# MomeDTA: Improving Generalizability in Drug-Target Affinity Prediction by Mixture of Multi-view Experts

**Qingyu Yang[1, e], Yue Teng[1, e], Jie Yang[2, e], Tao Zhang[1], Jiale Yu[1], Jie Zheng[1, 3, *]**
[1]School of Information Science and Technology, ShanghaiTech University
[2]School of Biomedical Engineering, ShanghaiTech University
[3]Shanghai Engineering Research Center of Intelligent Vision and Imaging
[*]Correspondence: zhengjie@shanghaitech.edu.cn
[e]Equal contribution

## Abstract

Accurate prediction of drug-target affinity (DTA) plays an important role in targeted therapy and drug discovery by enabling efficient virtual screening. However, many computational methods for DTA prediction show poor accuracy on novel drug-target pairs and inconsistent performance across different datasets because they only focus on a single view. In this work, we present a novel method called MomeDTA, which utilizes mixture of multi-view expert models to adaptively fuse 1D, 2D and 3D information. Our method achieves state-of-the-art (SOTA) performance in multiple scenarios on different datasets, demonstrating that we successfully improve model generalizability in DTA prediction. All of our code is available at https://github.com/Yangqy-16/MomeDTA.

## 1 Introduction

In conventional drug discovery, the identification of ligands that selectively bind to a target protein relies on labor-intensive manual design and experimental validation. However, a significant proportion of drug candidates will be eliminated in the experimental stages due to insufficient binding affinity, which leads to enormous expenditures of time and financial resources. Therefore, virtual screening serves as a key step in novel drug development by leveraging deep learning methods for rapid, large-scale prediction of drug-target binding affinities.

Recent deep learning approaches for predicting drug-target binding affinity can be categorized by how they represent drugs and targets on different views. Sequence-based (1D) models use amino acid sequences for proteins and SMILES for drugs. Early models such as DeepDTA [1] apply convolutional neural networks (CNNs) [2] to one-hot encoded inputs. More recent work employs pretrained language models (PLMs) for DTA prediction. For example, LLMDTA [3] uses ESM-2 [4] for proteins and Mol2Vec [5] for drugs. Graph-based (2D) approaches represent drugs and proteins as graphs to capture topological information. GraphDTA [6] extends DeepDTA by applying graph neural networks (GNNs) [7, 8] to molecular graphs. MMSG-DTA [9] integrates the GINE network for molecular graph representation and further constructs protein contact maps to facilitate feature learning through GateGAT. Structure-based (3D) methods [10] leverage 3D conformational data, including the use of structural PLMs such as Uni-Mol [11].

However, methods based on a single view have potential flaws related to generalizability. First, while practical drug discovery scenarios require computational methods to have good performance on novel samples, the prediction accuracy of these methods on unseen drugs or proteins remains poor, probably

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Machine Learning and Structural Biology.

due to limited input features of the single view. Second, most existing studies use the same modal features for all datasets, yet they overlook data bias: due to variation in their data collection scenarios, different datasets inherently exhibit differences in their distributions and traits. According to our experiments, the traits of different datasets lead to distinct dependencies on specific data views (see Appendix C.2) and unstable performance of single-view approaches (see Section 3.3).

Multimodal learning, which trains models by combining inputs from multiple modalities/views to capture complementary information and learn a more comprehensive representation [12], has found many applications in biomolecule representation learning and property prediction [13–16]. In this work, we present **MomeDTA**, a novel deep learning method that utilizes **M**ixture **o**f **m**ulti-view **e**xpert models to predict **D**rug-**T**arget **A**ffinity. Notably, we are the first to combine 1D, 2D, and 3D inputs of both drugs and proteins, and achieve SOTA performance in warm and cold-start scenarios on multiple datasets, effectively improving model generalizability in DTA prediction. In addition, we validate the necessity and effectiveness of integrating multi-view inputs, and demonstrate that the affinity prediction for different drug-target pairs have distinct dependencies on different views.

## 2 Methods

### 2.1 Model Architecture

Typically, a DTA prediction model follows a framework comprising three components: encoders for extracting representations from the protein and drug inputs, a fusion module to combine these representations, and a prediction head for estimating the binding affinity [17]. Concerned with the generalizability and inspired by previous MoE-like architectures [18] in DTA prediction, we design three tracks that process different views of the input drug and protein, with each track following the above framework. Specifically, in the 1D and 3D tracks, we first utilize pretrained language models (PLMs) with rich prior biological knowledge to encode drugs and proteins, which enhances the generalizability and robustness of our model. Then, a modified mutual attention (MAN) [10] and a fusion block are applied to fuse and refine the information of the two molecules. In the 2D track, we adopt multiscale GNNs for drugs and proteins to capture both global and local information. Finally, a gating architecture named ViewMix integrates the three processed embeddings to get the affinity score. The whole pipeline is shown in Figure 1 and the details can be found in Appendix A.

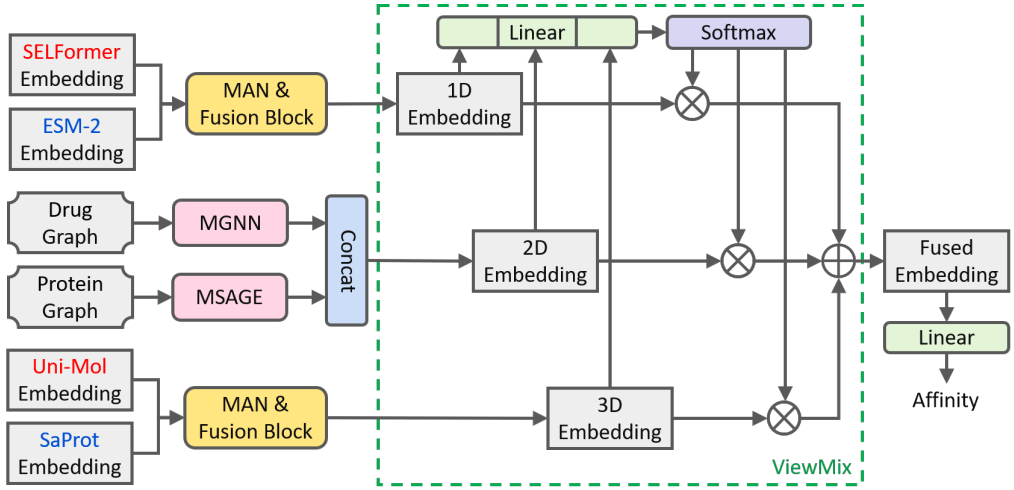


Figure 1: The pipeline of MomeDTA. Pretrained drug models are shown in red text and pretrained protein models are shown in blue text.

### 2.2 View-aware Mixture Block

Mixture of Experts (MoE) is a common approach to take advantage of the ability of multiple models [19] and has found some applications in DTI prediction [18]. In this work, we design a view-aware mixture block (ViewMix). It imitates and refines the gating architecture, where gating refers to the

technique that, given an input $x$, the proportion of each model's prediction logit to the final prediction logit is determined by a mapping of the combination of $x$'s embeddings of different models. Here, we take a step further: the final prediction logit depends on the weighted sum of the embeddings of each model, where the weights are predicted through a simple MLP from these embeddings. This allows the fusion to be aware of different features or positions in a sample instead of giving each model a fixed weight. The detailed process is shown in Algorithm 6 in Appendix A.

