MDM-DTA: Message Passing Neural Network with Molecular Descriptors and Mixture of Experts for Drug-Target Affinity Prediction

Yang Dai[a,1], Xiaoyu Tan[a,b,1], Haoyu Wang[a,1], Gengchen Ma[a], Yujie Xiong[a] and Xihe Qiu [a,*]

[a]School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China
[b]INF Technology (Shanghai) Co., Ltd., Shanghai, China

ARTICLE INFO

Keywords:
drug discovery
drug–target binding affinity
mixture of experts
forecast revision

ABSTRACT

Background and Objective: Drug-target affinity (DTA) prediction is a pivotal task in computational drug discovery, enabling the estimation of binding affinities between small molecules and their target proteins. This process is essential for reducing the costs, development time, and risks inherent in traditional drug development pipelines. Current DTA prediction models primarily rely on separate extraction and concatenation of drug and protein features. However, these models often fail to account for the complex semantic relationships within protein sequences, which limits their ability to accurately predict affinity.

Methods: In response to these challenges, we propose MDM-DTA, a novel framework leveraging a Mixture of Experts (MoE) strategy to integrate diverse molecular and protein representations. For drug representation, MDM-DTA utilizes molecular graphs, which are processed via Message Passing Neural Networks (MPNNs), alongside molecular descriptors that are passed through a three-layer convolutional neural network (CNN). Protein features are extracted using a deep convolutional network enhanced with Squeeze-and-Excitation (SE) mechanisms to capture inter-channel dependencies. Furthermore, protein sequence semantics are encoded through pre-trained embeddings from a knowledge-guided Bidirectional Encoder Representations from Transformers (BERT) model and the Evolutionary Scale Modeling 2 (ESM2) model, enabling the model to capture contextual relationships within protein sequences.

Results: Extensive experiments on three benchmark datasets demonstrate that MDM-DTA consistently outperforms state-of-the-art models of similar complexity in terms of predictive accuracy. The incorporation of both structural and semantic features significantly enhances the model's ability to predict drug-target binding affinities, highlighting the importance of a multi-modal representation approach.

Conclusions: The proposed MDM-DTA framework effectively integrates both molecular and semantic protein representations, providing superior performance in DTA prediction tasks. The results underscore the potential of MDM-DTA to improve the accuracy of computational drug discovery models, facilitating the identification of novel drug candidates and advancing the field of in silico drug development.

1. Introduction

The drug development process typically takes ten years to complete, starting with the initial research stage, which lays the foundation for further studies[1]. Throughout this process, the costs accumulate, often exceeding one billion dollars[2]. These expenses are incurred from the early stages all the way to commercialization[3, 4]. Most drug development failures in clinical trials are due to insufficient target validation[5]. This highlights the importance and significance of accurately screening drugs for specific targets. Studies have shown that a higher DTA indicates a stronger binding between the drug and its target, resulting in better therapeutic effects[6]. Furthermore, the strength of DTA has become one of the critical indicators in drug screening[7, 8], and the identification of DTA is crucial in drug design and development[9, 10].

As shown in Figure 1, DTA prediction methods have evolved in three stages: starting with single-feature approaches, progressing to simple feature fusion, and finally enhancing fusion using advanced mechanisms to improve prediction accuracy. The first two stages of DTA prediction methods have notable limitations. Early single-feature methods primarily focused on a single modality of information, such as sequence data, molecular graphs, or drug-target interaction networks, neglecting the complementary nature of different modalities. While effective in specific contexts, these methods fail to fully capture the complex interactions between drugs and targets due to limited information, resulting in restricted predictive power. As the field progressed, simple feature fusion methods emerged, combining multiple modalities like molecular sequence and structure. However, these methods typically employed shallow fusion strategies, such as concatenation or weighted averaging, which were unable to fully capture the deep relationships between modalities. This approach often led to information redundancy, modality mismatch, or imbalanced weighting of key features, ultimately affecting the stability and generalization of predictions.

To systematically address these limitations, a layered feature integration framework can be employed to facilitate multi-modal biomolecular characterization. At the foundational level, traditional architectures such as CNNs [11], LSTMs [12], and Transformers [13] contribute to sequence

*Corresponding author
[1]These authors contributed significantly to this work.

and structural representation learning. Specifically, CNNs [11] utilize grid-based convolution to process protein contact graphs [14]; LSTMs [12] capture sequential dependencies within biomolecular sequences; and Transformers [13] leverage self-attention mechanisms to model long-range interactions.

