Attentive Multi-Channel Molecular Representation in Drug–Target Affinity Prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Accurate drug–target affinity (DTA) prediction is fundamental to computational drug discovery. Drug–target binding affinity is influenced by multiple factors, including structural conformations, functional groups, and molecular flexibility. However, existing graph neural network (GNN)-based approaches often fail to explicitly capture these fine-grained features, leading to suboptimal performance and limited interpretability. To address these issues, we propose a framework that integrates structure-aware protein embeddings with muti-channel molecular representations across global, scaffold, and local levels. The key innovation lies in a Cross-Channel Attention (CCA) mechanism, which dynamically aligns protein features with molecular channels and assigns adaptive weights, thereby selectively emphasizing binding-relevant information while reducing redundancy. Experiments on the Davis dataset demonstrate that our model consistently outperforms strong baseline methods. Beyond performance improvements, the cross-channel attention mechanism also enhances interpretability by highlighting structural and chemical determinants of binding. Overall, this work establishes cross-channel attention as an effective and interpretable paradigm for advancing DTA prediction.

1 Introduction

Drug–target affinity (DTA) is a crucial task in drug design and development, playing a vital role in drug discovery, candidate screening, and dosage optimization. Traditional experimental approaches such as high-throughput screening (HTS) and surface plasmon resonance (SPR) provide accurate measurements but suffer from high cost, low throughput, and long processing time, making them insufficient for modern large-scale drug development [8]. In recent years, deep learning has emerged as a powerful tool, showing strong ability to model large-scale biomedical data and significantly accelerating progress in DTA prediction.

In DTA modeling, the representation of compounds and proteins is a crucial factor affecting model performance. Generally, three mainstream strategies are adopted: one-dimensional sequence representation (1D), two-dimensional structural graphs (2D), and three-dimensional spatial structures (3D) (Figure 1). At the 1D level, compounds are usually represented by SMILES strings or molecular fingerprints, while proteins are expressed as amino acid sequences encoded through descriptors such as k-mer or PSSM [3]. At the 2D level, compounds are modeled as molecular graphs and proteins as contact maps, capturing topological and spatial dependencies[7]. At the 3D level, with advances such as AlphaFold2, structural conformations and drug–target complexes can be modeled as point clouds, voxel grids, or atom-level graphs to better capture realistic molecular interactions [16].

Previous studies have proposed a variety of deep learning-based methods for DTA prediction, which can be broadly divided into three categories: sequence-based models, structure-based models, and hybrid-based models [22]. Sequence-driven models usually rely on SMILES strings for compounds

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

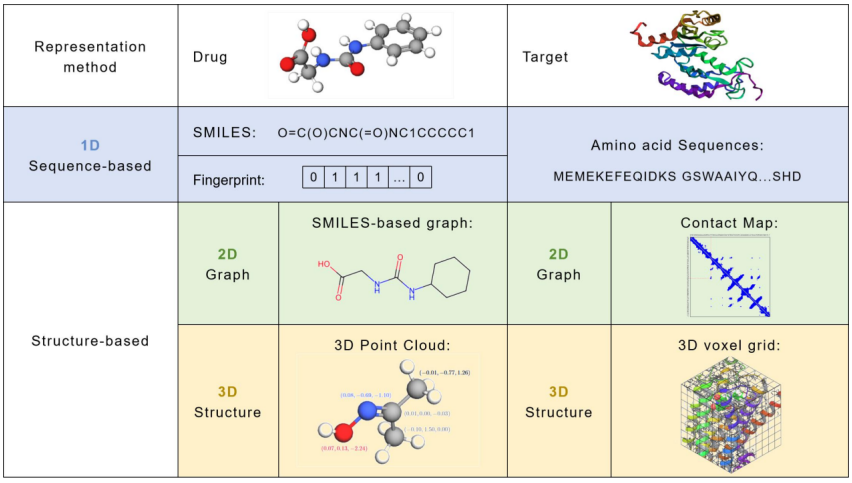

Figure 1: Illustration of compound and protein representations in DTA modeling.

and amino acid sequences for proteins, and employ CNNs, RNNs, or Transformer-based architectures for feature extraction. Representative examples include DeepDTA [24], DeepCDA [1], and AttentionDTA [23], which achieve strong baseline performance but fail to capture spatial structural information. Structure-aware models leverage molecular graphs, protein contact maps, or 3D complex structures to incorporate spatial dependencies. Notable approaches include GSAML-DTA [10], HGRL-DTA [4], and MSGNN-DTA [18]. Although these models offer better interpretability and physical consistency, they are heavily dependent on high-quality structural inputs, which are not always available in real-world scenarios [22]. Moreover, structure-aware models typically impose higher computational costs due to the complexity of processing 3D spatial information and large-scale graph structures. Hybrid models combine the strengths of both paradigms by jointly leveraging sequence and structural information. Representative methods include GraphDTA [14], MGraphDTA [20], ColdDTA [6], and MutualDTA [21], obatining the balance between performance and efficiency.