## 3 Experiments and Results

### 3.1 Data

We collect the labeled data of Davis [20] and Kiba [21] from DeepDTA and Metz [22] from the original paper, and no protein structures are provided in these three datasets. First, we search AlphaFold DB [23] for matched structures of all the protein sequences. For those without matched structures, we run AlphaFold2 [24] to predict their structures. Then, we encode drugs and proteins based on all the encoders in our model to obtain their embeddings. The encoding process of the 4 PLMs can be found in their corresponding codes [4, 11, 25, 26].

Next, we split the dataset into training, validation, and test set under 2 scenarios: warm and cold-start. For the warm scenario, we randomly split all labeled drug-target pairs into 64:16:20. For the cold-start scenario, we first cluster drugs and proteins and divide them into different folds. Then, different folds are distributed to the training, validation, and test sets, which ensures molecular dissimilarity between these three sets. More details can be seen in Appendix C.1.

### 3.2 Experimental Setup

We collect baselines covering different methods based on sequence, graph, and multi-modal inputs. Their details are listed in Table 4 in Appendix B. We adopt MSE, CI, $R^2$, Pearson correlation coefficient, Spearman correlation coefficient, and MAE as evaluation metrics, all of which are defined and widely used in previous methods [1, 3, 6].

### 3.3 Results

Table 1: Results of the warm scenario. The method ranking first is in bold and second is underlined. "PCC" is short for Pearson correlation coefficient and "SCC" is short for Spearman correlation coefficient. Likewise for other tables.

| Dataset | Method | MSE↓ | CI↑ | $R^2$↑ | PCC↑ | SCC↑ | MAE↓ |
|---|---|---|---|---|---|---|---|
| Davis | DeepDTA | 0.4870 | 0.8229 | 0.2497 | 0.6334 | 0.5742 | 0.5314 |
| | GraphDTA | 0.2632 | 0.8603 | 0.5944 | 0.7738 | 0.6298 | <u>0.2820</u> |
| | LLMDTA | <u>0.2083</u> | <u>0.8777</u> | <u>0.6790</u> | <u>0.8261</u> | <u>0.6548</u> | 0.2861 |
| | MMSG-DTA | 0.3797 | 0.8233 | 0.4150 | 0.6479 | 0.5752 | 0.3840 |
| | MomeDTA | **0.1988** | **0.8837** | **0.6937** | **0.8360** | **0.6644** | **0.2185** |
| Kiba | DeepDTA | 0.1934 | 0.8630 | 0.7173 | 0.8477 | 0.8430 | 0.2614 |
| | GraphDTA | <u>0.1567</u> | <u>0.8821</u> | <u>0.7709</u> | <u>0.8793</u> | <u>0.8729</u> | <u>0.2225</u> |
| | LLMDTA | 0.1642 | 0.8681 | 0.7599 | 0.8743 | 0.8594 | 0.2478 |
| | MMSG-DTA | 0.3008 | 0.8090 | 0.5604 | 0.7489 | 0.7557 | 0.3406 |
| | MomeDTA | **0.1553** | **0.8867** | **0.7730** | **0.8798** | **0.8737** | **0.2103** |
| Metz | DeepDTA | 0.1660 | 0.8611 | 0.6986 | 0.8427 | 0.8366 | 0.2400 |
| | GraphDTA | **0.1275** | <u>0.8872</u> | **0.7683** | **0.8771** | **0.8759** | <u>0.1989</u> |
| | LLMDTA | 0.1347 | 0.8663 | 0.7553 | 0.8709 | 0.8532 | 0.2270 |
| | MMSG-DTA | 0.2585 | 0.8068 | 0.5306 | 0.7285 | 0.7489 | 0.3219 |
| | MomeDTA | <u>0.1308</u> | **0.8899** | <u>0.7624</u> | <u>0.8742</u> | <u>0.8731</u> | **0.1943** |

The results in warm and cold-start scenarios on the Davis, Kiba, and Metz datasets are shown in Table 1 and Table 2, respectively. In general, MomeDTA achieves SOTA performance in almost all experiments. In the warm scenario, our method ranks first on all metrics on Davis and Kiba and is

comparable to the best baseline GraphDTA on Metz. In the cold-start scenario, our model ranks first on all metrics on Metz and performs best overall on Davis and Kiba. The performance of all methods drops significantly in the cold-start scenario, implying that prediction on novel samples is difficult, but our top performance on all the datasets and the rise of our performance rank from the warm to the cold-start scenario on Metz demonstrate that MomeDTA has better generalizability than baselines. The comparison of predicted versus true values of MomeDTA is shown in Appendix D.

Table 2: Results of the cold-start scenario.

| Dataset | Method | MSE↓ | CI↑ | $R^2$↑ | PCC↑ | SCC↑ | MAE↓ |
|---|---|---|---|---|---|---|---|
| Davis | DeepDTA | 0.4914 | 0.6351 | -0.0725 | 0.2771 | 0.2228 | 0.5758 |
| | GraphDTA | 0.4977 | 0.4973 | -0.0863 | 0.0226 | -0.0058 | 0.4596 |
| | LLMDTA | 0.4314 | **0.7093** | 0.0642 | 0.3618 | **0.3415** | 0.4384 |
| | MMSG-DTA | 0.4799 | 0.5820 | -0.0474 | 0.1396 | 0.1341 | 0.5362 |
| | MomeDTA | **0.4013** | 0.6884 | **0.1241** | **0.3735** | 0.3079 | **0.4137** |
| Kiba | DeepDTA | 0.8730 | 0.5620 | -0.2630 | 0.1942 | 0.1754 | 0.6865 |
| | GraphDTA | 0.6727 | 0.5783 | 0.0268 | 0.2078 | 0.2218 | 0.5936 |
| | LLMDTA | 0.8725 | **0.6136** | -0.2626 | 0.2174 | **0.3166** | 0.6048 |
| | MMSG-DTA | 0.7105 | 0.5715 | -0.0279 | 0.2115 | 0.2035 | 0.6041 |
| | MomeDTA | **0.6555** | 0.6014 | **0.0516** | **0.2865** | 0.2819 | **0.5903** |
| Metz | DeepDTA | 0.6933 | 0.5881 | -0.1704 | 0.1949 | 0.2427 | 0.5247 |
| | GraphDTA | 0.5707 | 0.5968 | 0.0365 | 0.2430 | 0.2650 | 0.5386 |
| | LLMDTA | 0.4482 | 0.6577 | 0.2439 | 0.4957 | 0.4257 | 0.4781 |
| | MMSG-DTA | 0.5383 | 0.6344 | 0.0912 | 0.3061 | 0.3673 | 0.5090 |
| | MomeDTA | **0.4357** | **0.6776** | **0.2644** | **0.5239** | **0.4731** | **0.4513** |

Note that the performance of these baseline methods varies between different datasets. In the warm scenario on Davis, some sequence-based methods show superior performance than graph-based methods (i.e. LLMDTA is better than GraphDTA). By contrast, in the warm scenario on Kiba and Metz, the graph-based method is superior to sequence-based methods (i.e. GraphDTA is the best). These results collectively reveal the presence of data bias, which refers to the situation that different datasets have distinct traits and thus tend to rely on different kinds of input features, possibly due to variations in their data collection process or data distribution. According to the above experiment results and our data analysis in Appendix C.2, sequence information can be more important to the data in Davis while drugs in Kiba and Metz have stronger correlations with topological information. Hence, it may be difficult for previous methods based on a single modality to work with datasets of all kinds. We alleviate the problem by combining information from 1D to 3D. These multiple views enable the model to be more comprehensive so that it can be used in a wider range, which is demonstrated in our experiments.