Building on this foundation, GNNs [15] enhance the encoding of drug molecule graphs [10] with topological precision, ensuring a more faithful representation of molecular structures. Meanwhile, large language models introduce semantic perception, effectively mapping drug SMILES sequences and protein amino acid chains into a shared latent vector space [16] to capture biochemical semantics. Notably, our previous work, G-K BertDTA [17], demonstrated the effectiveness of integrating these orthogonal representations, achieving state-of-the-art performance through graphical-language fusion.

Despite recent advancements, challenges remain. DMPNN-Des [18] has refined molecular property prediction using dynamic message passing networks, while large-scale protein structure prediction [19] has revealed strong connections between sequence, structure, and biochemical properties. However, existing methods often process different modalities independently, limiting cross-modal interactions. Moreover, fusion strategies struggle to balance information across modalities, leading to redundant or misaligned representations.

To address these challenges, we propose the Message Passing Neural Network with Molecular Descriptors and Mixture of Experts for Drug-Target Affinity Prediction (**MDM-DTA**). Our framework integrates sequence data, molecular graphs, and semantic embeddings from language models to provide a comprehensive representation. A mixture of experts (MoE) mechanism [20] dynamically selects the most relevant features using a top-k gating strategy [21], ensuring effective fusion. Additionally, we introduce isotropic regression correction [22] to reduce prediction variance caused by input sensitivity, leading to more stable and accurate affinity estimates. By unifying multi-modal insights with adaptive feature selection, **MDM-DTA** enhances drug-target interaction modeling, improving both prediction accuracy and generalization.

In summary, the main contributions of this paper are as follows:

- We are the first to combine a Message Passing Neural Network (MPNN) [23] framework with molecular descriptors in a DTA prediction task, significantly enhancing the model's performance and stability.
- We introduced a large protein language model to extract semantic information from proteins, enriching the model with additional biological prior knowledge, which greatly improved its performance and robustness.
- We incorporated the mixture of experts (MoE) [20] mechanism by leveraging protein convolutional layers, a large protein language model for semantic extraction, and a SMILES-based molecular model. Using the top-k principle, we employed a sparse expert gating network to allocate weights and fuse information from protein convolution, protein semantics, and SMILES semantics.
- We applied isotonic regression to address the issue of unreasonable ordering in model predictions, ensuring that the results exhibit clear monotonicity, thereby preserving their biological relevance and logical consistency.

The remainder of this article is organized as follows: We review the previous work and concepts in Section 2. The proposed method is described in detail in section 3. Section 4 presents the results of our experiment. Section 5 presents the ablation experiment and its analysis. Sections 6 and 7 discuss and summarize our work and look to the future.

2. Related work

In recent years, machine learning and deep learning have emerged as the predominant approaches for drug–target affinity (DTA) prediction, owing to their capability to model complex nonlinear relationships between molecular entities [24, 25]. Within deep neural network-based frameworks, the structural and sequential characteristics of drugs and proteins are typically represented as one-dimensional sequences, two-dimensional topological maps, or three-dimensional spatial grids. These diverse representations are then processed by specialized neural architectures to extract discriminative features, which serve as the foundation for regression-based affinity prediction [11].

2.1. Single-Modal Feature-Based DTA Prediction

To enhance predictive performance, various strategies have been developed. For example, *Li et al.* [26] employed a random forest model incorporating docking score functions, demonstrating robust predictive capability even when handling low-quality data. *Öztürk et al.* [27] proposed a dual CNNs framework, where drug SMILES sequences and protein sequences were processed independently to extract feature representations, followed by affinity prediction through fully connected layers. *Yuan et al.* [28] introduced a multi-head linear attention mechanism to integrate drug and target information, leveraging knowledge distillation during training to enhance model performance.Although these methods have achieved considerable success, they primarily rely on unimodal feature representations, which may constrain their ability to capture comprehensive contextual information. As a result, such models often fail to fully account for the intricate nature of drug–target interactions, potentially leading to suboptimal predictive performance.