Despite recent advances, existing DTA prediction methods still face several challenges as follows:

- **Insufficient feature representation:** Drug–target affinity depends on multiple factors such as conformations, functional groups, and molecular flexibility. However, many GNN-based models fail to capture these fine-grained structural signals, which limits prediction accuracy and generalization.
- **Limitations in feature fusion:** Multi-channel molecular features are often combined via simple concatenation, static weighting, or globally learnable weights. These strategies cannot adaptively assign channel importance for each specific drug–protein pair, and may encounter unstable weight convergence or even yield inferior performance compared to single-channel representations.
- **Lack of interpretability:** Many existing models operate as black-box predictors, offering little biological insight into the determinants of molecular binding, which restricts their utility in practical drug discovery scenarios.

Therefore, to address these challenges, we propose a novel framework for drug–target affinity prediction that integrates GNN-based multi-channel molecular representation with structure-aware protein embedding (Figure 2). On the molecular side, three complementary channels—global topology, scaffold backbone, and local functional groups—are extracted using a pretrained GNN, providing a hierarchical description of chemical features. On the protein side, we employ the pretrained ESM-2 language model to generate structure-aware embeddings that capture contextual and conformational information. To enable effective interaction between drug and protein features, we introduce a Cross-Channel Attention mechanism that adaptively balances the contributions of different molecular channels, thereby enhancing both predictive accuracy and interpretability.

In summary, our main contributions are as follows:

- We propose a drug–target affinity prediction model that incorporates **multi-channel molecular representations**, enabling a more comprehensive capture of global, scaffold, and local chemical features.
- We design a **Cross-Channel Attention (CCA) mechanism** that adaptively weights molecular channels against protein embeddings, effectively leveraging complementary information while mitigating redundancy.
- We conduct extensive experiments, including ablation studies and case analyses, which demonstrate that the proposed model not only outperforms mainstream baselines but also yields biologically meaningful interpretability under different binding modes.

2 Methodology

In this section, we introduce the proposed framework for DTA prediction. We first present the overall architecture of our proposed framework. We then describe the molecular and protein representation modules in detail, followed by the cross-channel attention mechanism, which enables adaptive integration of multi-channel molecular features.

2.1 Model Architecture

The proposed framework consists of three main components: a molecular feature extraction module, a protein feature extraction module, and a cross-channel attention fusion module. Concretely, a pretrained protein language model is employed to encode protein sequences into structure-aware embeddings. On the molecular side, we adopt the multi-channel representation scheme introduced in MolMCL [17], where graph neural networks generate three complementary embeddings capturing global, scaffold, and local features. Building on these representations, our key contribution is the cross-channel attention module, where protein embeddings serve as guiding signals to dynamically adjust the importance of different molecular channels. The fused representation is then passed through a multilayer perceptron to predict drug–target affinity. This design enables effective utilization of structural information and enhances interpretability, as illustrated in Figure 2.