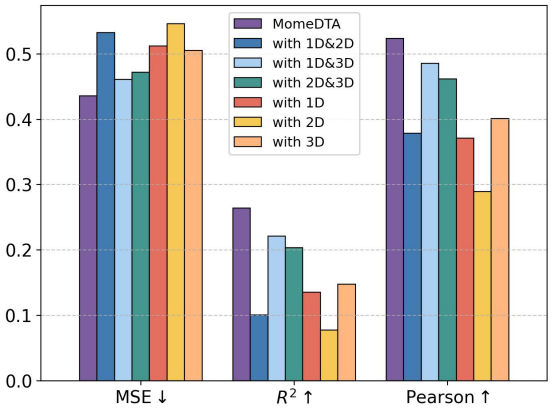


Figure 2: Ablation study in cold-start scenario on Metz.

### 3.4 Ablation Study

To validate the necessity of integrating the three views, we also perform an ablation study in the cold-start scenario on Metz, with results shown in Figure 2. It is clear that the performance of the entire MomeDTA exceeds that of any single view or combination of only 2 views. This result demonstrates the advantage of integrating information of multiple views within the model. When

working alone, the 2D track slightly lags behind because its optimal hyperparameters (e.g. learning rate, batch size, etc.) are different from those of MomeDTA, but we keep all hyperparameters the same when performing ablation study for fair comparison. Among the 2-view combinations, "1D&2D" performs the worst, which means 3D view makes the most contribution under this setting, followed by 1D and 2D.

### 3.5 Case Study

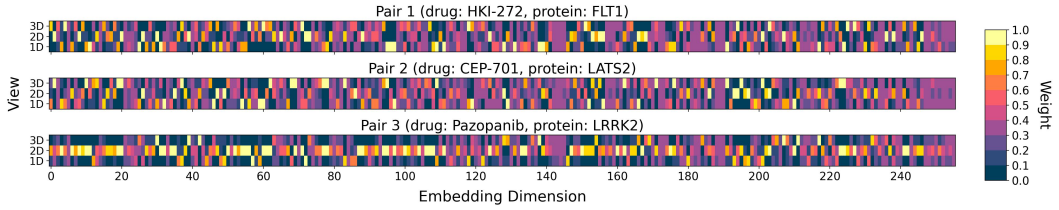


Figure 3: Weights $W$ in ViewMix for 3 different drug-target pairs in the test set of Davis.

To showcase our model is able to perceive the information of different views, we randomly select 3 drug-target pairs in Davis, perform model inference on them, and visualize the weights in ViewMix ($W$ in Algorithm 6 in Appendix A) in Figure 3, where lighter colors represent larger weights. According to the results, the weights of the 3 modalities all appear large somewhere, meaning that all the 3 views are useful for the prediction to some extent. More importantly, the numerical distribution of the weights varies between different views and between different drug-target pairs, which demonstrates that different drug-target pairs exhibit distinct dependencies on specific views, and our ViewMix is capable of capturing such differences. The reason is that ViewMix decides the contribution of each view to the prediction of binding affinity by controlling "dynamic" weights, where the weights are calculated according to the input features and thereby "dynamic" for different samples. This is different from "static" fusion techniques such as concatenation or weighted average where the same proportion of each modality is applied to all the samples.

## 4 Discussion

In conclusion, we have designed a new method called MomeDTA which exploits mixture of multi-view expert models to predict drug-target affinity. We achieve SOTA performance in warm and cold-start scenarios on datasets with different traits, effectively improving generalizability both on novel samples and against data bias in DTA prediction. The ablation study demonstrates the necessity of integrating multi-view inputs, and the case study shows that our model is aware of the distinct dependencies on specific views across different drug-target pairs. Moreover, our model can identify the binding region within the drug-target pair (see Appendix E).

The superiority of MomeDTA over its single tracks can be attributed to the distinct and complementary roles played by the different views in modeling interactions. In the 1D view, pretrained BERT-style [27] models are employed. For proteins, such models extract evolutionary information closely related to their biological function and binding propensity [4]. For drugs, semantic patterns in SMILES/SELFIES allow the model to capture functionally relevant atomics, even over long ranges, and reveal general binding tendencies [28]. The 2D view is well suited to represent relational structures such as chemical bonds and functional groups in drugs as well as residue interaction networks in proteins, allowing identification of the topological structures in both entities [29]. The 3D view provides essential spatial contexts including potential binding pockets in proteins [30] and fine-grained conformations of drugs [11]. These factors directly determine intermolecular affinity by defining shape complementarity and interaction feasibility. Integrating these three views enriches the feature space with non-redundant information, enabling the model to capture complex drug-target interaction mechanisms that no single view can fully represent.

Still, our model has limitations. Our prediction accuracy is not very high on some datasets or in some scenarios. Specifically, the 2D track possibly drags behind the model capability due to incompatible batch size. In addition, our training speed is slower than single-modality methods due to our multiple tracks (see Appendix F). These limitations are our potential places for improvement. The future of

DTA prediction probably lies in developing more efficient methods for modal fusion and continuous improvement on model generalizability and interpretability.

## References


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

[2] Stanisław Jastrzębski, Damian Leśniak, and Wojciech Marian Czarnecki. Learning to SMILE(S). *arXiv preprint arXiv:1602.06289*, 2016.

[3] Wuguo Tang, Qichang Zhao, and Jianxin Wang. LLMDTA: improving cold-start prediction in drug-target affinity with biological LLM. *IEEE Transactions on Computational Biology and Bioinformatics*, 2025.

[4] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[5] Sabrina Jaeger, Simone Fulle, and Samo Turk. Mol2Vec: unsupervised machine learning approach with chemical intuition. *Journal of Chemical Information and Modeling*, 58(1):27–35, 2018.

[6] Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. GraphDTA: predicting drug-target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.

[7] Kyle Yingkai Gao, Achille Fokoue, Heng Luo, Arun Iyengar, Sanjoy Dey, Ping Zhang, et al. Interpretable drug target prediction using deep neural representation. In *IJCAI*, volume 2018, pages 3371–3377, 2018.

[8] Wen Torng and Russ B Altman. Graph convolutional neural networks for predicting drug-target interactions. *Journal of Chemical Information and Modeling*, 59(10):4131–4149, 2019.

[9] Jiahao Xu, Lei Ci, Bo Zhu, Guanhua Zhang, Linhua Jiang, Shixin Ye-Lehmann, and Wei Long. MMSG-DTA: a multimodal, multiscale model based on sequence and graph modalities for drug-target affinity prediction. *Journal of Chemical Information and Modeling*, 65(2):981–996, 2025.