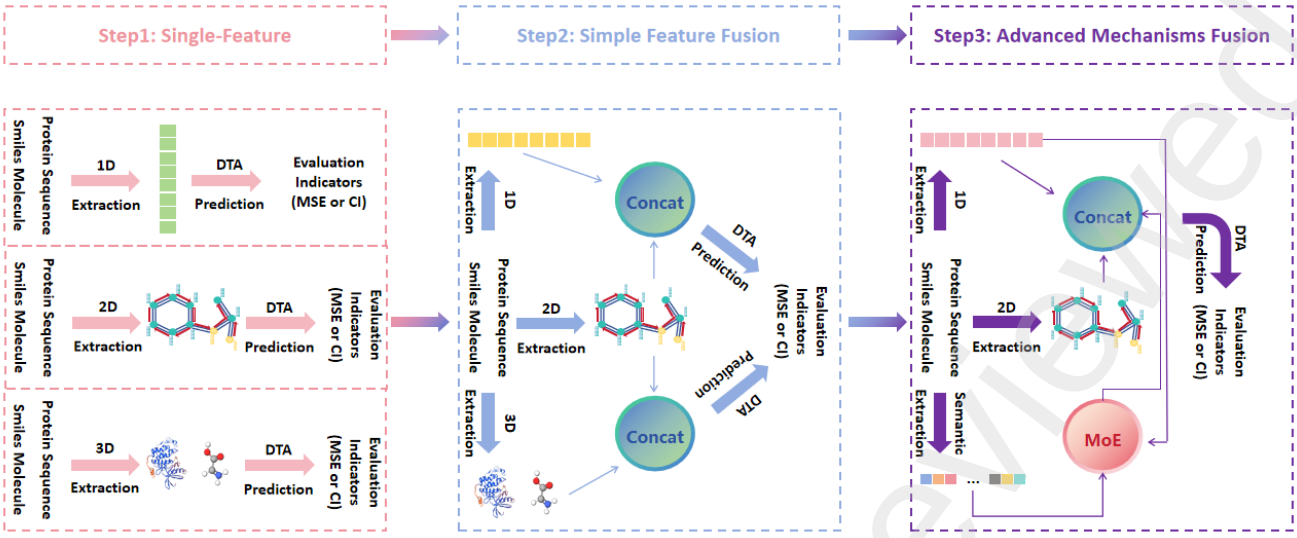


Figure 1: The approach to DTA prediction progresses through three stages: beginning with the use of a single feature, advancing to simple feature fusion, and ultimately improving fusion through the integration of advanced mechanisms for enhanced prediction accuracy. Step 1 involves the use of a single feature for DTA prediction tasks. Step 2 incorporates multiple feature types for DTA prediction, with feature fusion achieved through basic concatenation. Step 3 extends this by utilizing diverse features and refining feature fusion not only via concatenation but also through the integration of advanced mechanisms such as the MoE to further enhance fusion effectiveness.