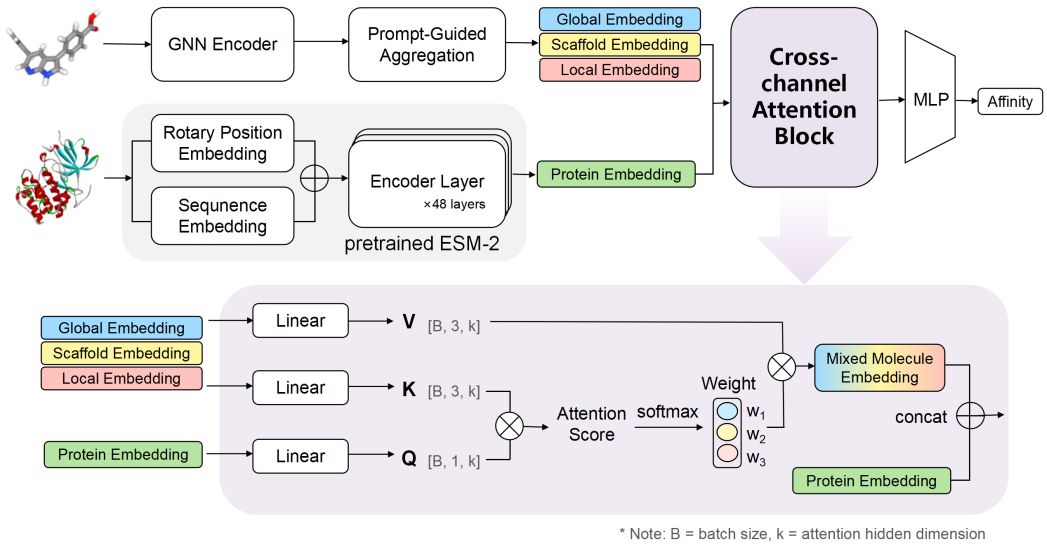

Figure 2: Overall framework. Molecular graphs are encoded by a single GNN and, via prompt-guided aggregation, pooled into three channel embeddings (global, scaffold, local) following MolMCL [17]. Protein sequences are encoded by pretrained ESM-2 [11]. Our key contribution is the **Cross-Channel Attention Block**, which uses the protein embedding to compute softmax weights over the three channels and fuse them for affinity prediction.

2.2 Molecular Representation

We adopt the MoLMCL framework [17] to construct hierarchical molecular embeddings from three complementary perspectives: global, scaffold, and local. Concretely, molecular SMILES are first converted into graphs and encoded by a GNN. Within each channel, a Prompt-Guided Aggregation module introduces a learnable prompt token to guide attention-based pooling over node features, producing a channel-specific graph embedding. Through this process, we obtain three embeddings that correspond to global topology, scaffold backbone, and local functional groups. Each embedding is channel with a dedicated self-supervised objective to capture multi-level structural information, as detailed in the following subsections.

Global Channel: Molecule Contrastive Distancing (MCD). The goal is to learn embeddings that preserve the overall molecular topology. Training for this channel is conducted through triplet contrastive learning, where each batch constructs {anchor, positive, negative} triplets. The anchor a is the original molecule, the positive p is generated by applying subgraph masking to a following the strategy proposed by MolCLR [19], and the negative n is randomly sampled from other molecules in the batch. To improve sensitivity to structural similarity, an adaptive-margin triplet loss is employed:

$$\mathcal{L}_{MCD} = \max\left(0, \alpha_{\text{MCD}} + d(a,p) - d(a,n)\right), \tag{1}$$

where $d(\cdot,\cdot)$ denotes the embedding distance between two molecules. The adaptive margin α_{MCD} is defined as:

$$\alpha_{\text{MCD}} = \alpha_{\text{offset}} \cdot \left(1 - \text{sim}_{\text{Tanimoto}}\left(\text{FP}_{\text{mol}}^{(a)},\ \text{FP}_{\text{mol}}^{(n)}\right)\right), \tag{2}$$

where α_{offset} is a hyper-parameter controlling the margin scale, $\text{sim}_{\text{Tanimoto}}(\cdot,\cdot)$ is the Tanimoto similarity [2], and $\text{FP}_{\text{mol}}^{(a)}$ and $\text{FP}_{\text{mol}}^{(n)}$ are molecular fingerprints of the anchor and negative molecules, respectively.

Sacffold Channel: Scaffold Contrastive Distancing (SCD). This channel focuses on backbone-level invariance. The anchor a is the original molecule, the positive p is generated by applying scaffold-invariant perturbations using CReM (an open-source framework for chemically reasonable mutations), and the negative n is sampled from the batch. A similar adaptive-margin triplet loss is applied:

$$\mathcal{L}_{SCD} = \max\left(0, \alpha_{\text{SCD}} + d(a,p) - d(a,n)\right), \tag{3}$$

$$\alpha_{\text{SCD}} = \alpha_{\text{offset}} \cdot \left(1 - \text{sim}_{\text{Tanimoto}}\left(\text{FP}_{\text{scaff}}^{(a)},\ \text{FP}_{\text{scaff}}^{(n)}\right)\right), \tag{4}$$

where $\text{FP}_{\text{scaff}}^{(a)}$ and $\text{FP}_{\text{scaff}}^{(n)}$ represent scaffold-level fingerprints of the anchor and negative molecules.

