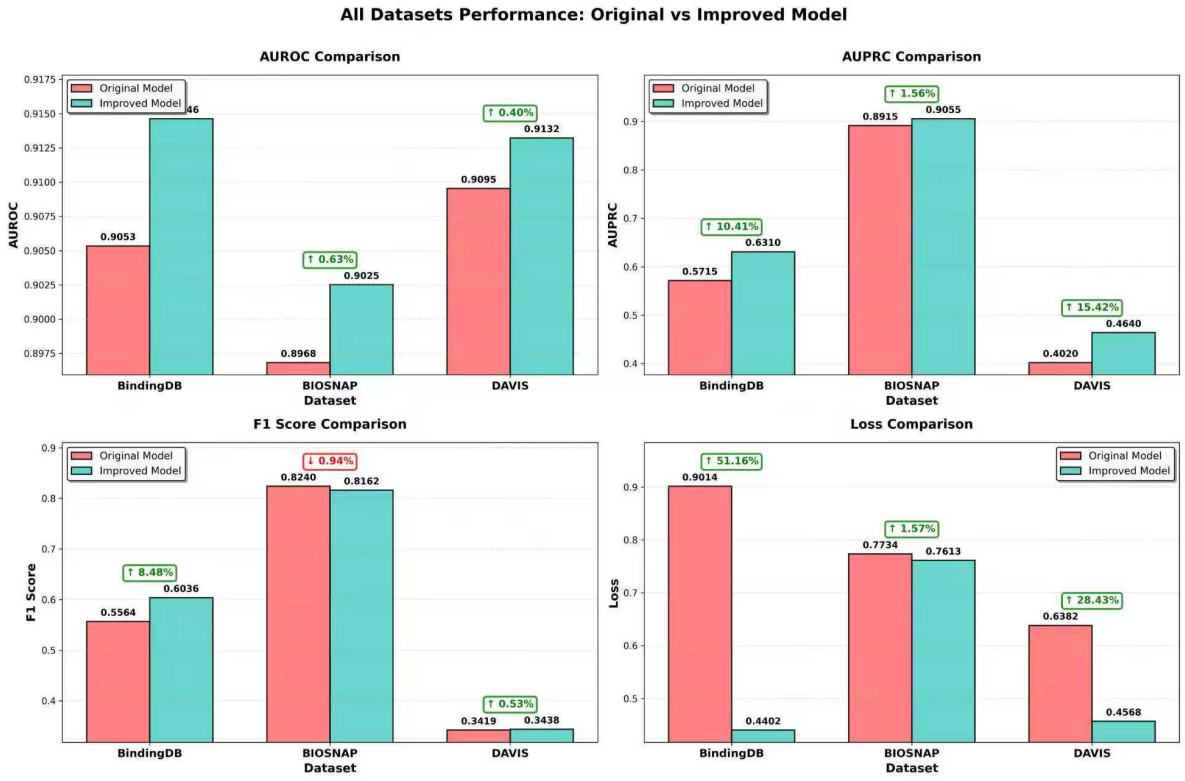


Figure 3: Comparison of AUROC, AUPRC, F1-score, and loss between the original MolTrans model and the proposed framework across the three benchmark datasets.

Table 2
Performance comparison of different improvements

Model	AUROC	AUPRC	F1
BindingDB			
Original	0.9073	0.5899	0.5724
Cross-Attention	0.9125	0.6312	0.5768
Multi-scale Feat.	0.9143	0.6319	0.5536
Enhanced MLP	0.8989	0.6003	0.5686
All Improvements	**0.9202**	**0.6344**	**0.6106**
BIOSNAP			
Original	0.8884	0.8968	0.8157
Cross-Attention	0.8870	0.8892	0.8145
Multi-scale Feat.	0.8847	0.8876	0.8200
Enhanced MLP	0.8768	0.8852	0.7973
All Improvements	**0.8927**	**0.9023**	**0.8183**
DAVIS			
Original	0.9022	0.3721	0.3172
Cross-Attention	0.9127	0.4374	0.3284
Multi-scale Feat.	0.9166	0.4373	0.3474
Enhanced MLP	0.8968	0.3301	0.3200
All Improvements	**0.9262**	**0.4898**	**0.3771**

On all datasets the combined model achieves the highest AUROC and AUPRC, confirming that the three modules are complementary when used together. The largest relative improvements are observed on DAVIS, where the data are scarce and noisy; here, the AUPRC increase of more than 11 percentage points suggests that the improved architecture and training strategy allow the model to better capture rare positive interactions.