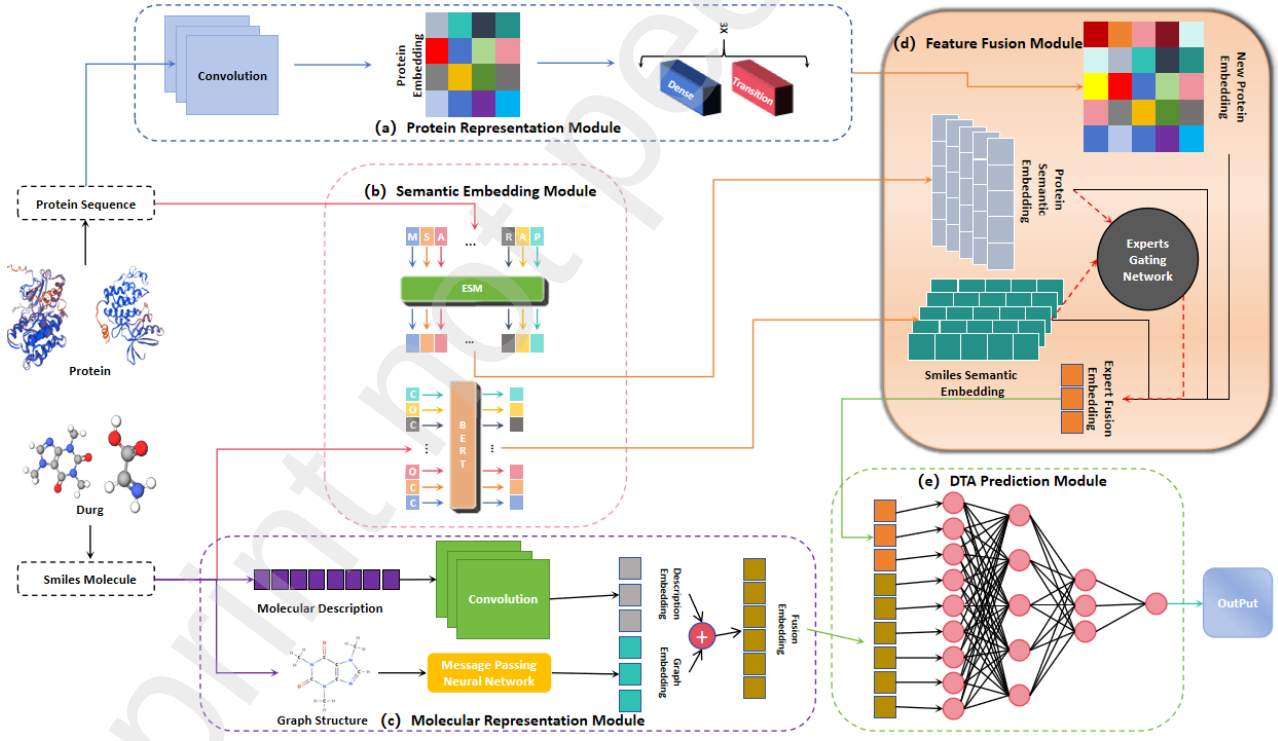


Figure 2: The overall architecture of MDM-DTA: (a) Protein Representation Module, (b) Semantic Embedding Module, (c) Molecular Representation Module, (d) Feature Fusion Module, and (e) DTA Prediction Module.

2.2. Multi-Modal Feature Fusion for DTA Prediction

To overcome the limitations of single-modal representations, researchers have increasingly explored multi-modal fusion techniques, leveraging the inherent graph-structured properties of drug molecules and proteins. The emergence of graph neural networks (GNNs) [15] has further established graph-based representation learning as a promising approach for drug–target affinity (DTA) prediction. Notable contributions in this domain include *Ruan et al.* [10],

who developed a hybrid model integrating molecular graph-based GNNs for drug representation with sequence-based CNNs for protein characterization; *Li et al.* [11], who introduced PocketDTA to model spatial drug-binding site structures; and *Wang et al.* [29], whose MSGNN-DTA employed gated skip connections to facilitate multi-scale topological fusion. Additionally, *Lin et al.* [30] proposed a comprehensive framework that incorporates molecular topology, SMILES strings, and protein sequences, while *Jiang et al.* [14] addressed the reliance on complex multiple sequence alignments in WGNN-DTA [31] by utilizing amino acid contact maps.Despite these advancements, existing multimodal models frequently adopt simplistic feature concatenation rather than fully exploiting cross-modal complementarity. This limitation hinders their capacity to capture high-order drug–target dependencies and generalize effectively to complex molecular systems.

3. Methods

In this section, we first define the DTA prediction task as a regression problem and propose **MDM-DTA**, a novel prediction model that integrates message passing neural networks, molecular descriptors, and a hybrid expert model. The overall framework of our model is illustrated in Figure 2, which consists of five key components: (a) the Protein Representation Module, which encodes protein sequences into meaningful feature representations; (b) the Semantic Embedding Module, which maps the semantic information of the compounds and proteins into continuous vector spaces to capture their underlying relationships; (c) the Molecular Representation Module, which encodes two primary structural aspects of the compounds (i.e., one-dimensional sequences and two-dimensional topological structures) to extract their structural features; (d) the Feature Fusion Module, which integrates the protein, semantic, and molecular features into a unified feature vector for subsequent processing; and (e) the DTA Prediction Component, which uses the fused feature vector to make accurate predictions of drug-target affinity.

3.1. Protein Representation Module

To represent a protein, we use multiple capital letters to represent multiple amino acids in the protein. To facilitate protein feature extraction, and building on our previous work, G-K BertDTA [17], we encode the protein sequence as a fixed-length numerical vector. Specifically, each amino acid sequence is mapped to a corresponding integer sequence of uniform length, with a maximum sequence length of 1000. Sequences exceeding this length are truncated, while shorter sequences are padded with zeros. The encoding process is as follows:

$$\text{convert} = (x \to i;\ x \in X;\ i \in I) \tag{1}$$

Here, X represents the set of amino acids, I denotes the set of integers from 1 to 25, $x \in X$ is an amino acid, and $i \in I$ is its corresponding index.