Local Channel: Context Prediction (CP). The local channel enhances context understanding by jointly modeling subgraph-level structures and functional group descriptors through a multi-task learning scheme with two objectives:

- *Masked subgraph prediction:* A random atom and its 1-hop neighbors are masked, and the model predicts the missing features as a multi-label classification task optimized with cross-entropy loss:

$$\mathcal{L}_{mask} = -\sum_{i=1}^{C} y_i \log p_i, \tag{5}$$

  where C is the number of classes, p_i is the predicted probability for class i, and $y_i \in \{0,1\}$ is the ground-truth label.

- *Functional group prediction:* Each molecule is represented by an 86-dimensional normalized functional group descriptor, and the task is formulated as regression with Smooth L1 loss:

$$\mathcal{L}_{FG} = \frac{1}{d}\sum_{i=1}^{d} \begin{cases} 0.5(\hat{y}_i - y_i)^2, & \text{if } |\hat{y}_i - y_i| < 1, \\ |\hat{y}_i - y_i| - 0.5, & \text{otherwise}, \end{cases} \tag{6}$$

  where y_i and $\hat{y}_i$ denote the ground-truth and predicted descriptor values of the i-th functional group.

The total context prediction loss is defined as:

$$\mathcal{L}_{CP} = \mathcal{L}_{mask} + \mathcal{L}_{FG}. \tag{7}$$

Prompt-Guided Aggregation. In each channel, a learnable prompt token guides a multi-head attention pooling over atom-level features: the prompt serves as the query, while node embeddings act as keys and values. This prompt-guided pooling enforces channel-wise structural focus and replaces uniform pooling with selective, structure-aware aggregation. Prompt-Guided Aggregation operates during channel encoding and is optimized jointly with the corresponding channel objective in the pretraining stage, yielding a unified yet channel-aware molecular representation for downstream prediction.

2.3 Protein Representation

We employ the ESM-2 model [11], a state-of-the-art protein language model, to encode protein sequences into structure-aware embeddings. ESM-2 was pretrained on 98M UniRef50 protein sequences, covering diverse evolutionary and functional contexts, which enables it to capture rich contextual and structural information from primary amino acid sequences.

The model adopts a masked language modeling objective: around 15% of residues are randomly masked or replaced, and the model is trained to recover the original amino acids. This task drives ESM-2 to learn rich contextual dependencies across residues, capturing long-range sequence relationships. In addition, embeddings from upper layers of ESM-2 have been shown to align closely with protein contact map, demonstrating that the model implicitly encodes structural features despite being trained without explicit 3D supervision.

In our work, we utilize the final-layer embeddings (layer 36) from ESM-2, where residue-level representations are averaged to form a global protein embedding. These embeddings are particularly suitable for our multi-channel framework, as they simultaneously encode complementary information at different levels: global sequence context, structural backbone organization, and local binding-pocket features. Such hierarchical protein information aligns naturally with the molecule-side multi-channel design, enabling effective cross-channel attention and more accurate affinity prediction.

2.3.1 Cross-Channel Attention Mechanism(CCA)

In the preliminary design, the fusion of the three molecular channels relied on fixed weights or static weighted summation. Such approaches lack sensitivity to protein features and cannot dynamically adapt to different protein–molecule pairs. To address these limitations, we introduce a Cross-Channel Attention mechanism that enables more flexible and context-aware integration of representations (Figure 2).

Specifically, the protein embedding is first projected into a query vector $Q \in \mathbb{R}^{B\times 1\times k}$, while the three molecular channel embeddings are projected into key and value vectors $K \in \mathbb{R}^{B\times 3\times k}$ and $V \in \mathbb{R}^{B\times 3\times k}$, respectively, where B is the batch size and k represents the hidden dimension of the attention space. The attention scores are then computed using the scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

where d_k represents the dimensionality of the key vectors, used to scale the dot product and prevent the attention scores from becoming excessively large. The softmax operation produces dynamic weights $\{w_1, w_2, w_3\}$ for the three channels, which are used to compute a weighted sum of molecular embeddings and form the Mixed Molecule Embedding. This fused embedding is subsequently concatenated with the protein embedding and passed into an MLP predictor for affinity regression.

The CCA mechanism offers several advantages compared to other fusion methods:

- **Adaptive and stable fusion**: Attention weights are computed dynamically for each protein–molecule pair, enabling the model to adjust the fusion strategy in a context-specific manner. Unlike approaches that assign a fixed learnable scalar weight to each channel, which may lead to unstable or biased convergence, this dynamic weighting ensures more reliable and consistent training behavior.