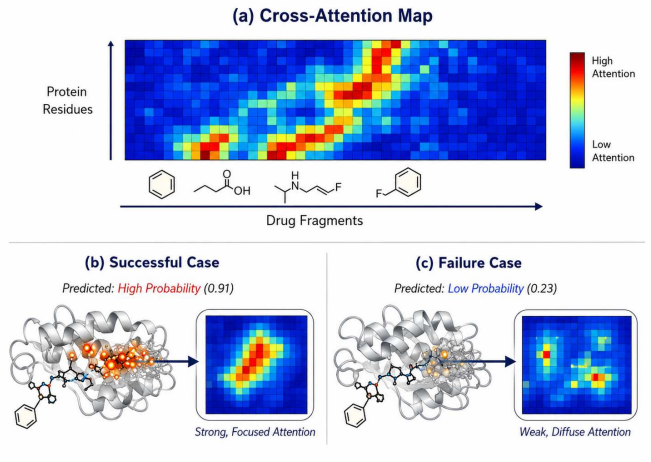


Figure 4: Visualisation of cross-attention between drug fragments and protein residues. (a) Cross-attention map showing interaction patterns learned by the model. (b) Example of a successful prediction with concentrated attention regions. (c) Example of a failure case exhibiting weaker and more diffuse attention patterns.

Table 3
Performance under unseen drug and protein settings on the BIOSNAP dataset (ROC-AUC averaged over five runs)

Setting	DeepDTA	DeepConv-DTI	GNN-CPI	Ours
Unseen drugs	0.861 ± 0.006	0.872 ± 0.005	0.879 ± 0.004	**0.893 ± 0.006**
Unseen proteins	0.781 ± 0.021	0.795 ± 0.018	0.801 ± 0.017	**0.818 ± 0.021**

Table 4
Performance under different missing-data ratios (ROC-AUC averaged over five runs)

Missing (%)	DeepDTA	DeepConv-DTI	GNN-CPI	Ours
70	0.895 ± 0.004	0.901 ± 0.003	0.906 ± 0.003	**0.916 ± 0.004**
80	0.872 ± 0.006	0.881 ± 0.004	0.887 ± 0.004	**0.902 ± 0.005**
90	0.821 ± 0.009	0.835 ± 0.006	0.842 ± 0.005	**0.865 ± 0.006**
95	0.742 ± 0.012	0.765 ± 0.009	0.772 ± 0.008	**0.801 ± 0.007**

5.4. Attention Visualisation and Case Analysis

To better understand the interaction modelling behaviour of the proposed cross-attention mechanism, we visualise representative attention maps between drug fragments and protein residues, as shown in Figure 4.

The attention maps reveal that the proposed framework learns focused and structured interaction patterns rather than diffuse correlations. In successful prediction cases, the model exhibits concentrated attention over specific drug fragments and protein residue regions, suggesting that the cross-attention mechanism effectively captures biologically meaningful interactions.

In contrast, failure cases tend to exhibit weaker and more dispersed attention patterns, suggesting reduced confidence and less discriminative cross-modal interaction modelling. These cases may reflect ambiguous biochemical interaction signals or insufficient contextual representation under challenging data conditions. This observation is consistent with the lower predicted interaction probabilities produced by the model.

These visualisations provide qualitative evidence that the proposed bidirectional cross-attention mechanism enhances cross-modal alignment between drug and protein representations and contributes to improved predictive performance. In addition, the attention maps improve interpretability by providing a tractable view of the learned interaction patterns between drug fragments and protein residues. Concentrated attention regions in successful cases suggest that the model learns biologically meaningful cross-modal dependencies.

5.5. Comparison with Existing Methods

We further compare the proposed model with representative DTI prediction methods, including DeepDTA, DeepConv-DTI, and GNN-CPI.

In the evaluated generalisation settings, the proposed framework achieves superior performance compared with representative DTI prediction methods.

The results in Table 3 further demonstrate strong generalisation capability under realistic drug discovery settings involving unseen drugs and unseen protein targets.

In addition, the proposed model exhibits robust performance under high levels of missing data, as shown in Table 4. Even when up to 95% of interaction data is removed, the model maintains a significant performance advantage over baseline approaches.

Compared to MolTrans, the improvements can be attributed to three key factors: (i) explicit cross-modal interaction modelling via bidirectional attention, (ii) integration of local and global feature representations, and (iii) improved optimisation and classification design.