Subsequently, each integer in the encoded protein sequence is mapped to a 128-dimensional vector. As shown in Figure 3, this mapped vector is then processed through a module consisting of three layers of CNNs [11] to extract protein feature representations. To further enhance feature extraction, we integrate DenseSEnet [17], which refines the protein representation. DenseSEnet [17] comprises three Dense Blocks and Transition Blocks, where each Dense Block consists of multiple DenseLayers. The output of each DenseLayer is concatenated with the input features, establishing a tightly connected structure. The encoding process is as follows:

$$Y_k = \text{DeseLayer}_k\left(\text{concat}(x_0, y_1, y_2, \ldots, y_{k-1})\right) \tag{2}$$

Here, Y_k represents the output of the k-th DenseLayer. The function concat$(\cdot)$ concatenates the initial input x_0 and the outputs of all previous layers $y_1, y_2, \ldots, y_{k-1}$ as the input to the current DenseLayer.

Squeeze-and-Excitation (SE)[32] blocks are incorporated into each DenseLayer to enhance channel-wise feature recalibration. The process is performed as follows:

$$\text{Squeeze phase: } z_c = \text{GlobalAveragePooling}(x_c) \tag{3}$$

$$\text{Excitation phase: } \hat{z}_c = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot z_c)) \tag{4}$$

Where x_c represents the feature map of the c-th channel, W_1 and W_2 are the weights of the fully connected layers, ReLU$(\cdot)$ is the ReLU activation function, and σ is the Sigmoid activation function.

This method effectively captures local features in sequence data using CNNs and enhances feature extraction by incorporating DenseSEnet [17]. By leveraging the power of DenseLayer and Squeeze-and-Excitation (SE) [32] blocks, it strengthens the overall feature extraction process, resulting in a more refined and comprehensive protein representation.

3.2. Semantic Embedding Module

In the semantic feature extraction process of large language models, one-dimensional sequences are first converted into strings and then segmented into text. These

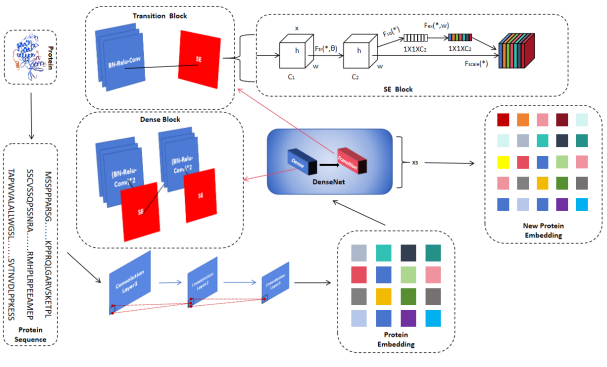


Figure 3: Architecture of the Protein Representation Module.

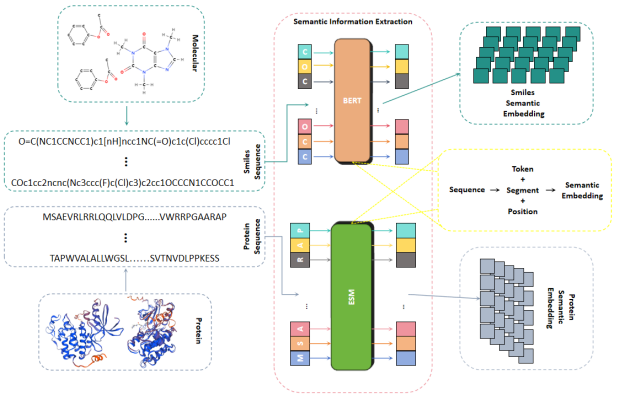


Figure 4: Architecture of the Semantic Embedding Module.