- **Context-guided alignment**: Protein embeddings guide the computation of fusion weights, enabling the model to capture dependencies between protein context and molecular structural levels.

3 Experiments

In this section, we evaluate the effectiveness of our proposed model on the Davis dataset and compare it with representative baselines. We further conduct ablation studies and case analyses to validate the contribution of each module and the interpretability of our framework.

3.1 Experimental Setup

Objective. The primary goal of our experiments is to assess the predictive performance of the proposed model on DTA prediction tasks and benchmark it against state-of-the-art baselines.

Evaluation Metrics. We adopt three widely used regression metrics. Let y_i denote the ground-truth affinity value, $\hat{y}_i$ the predicted value, and N the number of samples.

(1) Mean Squared Error (MSE). MSE evaluates the absolute prediction accuracy by measuring the average squared difference between predictions and true values:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{9}$$

(2) Concordance Index (CI). CI measures the consistency of ranking between predicted and true affinities. For all comparable pairs (i, j) where $y_i > y_j$, CI is defined as:

$$\text{CI} = \frac{1}{Z}\sum_{y_i > y_j} h(\hat{y}_i - \hat{y}_j), \tag{10}$$

where Z is the total number of such pairs and

$$h(x) = \begin{cases} 1 & x > 0, \\ 0.5 & x = 0, \\ 0 & x < 0. \end{cases}$$

(3) Coefficient of Determination (R^2). R^2 assesses the proportion of variance in the ground-truth values explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{11}$$

where $\bar{y}$ is the mean of the ground-truth values.

3.2 Dataset

We conduct experiments on the Davis dataset, which is a widely adopted benchmark for DTA prediction [5]. The dataset consists of 68 kinase inhibitors and 442 protein kinases, resulting in 30,056 interaction pairs with experimentally measured binding affinities (K_d values). These values are provided in log-transformed form, making them suitable for regression-based modeling. The Davis dataset is characterized by its large coverage, containing more than 30k compound–protein pairs, as well as its continuous affinity labels that enable quantitative evaluation.

Split strategy. Following common practice, the dataset is split into 80% training, 10% validation, and 10% testing, controlled by a fixed random seed to ensure reproducibility.

Table 1: Performance comparison on the Davis dataset. The best results are in bold.

Model	MSE ↓	CI ↑	R^2 ↑
KronRLS	0.379	0.871	0.407
DeepDTA	0.261	0.878	0.630
GraphDTA	0.229	0.893	0.670
ColdDTA	0.250	0.884	0.652
NTMFF-DTA	0.237	**0.896**	0.684
Ours	**0.221**	0.884	**0.732**

3.3 Baselines

We compare our proposed model with six representative baseline methods, covering traditional kernel-based approaches, early deep learning models, and the latest multi-scale frameworks:

- **KronRLS** [15]: a classical kernelized regression method that models drug–target pairs via the Kronecker product of drug kernels and protein kernels. Despite its simplicity, KronRLS remains a strong traditional baseline due to its efficiency and effectiveness on small-scale datasets.
- **DeepDTA** [24]: the first end-to-end deep learning framework for DTA prediction. It encodes drug SMILES and protein amino acid sequences separately with convolutional neural networks (CNNs) and learns their joint interactions through fully connected layers.
- **GraphDTA** [14]: extends DeepDTA by representing drugs as molecular graphs instead of SMILES strings, thereby capturing richer structural information with graph neural networks (GNNs). Proteins are still modeled using CNNs over their sequences.
- **ColdDTA** [6]: designed to address the cold-start problem where unseen drugs or proteins appear in the test set. It introduces a cold-start aware loss function that forces the model to generalize better across novel compounds and targets.
- **NTMFF-DTA** [12]: a recent state-of-the-art model that incorporates neural temporal memory modules and multi-scale feature fusion. This design enables the model to capture both local and global dependencies in drug and protein features.

3.4 Implementation Details

Our model is implemented in `PyTorch 2.1.2` with CUDA 11.8 and trained on a single NVIDIA RTX 3090 GPU (24GB). We adopt the Adam optimizer with a cosine learning rate scheduler and warm-up strategy, and train for up to 100 epochs with early stopping to prevent overfitting.