[10] 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.

[11] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: a universal 3D molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023.

[12] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In *ICML*, volume 11, pages 689–696, 2011.

[13] Rong Yin, Ruyue Liu, Xiaoshuai Hao, Xingrui Zhou, Yong Liu, Can Ma, and Weiping Wang. Multi-modal molecular representation learning via structure awareness. *IEEE Transactions on Image Processing*, 2025.

[14] Guishen Wang, Zhitong Guo, Guilin You, Ming Xu, Chen Cao, and Xiaowen Hu. MMDDI-MGPFF: multi-modal drug representation learning with molecular graph and pharmacological feature fusion for drug-drug interaction event prediction. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 467–470. IEEE, 2024.

[15] Viet Thanh Duy Nguyen and Truong Son Hy. Multimodal pretraining for unsupervised protein representation learning. *Biology Methods and Protocols*, 9(1):bpae043, 2024.

[16] Xuefeng Liu, Songhao Jiang, Chih-chan Tien, Jinbo Xu, and Rick Stevens. Bidirectional hierarchical protein multi-modal representation learning. *arXiv preprint arXiv:2504.04770*, 2025.

[17] Zhaohan Meng, Zaiqiao Meng, Ke Yuan, and Iadh Ounis. FusionDTI: fine-grained binding discovery with token-level fusion for drug-target interaction. *arXiv preprint arXiv:2406.01651*, 2024.

[18] Xinlong Zhai, Chunchen Wang, Ruijia Wang, Jiazheng Kang, Shujie Li, Boyu Chen, Tengfei Ma, Zikai Zhou, Cheng Yang, and Chuan Shi. Blend the separated: mixture of synergistic experts for data-scarcity drug-target interaction prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22336–22344, 2025.

[19] Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293, 2014.

[20] 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.

[21] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3):735–743, 2014.

[22] James T Metz, Eric F Johnson, Niru B Soni, Philip J Merta, Lemma Kifle, and Philip J Hajduk. Navigating the kinome. *Nature Chemical Biology*, 7(4):200–202, 2011.

[23] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 2022.

[24] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[25] Atakan Yüksel, Erva Ulusoy, Atabey Ünlü, and Tunca Doğan. SELFormer: molecular representation learning via SELFIES language models. *Machine Learning: Science and Technology*, 4(2):025035, 2023.

[26] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: protein language modeling with structure-aware vocabulary. *bioRxiv*, pages 2023–10, 2023.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[28] Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, et al. SELFIES and the future of molecular string representations. *Patterns*, 3(10), 2022.

[29] Hengliang Guo, Congxiang Zhang, Jiandong Shang, Dujuan Zhang, Yang Guo, Kang Gao, Kecheng Yang, Xu Gao, Dezhong Yao, Wanting Chen, et al. Drug-target affinity prediction based on topological enhanced graph neural networks. *Journal of Chemical Information and Modeling*, 65(7):3749–3760, 2025.

[30] Ding Luo, Dandan Liu, Xiaoyang Qu, Lina Dong, and Binju Wang. Enhancing generalizability in protein-ligand binding affinity prediction with multimodal contrastive learning. *Journal of Chemical Information and Modeling*, 64(6):1892–1906, 2024.

[31] Zhen Wang, Zhanfeng Wang, Maohua Yang, Long Pang, Fangyuan Nie, Siyuan Liu, Zhifeng Gao, Guojiang Zhao, Xiaohong Ji, Dandan Huang, et al. Enhancing challenging target screening via multimodal protein-ligand contrastive learning. *bioRxiv*, pages 2024–08, 2024.

[32] Gregory W Kyro, Anthony M Smaldone, Yu Shee, Chuzhi Xu, and Victor S Batista. T-ALPHA: a hierarchical Transformer-based deep neural network for protein-ligand binding affinity prediction with uncertainty-aware self-learning for protein-specific alignment. *Journal of Chemical Information and Modeling*, 65(5):2395–2415, 2025.

[33] 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.

[34] Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *bioRxiv*, pages 2022–02, 2022.

[35] Hamid Hadipour, Yan Yi Li, Yan Sun, Chutong Deng, Leann Lac, Rebecca Davis, Silvia T Cardona, and Pingzhao Hu. GraphBAN: an inductive graph-based approach for enhanced prediction of compound-protein interactions. *Nature Communications*, 16(1):2541, 2025.

[36] Raihana Ferdous et al. An efficient k-means algorithm integrated with Jaccard distance measure for document clustering. In *2009 first Asian Himalayas International Conference on Internet*, pages 1–6. IEEE, 2009.

[37] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

[38] Steven H Bertz. The first general index of molecular complexity. *Journal of the American Chemical Society*, 103(12):3599–3601, 1981.

[39] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024.

[40] Warren L DeLano et al. PyMOL: an open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography*, 40(1):82–92, 2002.

# A Methods

## A.1 Model Architecture

The whole process of MomeDTA can be formulated as Algorithm 1. The algorithms in each line are elucidated in the following subsections.

**Algorithm 1** MomeDTA

**Input:** $d_1, p_1, d_3, p_3$ defined in following algorithms, drug graph $d_2$, protein graph $p_2$;
**Output:** Affinity score $s$;
1: $F_1 = \text{1D-Track}(d_1, p_1)$; // see Algorithm 4
2: $F_2 = \text{concat}([\text{MGNN}(d_2), \text{MSAGE}(p_2)])$; // see Section A.3
3: $F_3 = \text{3D-Track}(d_3, p_3)$; // see Algorithm 5
4: $s = \text{ViewMix}([F_1, F_2, F_3])$. // see Algorithm 6

## A.2 1D Track

In the 1D track, the input is the one-dimensional SMILES and amino acid sequence. Inspired by [17], we convert SMILES to SELFIES [28] and adopt the pretrained SELFormer [25] as the encoder. SELFIES is a string-based representation that circumvents the issue of robustness and that always generates valid molecular tokens for each character [17, 28]. For proteins, we use ESM-2 [4] as encoder since it is one of the most widely used protein encoders in DTA prediction [3, 10, 31, 32]. Trained on a great number of unlabeled protein sequences, ESM-2 efficiently captures proteins' sequential and evolutionary information and shows great generalizability in many downstream tasks [4]. After getting the outputs of the PLMs, we need to align them to the same size. Here, we adopt the one-dimensional convolutional neural network (1D-CNN) as in [3], which down-samples the hidden dimension of the original embedding while retaining multiscale neighborhood information.

**Algorithm 2** MAN

**Input:** Drug representation $D \in \mathbb{R}^{e \times N}$, drug mask $D_m \in \mathbb{R}^N$,
protein representation $P \in \mathbb{R}^{e \times L}$, and protein mask $P_m \in \mathbb{R}^L$;
**Output:** Protein-aware drug representation $\hat{D} \in \mathbb{R}^e$ and drug-aware protein representation $\hat{P} \in \mathbb{R}^e$;
1: $C = \tanh(P^T W_b D)$, $W_b \in \mathbb{R}^{e \times e}$;
2: $H^d = \tanh(W_d D + (W_p P)C)$, $H^p = \tanh(W_p P + (W_d D)C^T)$, $W_d, W_p \in \mathbb{R}^{k \times e}$;
3: $A^d = \text{softmax}((W_{hd}^T H^d).\,\text{masked\_fill}(\neg D_m))$,
$A^p = \text{softmax}((W_{hp}^T H^p).\,\text{masked\_fill}(\neg P_m))$, $W_{hd}, W_{hp} \in \mathbb{R}^{k \times 1}$;
4: $\hat{D} = A^d D$, $\hat{P} = A^p P$.