sequences are then encoded into the model, where each sequence character is transformed into a vector representation using maximum pooling. This results in a fixed-length feature vector that captures the semantic features of the sequence. To extract semantic features from SMILES and protein sequences, we incorporate two language models: KB-BERT for SMILES and ESM2 [33] for protein sequences, as shown in Figure 4. For SMILES sequences, the KB-BERT model represents the molecular structure as a one-dimensional string, where each character corresponds to a component of the molecular structure (such as atoms, bonds, etc.). These strings are first converted into sequences of integers through a tokenization process and subsequently mapped to dense vector representations to capture the chemical information represented by SMILES. This processing method is similar to that of protein sequences.Specifically, for the ESM2 [33] model, protein sequences are converted into a set of token IDs, which are then mapped to dense vector representations. The encoding process is as follows:

$$S = (s_1, s_2, \ldots, s_N) \tag{5}$$

$$T = (t_1, t_2, \ldots, t_N) = \text{Tokenizer}(S) \tag{6}$$

Here, S represents the protein sequence, consisting of symbols $s_1, s_2, \ldots, s_N$, where each s_i corresponds to an amino acid (e.g., a one-letter code such as "A", "C", etc.). The sequence $T = (t_1, t_2, \ldots, t_N)$ is the integer sequence obtained after applying tokenization step to S. The tokenizer converts each amino acid symbol s_i into a corresponding integer t_i, typically using a predefined mapping or dictionary. This transformation allows the protein sequence S to be represented numerically as T, making it suitable for computational models and further analysis.

The core component of the ESM2 [33] model is the Transformer encoder, which consists of multiple layers of self-attention mechanisms and feedforward neural networks stacked together. The process is as follows:

$$E = \text{ESM2Embedding}(T) \tag{7}$$

$$X_{\text{out}} = \text{TransformerEncoder}(E) \tag{8}$$

Where E is the embedding matrix of size $N \times d$, where N is the length of the protein sequence and d is the embedding dimension. Each row of E corresponds to the embedding representation of an amino acid in the protein sequence. X_{out} is the final output of the embedded E after being processed by multiple layers of Transformer encoders.

This method employs the ESM2 pretrained model to process protein sequences, converting them into integer representations and inputting them as numerical data into the computational model, thus enabling accurate semantic representation. The Transformer encoder in ESM2 effectively captures long-range dependencies within the sequence through its self-attention mechanism, enhancing the understanding of protein structure and function. The embedding layer transforms the sequence into dense vector representations, reducing the dimensionality of the data and improving processing efficiency.

3.3. Molecular Representation Module

To enhance the feature representation of drugs and address the issue of molecular information loss, we adopted a similar architecture to *DMPNN-Des*[18], which integrates molecular graphs and molecular descriptors for drug representation, as shown in Figure 5. Ablation experiments demonstrate that combining these two methods improves the predictive capability of the model.

The *RDKit* library [34] offers a comprehensive set of molecular descriptors, including molecular weight, *ALogP*, number of hydrogen bond donors, number of hydrogen bond acceptors, polar surface area, number of rotatable bonds, number of aromatic rings, number of warning fragments, and more. The extracted molecular descriptors are then passed through a three-layer CNNs [11] module to obtain more accurate molecular feature representations.

In addition, the *RDKit* library can be utilized to generate a graph structure, complete with node features and an adjacency matrix. Using this graph structure, molecular graph features are extracted through the Message Passing Neural Network. After performing a linear transformation on the input node features x, messages are propagated by aggregating the features of adjacent nodes, based on the edge information x_j in the graph. The aggregated messages are then updated at the MLP level.

Specifically, the MPNN consists of two main phases: the message passing phase and the readout phase. During

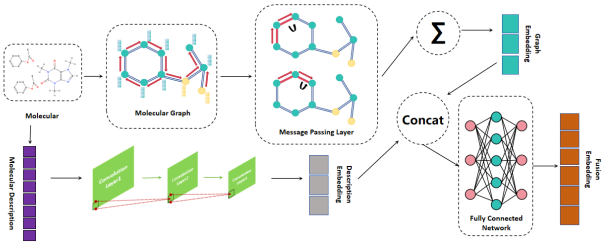


Figure 5: Architecture of the Molecular Representation Module.