3.5 Results

We evaluate our model against five representative baselines on the Davis dataset, and the results are summarized in Table 1. Traditional kernel methods such as KronRLS achieve reasonable performance but are limited by shallow feature representations. DeepDTA improves substantially by introducing CNN encoders for SMILES and protein sequences, while GraphDTA further reduces error by leveraging molecular graph structures. ColdDTA enhances generalization in cold-start scenarios with specialized loss design, and more recent approaches such as NTMFF-DTA demonstrate the benefits of incorporating attention mechanisms and multi-scale modeling. Compared with these baselines, our model achieves the lowest MSE (0.221) and the highest R^2 (0.732), indicating superior regression accuracy and stronger explanatory power. Although NTMFF-DTA attains a slightly higher CI, our approach strikes a better balance between error minimization and predictive stability. These results highlight the effectiveness of combining hierarchical molecular representations, protein language embeddings, and cross-channel attention for prediction.

3.6 Ablation Studies

To investigate the contribution of each component in our proposed model to DTA prediction, we conducted a series of ablation experiments. Specifically, we examined the effect of different molecular

Table 2: Ablation study results on the Davis dataset.

Model Variant	MSE	CI	R^2
Single-channel (CP)	0.238	0.881	0.711
Single-channel (SCD)	0.284	0.868	0.655
Single-channel (MCD)	0.277	0.869	0.664
Mean weight	0.251	0.880	0.695
Full model	**0.214**	**0.886**	**0.740**

channel settings by removing or simplifying the cross-channel fusion mechanism, while keeping the other parts of the architecture unchanged. The experimental settings follow the same data splitting strategy as in the main experiments, and all results were obtained with random seed fixed at the same value for reproducibility.

Regarding the molecular channel representation, we evaluated three reduced configurations: (i) **Single-channel (MCD/SCD/CP)**, where only one type of molecular representation was preserved and the other channels were discarded; (ii) **Mean weight**, where all three molecular channels were averaged with equal weights, discarding the dynamic weighting mechanism; and (iii) **Full model**, where all channels were fused via cross-channel attention (results provided in Table 2).

As shown in Table 2, the full model achieves the best performance across all metrics, demonstrating the effectiveness of integrating complementary information from multiple channels with adaptive attention. In contrast, the single-channel variants exhibited substantial performance degradation, with the CP-only variant performing relatively better than MCD-only and SCD-only settings. This result indicates that the CP channel preserves more fine-grained chemical property features that are directly related to ligand–protein binding, thereby outperforming other single-channel variants. Mean pooling achieved only moderate results, performing worse than the CP-only variant, since although it integrates features from all three channels, it fails to adaptively emphasize the most informative chemical property features and thus cannot fully leverage their complementarity. Overall, these results verify that both multi-channel representation and cross-channel attention are critical to the predictive power of the model.

3.6.1 Case Study and Model Interpretability

To further validate the effectiveness of multi-channel fusion and cross-channel attention in practical tasks, we selected two representative protein–ligand complexes with distinct conformational states of the Abl tyrosine kinase. The chosen complexes are the Abl tyrosine kinase structures with PDB IDs **3UE4** and **4XEY**, which exhibit substantially different ligand binding patterns. The visualization results are shown in Figure 3, where we analyze how distinct binding patterns influence the attention weights allocation across molecular channels, demonstrating the interpretability of the model.

Figure 3(a–c) illustrates the complex structure of 3UE4, originally reported by Levinson and Boxer [9], which depicts the Abl kinase domain bound to the ATP-competitive inhibitor bosutinib. Bosutinib inserts deeply into the conserved ATP-binding pocket, where its rigid backbone forms extensive hydrophobic contacts with residues such as Leu248A, Phe317A, and Met318A, ensuring a stable geometric fit within the cavity. The interaction pattern is dominated by scaffold-driven hydrophobic embedding, while only a limited number of hydrogen bonds (e.g., with Met318A) contribute modestly to the stabilization of the complex. This binding mode is faithfully captured by our model (Figure 3c), which assigns high weight to the *scaffold* channel and minimal weight to the *local* channel. This observation highlights the ability of the model to recognize scaffold-driven binding mechanisms and to adapt its attention toward structural features underpinning ligand affinity.