After obtaining the above representations, we use a refined mutual attention (MAN) for information fusion, as shown in Algorithm 2). Note that $softmax$ is calculated along the embedding dimension, and $masked\_fill$ is the same as that in the pytorch library. There are also other similar cross-attention blocks widely used in DTA prediction including BAN [3] and CAN [17]. However, in our experiments, BAN tends to have too large attention logits during forward propagation, which may cause unstable loss curve during training, and CAN has large computation cost, while MAN is relatively stable and efficient with minor accuracy loss, so we choose MAN in our work. Compared to original mutual attention [10], we consider drug and protein masks in the process, which can eliminate noise in the padding region that affects prediction accuracy.

Next, similar to [3], the outputs of MAN and encoder blocks are fed into a fusion block together (see Algorithm 3). This approach enables the model to learn both the global feature of the drug-protein complex and the features of the intermolecular interaction. Lastly, the fused feature is passed through an MLP to get the final affinity score. The whole process can be summarized in Algorithm 4.

**Algorithm 3** FusionBlock

**Input:** Drug representation $D$, protein representation $P$,
protein-aware drug representation $\hat{D}$, and drug-aware protein representation $\hat{P}$;
**Output:** Fused feature $F$;

$PreFeat = \text{LayerNorm}(\text{GELU}(\text{Linear}(\text{concat}([\text{MaxPool}(D), \text{MaxPool}(P)]))))$;
$PostFeat = \text{GELU}(\text{Linear}(\text{concat}([\hat{D}, \hat{P}])))$;
$F = \text{Linear}(PreFeat + PostFeat)$.

**Algorithm 4** 1D-Track

**Input:** SELFIES $d_1$ and amino acid sequence $p_1$;
**Output:** 1D output feature $F_1$;

$d_1 = \text{Embed}_{\text{SELFormer}}(d_1)$, $p_1 = \text{Embed}_{\text{ESM-2}}(p_1)$;
$(D, D_m) = \text{Pad}(d_1)$, $(P, P_m) = \text{Pad}(p_1)$;
$D = \text{1D-CNN}(\text{SELFormer}(d_1)), P = \text{1D-CNN}(\text{ESM-2}(p_1))$;
$\hat{D}, \hat{P} = \text{MAN}(D, P, D_m, P_m)$;
$F_1 = \text{FusionBlock}(D, P, \hat{D}, \hat{P})$.

### A.3 2D Track

In the 2D track, the input is the graph of the drug and the protein. A graph has a topological structure and is often considered as a modality between the 1D sequence and the 3D structure, providing a unique view related to drug-target binding [6, 33].

Table 3: Residue features used in protein graph.

| Feature | Size |
|---|---|
| Type | 21 |
| Hydrophobicity | 1 |
| Charge | 1 |
| Polarity | 1 |
| Molecular weight | 1 |

In this track, we use GNN to encode both the drug and protein graphs. For the drug graph, we utilize the MGNN proposed in [33], which contains multiscale blocks and dense connection to learn both local and global structural information in the molecule. For the protein graph, for simplicity, we choose some basic residual features as node features, as shown in Table 3. The edges are determined by the distance between the $C_\alpha$ atom of the residues. Then, similar to the idea of MGNN, we design a multiscale GNN called MSAGE consisting of 3 SAGEConv blocks with residual connection and pooling, as depicted in Figure 4. After going through GNNs, drug and protein representations are concatenated as fed into the final MLP.

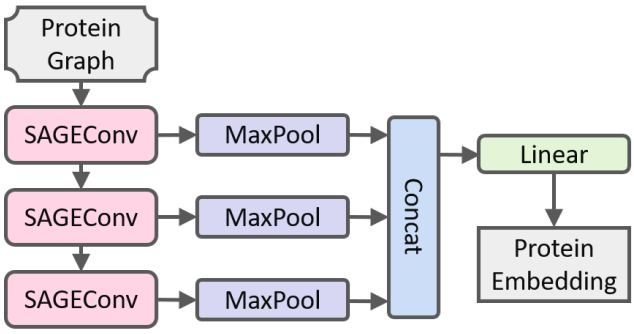


Figure 4: MSAGE architecture.

### A.4 3D Track

The process of 3D track is the same as that of 1D track except that the inputs are different and the encoders are replaced with 3D PLMs (see Algorithm 5). In this track, we focus on the coordinates of the drugs and proteins since they provide fine-grained structure information. The coordinates of the drugs can be obtained by RDKit. For proteins, we use AlphaFold2 to predict their 3D structures, and the details are in Section 3.1.

**Algorithm 5** 3D-Track

**Input:** Drug conformation $d_3$ (obtained by RDKit) and protein structure $p_3$ (predicted by AlphaFold2);
**Output:** 3D output feature $F_3$;

$d_3 = \text{Embed}_{\text{Uni-Mol}}(d_3)$, $p_3 = \text{Embed}_{\text{SaProt}}(p_3)$.
$(D, D_m) = \text{Pad}(d_3)$, $(P, P_m) = \text{Pad}(p_3)$;
$D = \text{1D-CNN}(\text{Uni-Mol}(d_3)), P = \text{1D-CNN}(\text{SaProt}(p_3))$;
$\hat{D}, \hat{P} = \text{MAN}(D, P, D_m, P_m)$;
$F_3 = \text{FusionBlock}(D, P, \hat{D}, \hat{P})$.

Inspired by previous works [10, 31], we choose Uni-Mol [11] and SaProt [26] as the encoders for drug and protein respectively. Uni-Mol was trained on millions of molecules and their 3D conformations, significantly enhancing the quality of molecular representations [10, 11]. SaProt uses Foldseek [34] to convert local structures to tokens, and then combines it with amino acid letters to form a "structure-aware" protein sequence as input to ESM-2 [26], which is a convenient way to fuse structure information into sequence encoder.

### A.5 ViewMix

The detailed computing process of ViewMix is shown below.

**Algorithm 6** ViewMix

**Input:** 1D output feature $F_1$, 2D output feature $F_2$, and 3D output feature $F_3$;
**Output:** Affinity score $s$;

$F = \text{concat}([F_1, F_2, F_3])$;
$F_g = \text{LayerNorm}(\text{GELU}(\text{Linear}(F)))$;
$W = \text{softmax}(\text{Unflatten}(\text{Linear}(F_g)))$;
$H = \text{stack}(F)$;
$F_{new} = HW$;
$s = \text{MLP}(F_{new})$.