These results confirm that the proposed framework provides a more expressive, interpretable, and robust solution for DTI prediction by explicitly modelling cross-modal dependencies and integrating complementary local and global interaction representations.

6. Discussion

The experimental results provide several important insights into the design of effective Transformer-based models for drug–target interaction prediction.

First, the introduction of bidirectional cross-attention plays a central role in improving model performance. By explicitly modelling interactions between drug substructures and protein residues, the proposed framework moves beyond the implicit, element-wise correlation mechanisms used in prior methods. This enables the model to capture higher-order and long-range dependencies, which are particularly important for complex biochemical interactions. The consistent AUROC improvements across all datasets, especially on the DAVIS dataset, highlight the effectiveness of this explicit cross-modal alignment strategy.

Second, the hybrid local–global representation proves essential for robust performance. While the CNN-based interaction map effectively captures fine-grained local patterns, it alone is insufficient to represent global sequence-level context. The integration of multi-scale pooling introduces complementary global information: average pooling encodes overall context, max pooling emphasises salient features, and attention pooling adaptively focuses on informative regions. The observed improvements in AUPRC, particularly under class imbalance, suggest that global semantic aggregation enhances the model's ability to correctly identify positive interactions.

Third, the results demonstrate that architectural improvements must be coupled with stable optimisation strategies. The use of BCEWithLogitsLoss, AdamW, and cosine learning-rate scheduling leads to smoother convergence and reduced variance across training runs. This effect is most evident on the DAVIS dataset, where the baseline model exhibits unstable behaviour, while the proposed framework achieves more consistent and reliable performance.

The results further suggest that the proposed framework is more robust under scarce-data settings, where stable representation learning and explicit interaction modelling become increasingly important.

The ablation study further reveals that the proposed components contribute in a complementary manner. Cross-attention primarily enhances cross-modal alignment and discriminative capability, while global pooling improves representation robustness and recall. The enhanced classification head, although less impactful in isolation, plays a supporting role by refining decision boundaries when combined with improved feature representations. These findings highlight the importance of balancing representation learning and model capacity.

Despite these improvements, several limitations remain. The current framework primarily relies on sequence-based representations and does not explicitly incorporate structural priors such as molecular graphs, binding pockets, or 3D protein conformations, which may contain additional biochemically relevant interaction information. In addition, the cross-attention mechanism introduces additional computational overhead, which may limit scalability for very large datasets.

Future work will explore integrating graph neural networks, structure-aware Transformers, and 3D geometric representations to better capture molecular topology, binding-site information, and protein structural dependencies. Incorporating 3D structural information, for example from AlphaFold predictions, may further enhance interaction modelling. In addition, techniques such as knowledge distillation and model compression could improve deployment efficiency, while active learning strategies may reduce the reliance on large labelled datasets.

Overall, the proposed framework demonstrates that explicitly modelling cross-modal interactions, combined with hybrid representation learning and stabilised optimisation, provides a more expressive, interpretable, and robust approach to DTI prediction across diverse and realistic drug discovery scenarios.

7. Conclusion

This work proposed a bidirectional cross-attention and hybrid local–global representation framework for Transformer-based DTI prediction that builds on the sub-structural foundations of MolTrans. By redesigning the interaction and decoding stages and stabilising training, the method achieves consistent improvements across three benchmark datasets. The ablation study shows that cross-attention, multi-scale global pooling and residual-enhanced classification contribute complementary benefits, and that the combination of all components yields the best trade-off between predictive accuracy, robustness, and generalisation capability.Overall, the proposed framework demonstrates that explicit cross-modal interaction modelling, combined with hybrid local–global representation learning and stabilised optimisation, provides an effective and interpretable solution for robust DTI prediction under realistic drug discovery settings.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All datasets used in this study are publicly available. BIOSNAP was obtained from the Stanford Network Analysis Project (SNAP) Biomedical Dataset Collection: https://snap.stanford.edu/biodata. DAVIS was obtained from the benchmark dataset introduced by Davis et al. (2011). and is publicly available at: http://staff.cs.utu.fi/~aatapa/data/DrugTarget/. BindingDB was obtained from the BindingDB database: https://www.bindingdb.org.

The source code, configuration files, and data processing scripts used in this study are available in the BCAG-DTI repository: https://github.com/Falmi/BCAG-DTI

Competing interests

The authors declare that they have no competing interests.