In contrast, Figure 3(d–f) presents the 4XEY structure, originally reported by Lorenz et al. [13], which reveals the Abl SH2-kinase domain in complex with dasatinib. Dasatinib binds in a mode distinct from bosutinib, where local functional groups rather than the rigid scaffold play the dominant role. The ligand inserts parallel to the protein surface, forming multiple hydrogen bonds with residues such as Met337A, Thr334A, and Met309A, while also engaging in hydrophobic contacts that provide additional stabilization. The interaction pattern is thus driven primarily by fine-grained chemical group recognition, with hydrogen bonds serving as key determinants of binding specificity. This

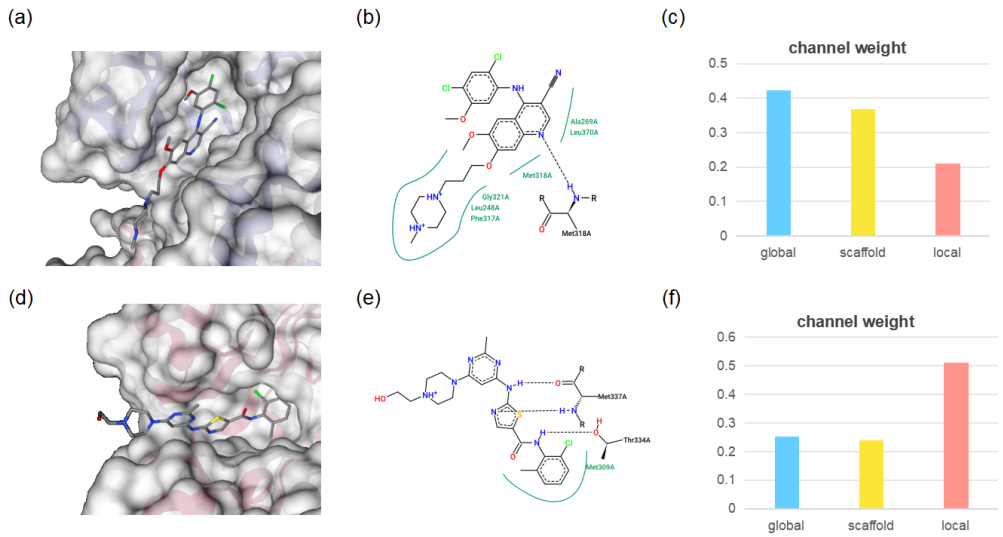

Figure 3: Case study of model interpretability on 3UE4 and 4XEY. (a–c) 3UE4 complex: (a) 3D binding pose, (b) 2D interaction diagram, (c) channel weight distribution. (d–f) 4XEY complex: (d) 3D binding pose, (e) 2D interaction diagram, (f) channel weight distribution.

binding mode is faithfully captured by our model (Figure 3f), which assigns dominant weight to the *local* channel and only minor contributions to the *scaffold* and *global* channels.

Together, these two cases demonstrate the capacity of our model to dynamically adjust attention across channels depending on the binding mode. In 3UE4, binding is scaffold-dominated and attention is focused on the *scaffold* channel, whereas in 4XEY, functional group interactions dominate and attention shifts to the *local* channel. This adaptive allocation not only confirms the rationality of the multi-channel design but also illustrates its structural interpretability. By dynamically allocating attention across multi-channel molecular representations, the framework provides mechanistic insights and enhances trustworthiness, addressing the common criticism of deep learning methods for their limited interpretability.

4 Conclusion

In this work, we presented an attentive multi-channel framework for DTA prediction that integrates structure-aware protein embeddings with muti-channel molecular representations. By employing complementary channels for global topology, scaffold backbone, and local functional groups, and fusing them through a cross-channel attention mechanism, our model effectively captures multi-level structural and chemical information. Extensive experiments on the Davis dataset demonstrated that our method outperforms representative baselines, while ablation studies and case analyses confirmed the importance of both multi-channel representation and adaptive attention for accuracy and interpretability. Beyond predictive performance, the framework provides mechanistic insights into distinct binding modes, enhancing the reliability of computational DTA modeling. These findings suggest that attentive multi-channel learning offers a promising direction for advancing interpretable and generalizable approaches in computational drug discovery.

References

[1] Karim Abbasi, Parvin Razzaghi, Antti Poso, Massoud Amanlou, Jahan B Ghasemi, and Ali Masoudi-Nejad. Deepcda: deep cross-domain compound–protein affinity prediction through lstm and convolutional neural networks. *Bioinformatics*, 36(17):4633–4642, 2020.

[2] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):20, 2015.

[3] Kuo-Chen Chou. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Bioinformatics*, 43(3):246–255, 2001.