To detect whether ViewMix learns the importance of different views, we analyze the training dynamics of the weight $W$ in ViewMix, as depicted in Figure 5. We choose the pair (AT-7519, FLT1) in the training set under cold-start setting on Davis, and visualize the change of the values of 3 different feature dimensions in $W$ (i.e. $W \in \mathbb{R}^{256\times3}$, and we extract $W[100]$, $W[150]$, and $W[200]$) with regard to the training epoch. Obviously, the weights are oscillating in the beginning but becomes stable when the number of epochs increases, which complies with the convergence of the whole model. In each dimension, the weight of one of the 3 views becomes dominant at last, indicating that it is the most crucial view related to this feature and is successfully captured by our model. Meanwhile, the dominating view differs across different dimensions, which shows that the importance of each view varies between different features, further validating the necessity of proportionally fusing weight matrices instead of prediction logits.

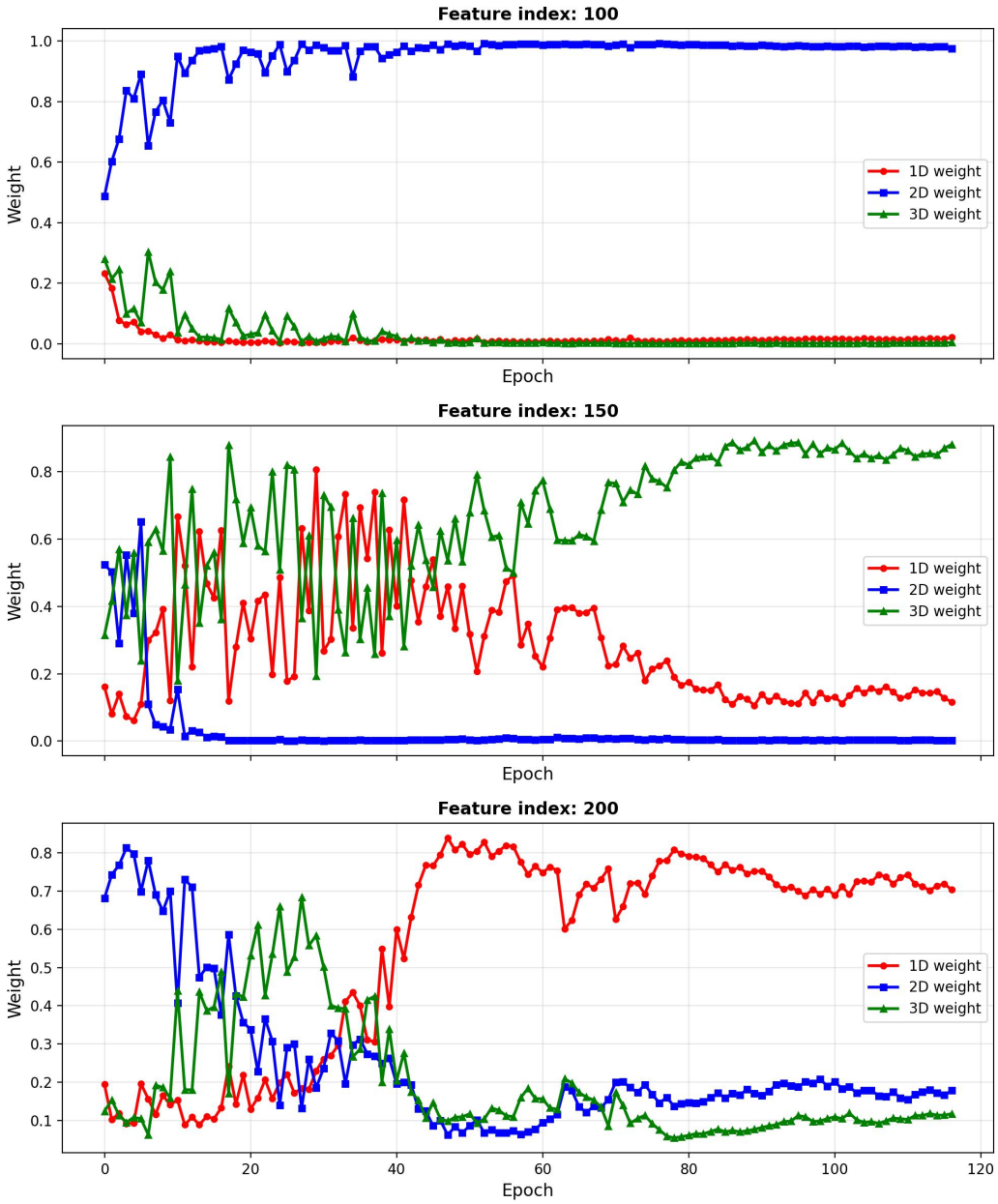


Figure 5: The training dynamics of the weight in ViewMix.

## B Baselines

Table 4: Architecture and the corresponding views of baseline methods.

| Methods | Drug Representation | Protein Representation | Fusion |
|---|---|---|---|
| DeepDTA (2018) [1] | CNN | CNN | - |
| GraphDTA (2020) [6] | GNN | CNN | - |
| LLMDTA (2024) [3] | Mol2Vec [5] | ESM-2 | BAN |
| MMSG-DTA (2025) [9] | GINE+Transformer | Sequence: MCNN;<br>Graph: ESM-2+GateGAT | Attention |

# C Data

## C.1 Data Split

For data split under cold-start settings, following GraphBAN [35], we first cluster all the drugs based on Jaccard distance [36] between their ECFP fingerprints [37] and cluster all the proteins based on cosine similarity between their 3-mer embeddings. Then, we split all the drugs into 5 folds such that the drugs within the same cluster must be in the same fold and the total number of drugs in each fold is nearly the same. The same method is applied to proteins. This ensures that drugs and proteins from different folds are biologically dissimilar. Next, we randomly select one fold of drugs (assume fold 0, and denote $\{D_0\}$) and one fold of proteins (denoted $\{P_0\}$), and let all data pairs $(d, p)$ where $d \in D_0$ and $p \in P_0$ be the test set. We select another fold (assume fold 1), and all $(d, p)$ where $d \in D_1$ and $p \in P_1$ are the validation set. Under this condition, all $(d, p)$ where $d \in D_2 \cup D_3 \cup D_4$ and $p \in P_2 \cup P_3 \cup P_4$ are the training set.

## C.2 Data Analysis

To demonstrate that different datasets exhibit data bias and thus have dependencies on modalities of different views, we conduct two experiments on the drugs in our datasets. Firstly, we show the presence of data bias, which means that data itself may have intrinsic properties suitable for methods based on different kinds of views. We extract the Morgan fingerprints of all the drugs in Davis, Kiba, and Metz, which encode the topological structure of molecules, and calculate the pairwise distance between the fingerprints within each dataset. As shown in Figure 6, Davis has the lowest molecular fingerprint diversity, while Kiba and Metz exhibit higher molecular fingerprint variability. This implies why graph-based methods, which primarily extract topological features, perform well on Kiba and Metz.

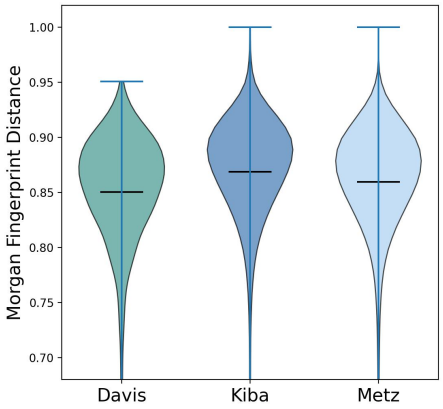


Figure 6: Distribution of Drug Pairwise Distance measured by Morgan Fingerprint in Davis, Kiba, and Metz.