Funding

No specific funding was received for this work.

CRediT authorship contribution statement

Hao Pang: Formal analysis, Methodology,Software, Visualization, Writing – original draft. **Fiseha Berhanu Tesema:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – review and editing . **Tianxiang Cui:** Writing – review and editing. **Yuan Cheng:** Writing – review and editing.

References

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. Elsevier. volume 4, pp. 217–241.

Broach, J.R., Thorner, J., et al., 1996. High-throughput screening for drug discovery. Nature 384, 14–16.

Cao, D.S., Xu, Q.S., Liang, Y.Z., 2013. propy: a tool to generate various modes of chou's pseaac. Bioinformatics 29, 960–962.

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, 1046–1051.

Dubchak, I., Muchnik, I., Holbrook, S.R., Kim, S.H., 1995. Prediction of protein folding class using global description of amino acid sequence. Proceedings of the National Academy of Sciences 92, 8700–8704.

Gao, M., Zhang, D., Chen, Y., Zhang, Y., Wang, Z., Wang, X., Li, S., Guo, Y., Webb, G.I., Nguyen, A.T., et al., 2024. Graphormerdti: a graph transformer-based approach for drug-target interaction prediction. Computers in Biology and Medicine 173, 108339.

Gong, Y., Luo, H., Zhang, J., 2017. Natural language inference over interaction space. arXiv preprint arXiv:1709.04348 .

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, 24.

Huang, K., Xiao, C., Glass, L.M., Sun, J., 2021. Moltrans: molecular interaction transformer for drug–target interaction prediction. Bioinformatics 37, 830–836.

Hughes, J.P., Rees, S., Kalindjian, S.B., Philpott, K.L., 2011. Principles of early drug discovery. British journal of pharmacology 162, 1239–1249.

Jia, J., Zhu, F., Ma, X., Cao, Z.W., Li, Y.X., Chen, Y.Z., 2009. Mechanisms of drug combinations: interaction and network perspectives. Nature reviews Drug discovery 8, 111–128.

Jin, Z., Wu, T., Chen, T., Pan, D., Wang, X., Xie, J., Quan, L., Lyu, Q., 2023. Capla: improved prediction of protein–ligand binding affinity by a deep learning approach based on a cross-attention mechanism. Bioinformatics 39, btad049.

Lee, I., Keum, J., Nam, H., 2019. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS computational biology 15, e1007129.

Li, F., Zhang, Z., Guan, J., Zhou, S., 2022. Effective drug–target interaction prediction with mutual interaction neural network. Bioinformatics 38, 3582–3589.

Liu, T., Lin, Y., Wen, X., Jorissen, R.N., Gilson, M.K., 2007. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic acids research 35, D198–D201.

Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J.K., Ceulemans, H., Clevert, D.A., Hochreiter, S., 2018. Large-scale comparison of machine learning methods for drug target prediction on chembl. Chemical science 9, 5441–5451.

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, 1140–1147.

Öztürk, H., Özgür, A., Ozkirimli, E., 2018. Deepdta: deep drug–target binding affinity prediction. Bioinformatics 34, i821–i829.

Öztürk, H., Ozkirimli, E., Özgür, A., 2019. Widedta: Prediction of drug-target binding affinity. arxiv 2019. arXiv preprint arXiv:1902.04166 .

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, 325–337.

Rogers, D., Hahn, M., 2010. Extended-connectivity fingerprints. Journal of chemical information and modeling 50, 742–754.

Schenone, M., Dančík, V., Wagner, B.K., Clemons, P.A., 2013. Target identification and mechanism of action in chemical biology and drug discovery. Nature chemical biology 9, 232–240.

Tsubaki, M., Tomii, K., Sese, J., 2019. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics 35, 309–318.

Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Wegner, J.K., Ceulemans, H., Hochreiter, S., 2014. Deep learning as an opportunity in virtual screening, in: Proceedings of the deep learning workshop at NIPS, MIT Press Cambridge, MA, United States. pp. 1–9.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., Lu, H., 2017. Deep-learning-based drug–target interaction prediction. Journal of proteome research 16, 1401–1409.

Zheng, X., Ding, H., Mamitsuka, H., Zhu, S., 2013. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1025–1033.

Zitnik, M., Sosic, R., Leskovec, J., 2018. Biosnap datasets: Stanford biomedical network dataset collection. Note: http://snap. stanford. edu/biodata Cited by 5.