[4] Zhaoyang Chu, Feng Huang, Haitao Fu, Yuan Quan, Xionghui Zhou, Shichao Liu, and Wen Zhang. Hierarchical graph representation learning for the prediction of drug-target binding affinity. *Information Sciences*, 613:507–523, 2022.

[5] Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

[6] Kejie Fang, Yiming Zhang, Shiyu Du, and Jian He. Colddta: Utilizing data augmentation and attention-based feature fusion for drug-target binding affinity prediction. *Computers in Biology and Medicine*, 164:107372, 2023.

[7] David T Jones and Shaun M Kandathil. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics*, 34(19):3308–3315, 2018.

[8] Visvaldas Kairys, Lina Baranauskiene, Migle Kazlauskiene, Daumantas Matulis, and Egidijus Kazlauskas. Binding affinity in drug design: experimental and computational techniques. *Expert Opinion on Drug Discovery*, 14(8):755–768, 2019.

[9] Nicholas M. Levinson and Steven G. Boxer. Structural and spectroscopic analysis of the kinase inhibitor bosutinib and an isomer of bosutinib binding to the abl tyrosine kinase domain. *PLOS ONE*, 7(4):1–10, 2012.

[10] Jiaqi Liao, Haoyang Chen, Lesong Wei, and Leyi Wei. Gsaml-dta: An interpretable drug-target binding affinity prediction model based on graph neural networks with self-attention mechanism and mutual information. *Computers in Biology and Medicine*, 150:106145, 2022.

[11] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*, 2022.

[12] Yuandong Liu, Youzhi Liu, Haoqin Yang, Longbo Zhang, Kai Che, and Linlin Xing. Ntmff-dta: Prediction of drug-target affinity based on network topology and multi-feature fusion. *Interdisciplinary Sciences: Computational Life Sciences*, pages 1–13, 2025.

[13] Sonja Lorenz, Patricia Deng, Oliver Hantschel, Giulio Superti-Furga, and John Kuriyan. Crystal structure of an sh2–kinase construct of c-abl and effect of the sh2 domain on kinase activity. *Biochemical Journal*, 468(2):283–291, 2015.

[14] Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2020.

[15] Tapio Pahikkala, Antti Airola, Sami Pietilä, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang, and Tero Aittokallio. Toward more realistic drug–target interaction predictions. *Briefings in Bioinformatics*, 16(2):325–337, 2014.

[16] Albert Ros-Lucas, Nieves Martinez-Peinado, Jaume Bastida, Joaquim Gascón, and Julio Alonso-Padilla. The use of alphafold for in silico exploration of drug targets in the parasite trypanosoma cruzi. *Frontiers in Cellular and Infection Microbiology*, Volume 12 - 2022:944748, 2022.

[17] Yue Wan, Jialu Wu, Tingjun Hou, Chang-Yu Hsieh, and Xiaowei Jia. Multi-channel learning for integrating structural hierarchies into context-dependent molecular representation. *Nature Communications*, 16(1):413, 2025.

[18] Shudong Wang, Xuanmo Song, Yuanyuan Zhang, Kuijie Zhang, Yingye Liu, Chuanru Ren, and Shanchen Pang. Msgnn-dta: Multi-scale topological feature fusion based on graph neural networks for drug–target binding affinity prediction. *International Journal of Molecular Sciences*, 24(9), 2023.

[19] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

[20] Ziduo Yang, Weihe Zhong, Lu Zhao, and Calvin Yu-Chian Chen. Mgraphdta: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical science*, 13(3):816–833, 2022.

[21] Yongna Yuan, Siming Chen, Rizhen Hu, and Xin Wang. Mutualdta: An interpretable drug–target affinity prediction model leveraging pretrained models and mutual attention. *Journal of Chemical Information and Modeling*, 65(3):1211–1227, 2025.

[22] Xin Zeng, Shu-Juan Li, Shuang-Qing Lv, Meng-Liang Wen, and Yi Li. A comprehensive review of the recent advances on predicting drug-target affinity based on deep learning. *Frontiers in Pharmacology*, Volume 15 - 2024, 2024.

[23] Qichang Zhao, Guihua Duan, Mengyun Yang, Zhongjian Cheng, Yaohang Li, and Jianxin Wang. Attentiondta: Drug–target binding affinity prediction by sequence-based deep learning with attention mechanism. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):852–863, 2023.

[24] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.