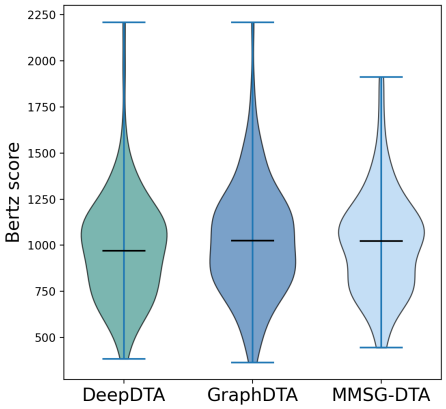


Figure 7: The distribution of Bertz scores of the best-predicting drugs of three baseline methods.

Secondly, we show that different methods perform differently on distinct drug-target pairs. We choose Kiba, which is the largest dataset in our study with multiple sources of drug-target pairs [21], and collect the predicted logits of DeepDTA, GraphDTA, and MMSG-DTA on it. Then, we rank drug-target pairs according to the difference between their predicted logits and ground-truth affinities in ascending order, and collect the drugs in the top 100 samples. These are the drugs on which the corresponding method has the best prediction performance. Next, we collect the Bertz scores of these drugs representing their topological complexity [38] and visualize them in Figure 7. The results show that topological complexity of the drugs of the graph-based methods (GraphDTA and MMSG-DTA which use GNN to encode drug graphs) is higher than that of the sequence-based method (DeepDTA which uses CNN to encode SMILES). This demonstrates that graph-based methods show superior

performance on molecules with more complex topological structure, which further indicates that different drug-target pairs have distinct traits and show dependencies on different modalities.

## D Result Analysis

Figures 8 and 9 show the distribution of predicted versus true affinity values of MMSG-DTA and MomeDTA in each dataset in the warm and cold-start scenario respectively. In both scenarios, the predicted values of MomeDTA are closer to ground-truth affinity values than those of MMSG-DTA in all datasets in that the data points of MomeDTA are closer to the $y = x$ line which means perfect prediction. However, in the cold-start scenario, many data points are far from the $y = x$ line for both methods, indicating that the generalizability of the model to novel samples still has room for improvement.

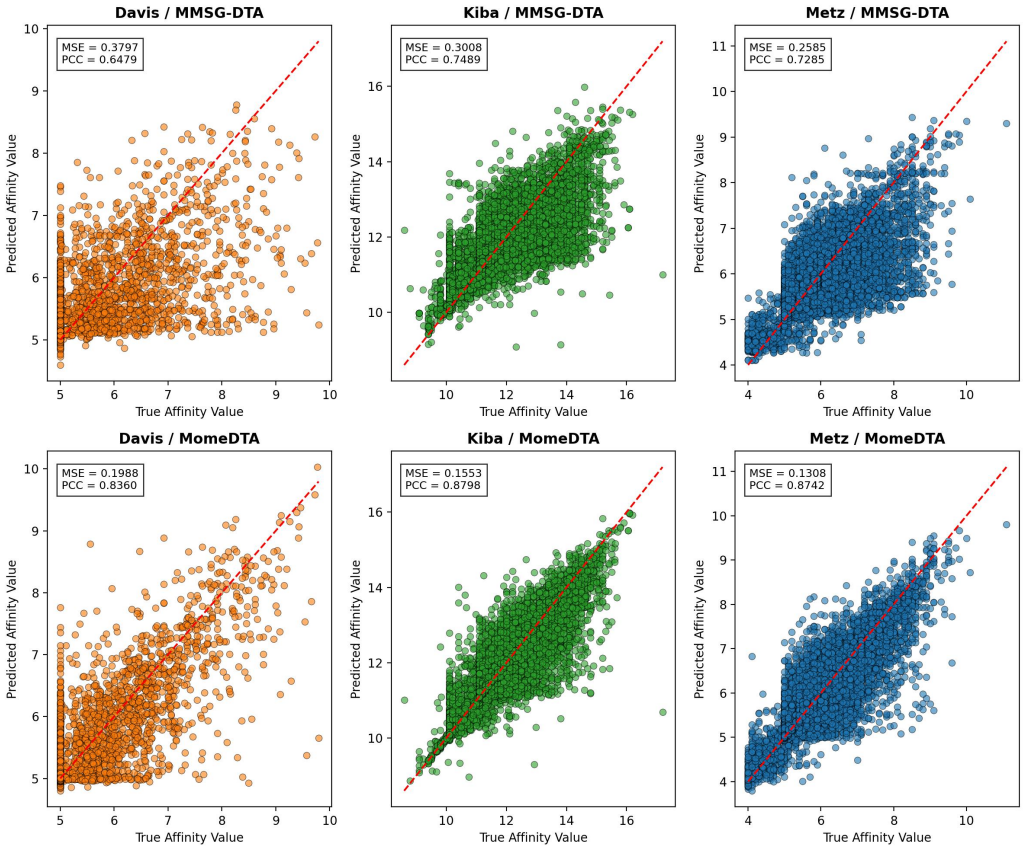


Figure 8: The distribution of predicted versus true affinity values of MMSG-DTA and MomeDTA in the warm scenario.

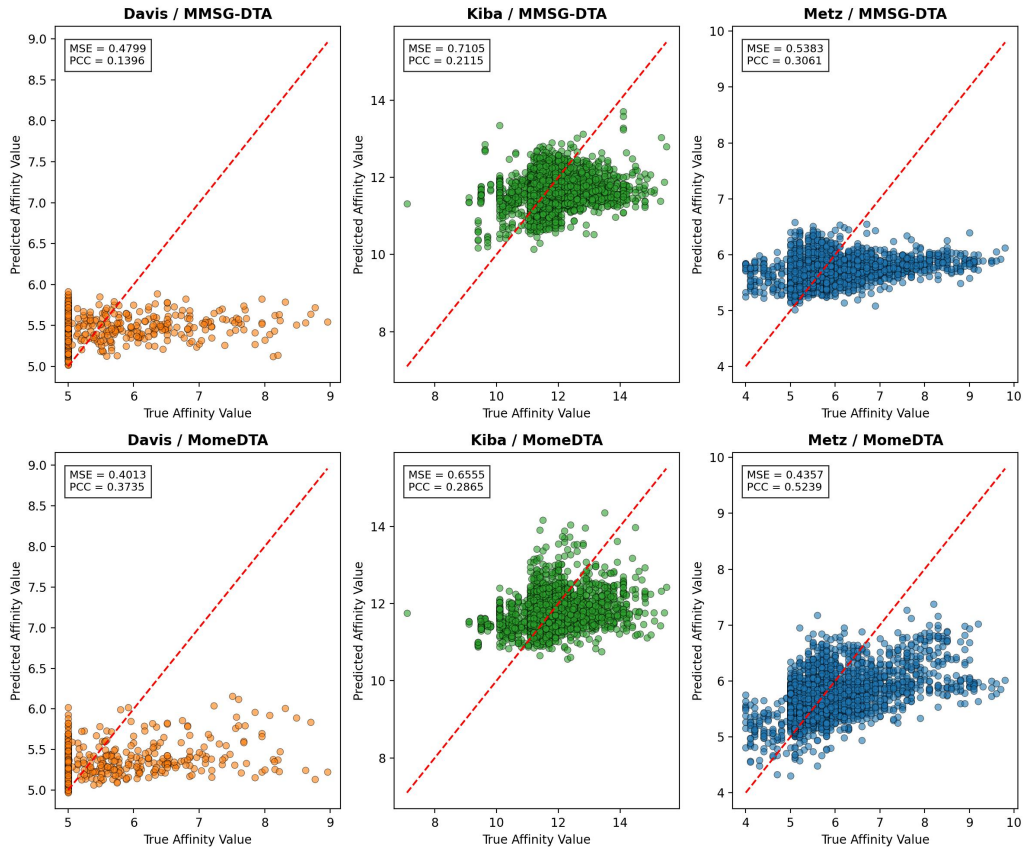


Figure 9: The distribution of predicted versus true affinity values of MMSG-DTA and MomeDTA in the cold-start scenario.

## E Interpretability Analysis

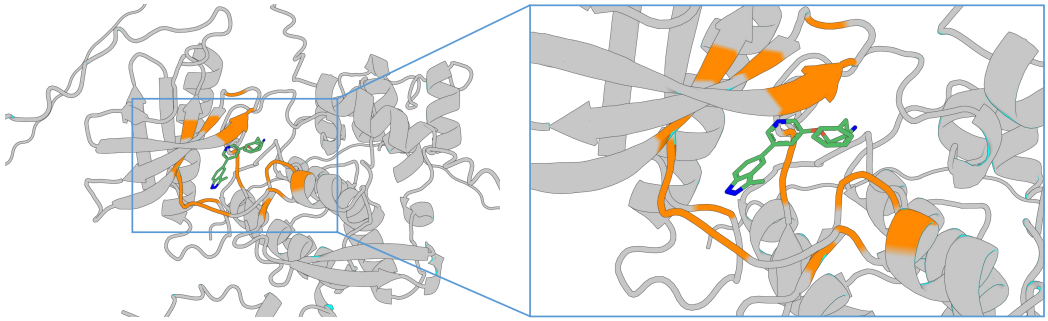

Figure 10: Binding sites predicted by AF3 for the studied drug-target pair. The protein is colored blue and the drug is colored green. The binding sites (mainly in residue region 310-318 and 384-393) are marked orange.

To figure out whether the model actually learns any binding pattern between drugs and targets, we select a drug-target pair in Davis, where the drug is A-674563 and the protein is CLK3, and pass this pair through the model. We extract the attention map of MAN ($C$ in Algorithm 2 in Appendix A) in the 3D track, and visualize it in Figure 11. The positions in the map are the attention logits with regards to the corresponding drug and protein positions for $x$-axis and $y$-axis respectively. Deeper color represents stronger signal. There are some deep "lines" in the map (i.e., target position near 300 and 390; drug position near 11, etc.) which indicates the potential binding sites of the drug and target.

Meanwhile, AlphaFold 3 [39] (AF3) is employed for predicting the structure of drug-target binding complex. We visualize the complex and mark its binding sites (mainly in residue region 310-318 and 384-393) using PyMol [40], as shown in Figure 10. Overall, the deep positions of the protein in the attention map correspond closely to the AF3-predicted binding sites, especially the regions around residue 300 and residue 390. This implies that our model learns the binding mechanisms between drugs and proteins.

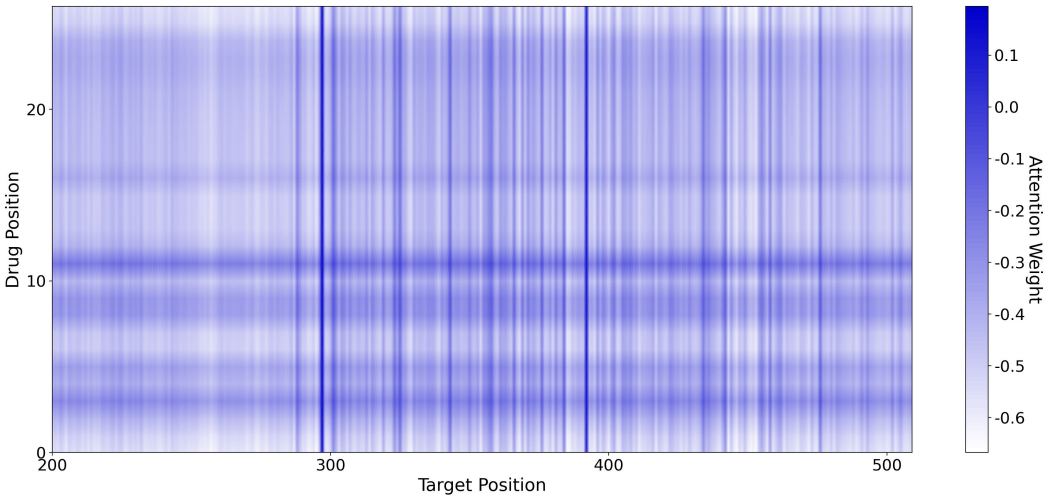


Figure 11: MAN's map in 3D track for the studied drug-target pair. The deepest lines are near 300 and 390.

## F Computational Cost

Finally, we analyze the computational cost of our method. Our memory requirement is at the same level as most baselines (see Table 5). However, the running time of our method is slower than the baselines, mainly because we use features from multiple views and train them together, and the details are shown in Table 7 with the number of data counted in Table 6 for reference. This is the main point to improve for our method, and we will design more concise and efficient way for modality fusion in the future.

Table 5: Average memory consumption of each method (in GB).

| Method | Memory |
|---|---|
| DeepDTA | 1.4 |
| GraphDTA | 1.9 |
| LLMDTA | 1.2 |
| MMSG-DTA | 22.8 |
| Ours | 2.8 |

Table 6: Number of data in each dataset.

| | Davis | | Kiba | | Metz | |
|---|---|---|---|---|---|---|
| | warm | novel_pair | warm | novel_pair | warm | novel_pair |
| Training | 16451 | 9307 | 75573 | 45850 | 68258 | 34889 |
| Validation | 4113 | 975 | 18894 | 4632 | 17065 | 4733 |
| Test | 5140 | 1064 | 23616 | 3572 | 21330 | 4966 |

Table 7: Runtime of each method (in hours).

| Method | Davis | | Kiba | | Metz | |
|---|---|---|---|---|---|---|
| | warm | novel_pair | warm | novel_pair | warm | novel_pair |
| DeepDTA | 0.6 | 0.4 | 1.9 | 1.2 | 1.8 | 1.0 |
| GraphDTA | 0.7 | 0.7 | 3.0 | 3.7 | 2.6 | 3.0 |
| LLMDTA | 4.0 | 1.4 | 33.8 | 6.1 | 16.4 | 3.3 |
| MMSG-DTA | 1.5 | 0.2 | 7.7 | 2.8 | 6.6 | 0.6 |
| Ours | 19.3 | 3.3 | 91.1 | 13.1 | 41.5 | 12.0 |