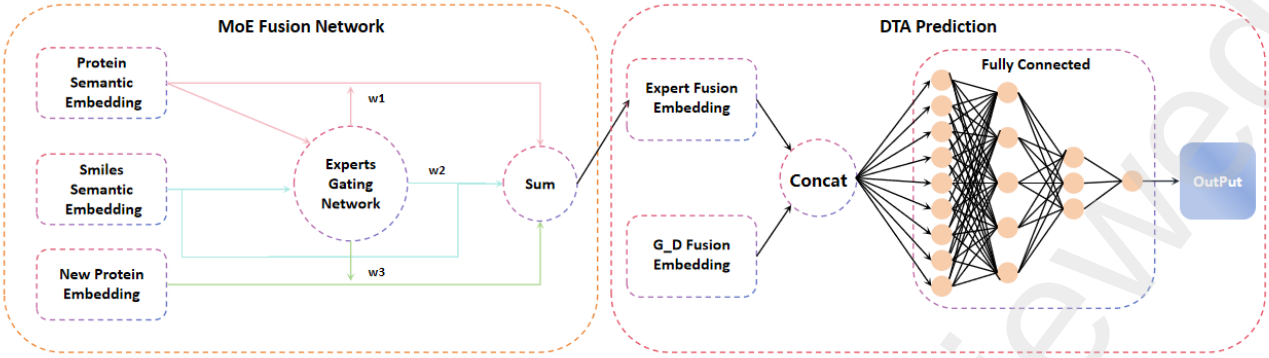


Figure 6: Architecture of the Feature Fusion and DTA Prediction Modules

the message passing phase, which is typically repeated for several iterations, each node gathers information from its neighboring nodes to form messages. These messages often depend on the edge features between nodes and are aggregated—commonly via a summation function—to integrate the features of neighboring nodes. In the subsequent update step, each node applies a multilayer perceptron (MLP) to the aggregated messages, performing a nonlinear transformation to update its hidden state. This enables the node to effectively capture local structural information.The update operation at each step can be expressed as:

$$h_i' = \mathrm{MLP}\left(\sum_{j \in \mathcal{N}(i)} x_j\right) \tag{9}$$

Where h_i' is the updated feature of node i after the transmission of the message, $\mathcal{N}(i)$ is the set of neighbors of node i, and x_j is the input feature of the neighboring node j. The summation $\sum_{j \in \mathcal{N}(i)} x_j$ aggregates the features of node i's neighbors, and the MLP processes this aggregated information to update the feature of node i.

After several rounds of message passing, a readout function is applied to aggregate the node-level representations into a graph-level embedding. This operation encodes the entire molecular graph into a fixed-dimensional vector, capturing its overall structural characteristics and serving as a comprehensive molecular representation.

After obtaining both the graph embedding features and the molecular descriptor features, the two are concatenated and passed through a fully connected network to obtain the fused feature representation of the SMILES.

This approach combines molecular graph features and molecular descriptors to build a more robust feature representation, providing the model with a comprehensive and accurate understanding of drug properties, which significantly enhances predictive performance.

3.4. Feature Fusion Module and DTA Prediction

We observed that simply concatenating protein semantic features and drug semantic features led to a decrease in the model's predictive performance. This is because the semantic features of proteins and drugs may contain irrelevant or redundant information, and simply concatenating them can introduce excessive noise to the model, resulting in a decrease in prediction performance. To address this, we introduced a hybrid expert model to integrate protein semantic features, drug semantic features, and protein features extracted by DenseSEnet [17], as shown in Figure 6. These three types of features are used as the outputs of the expert network. The gating mechanism, based on the top-k principle, allows the model to autonomously select and fuse different feature information. By incorporating gating, we combine the semantic features of protein sequences and drug molecules through a cross-attention mechanism.

The activation probabilities of the experts are determined when the weights are assigned. Then, using the top-k operation, the model selects the top two experts for activation and fusion. To prevent overfitting to the semantic information, we incorporate protein features into the weight allocation, which enhances the model's generalization ability. Experiments demonstrate that the model performs optimally when 2 out of 1-3 experts are activated. The operation is as follows:

$$g_{\text{topk}}, \text{indices}_{\text{topk}} = \text{topk}(g, k = 2) \tag{10}$$

$$\text{mask} = \text{scatter\_zeros\_like}(g_{\text{topk}}, \text{indices}_{\text{topk}}, 1) \tag{11}$$

$$\text{fused\_output} = \sum_i (E \cdot \text{mask})[i] \tag{12}$$

Where g_{topk} contains the top 2 values from g, and $\text{indices}_{\text{topk}}$ contains the corresponding indices of those top values. The function scatter_zeros_like creates a mask of the same shape as g_{topk}, with 1s at the selected indices from $\text{indices}_{\text{topk}}$ and 0s elsewhere. The embedding matrix E is then filtered by the mask, applied element-wise, to produce a filtered output, which is subsequently summed over the indices i.

After obtaining the features from expert fusion and DMPNN-Des fusion, the two are concatenated and passed through a fully connected network for DTA prediction, as shown in Figure 6. To refine the prediction, we apply isotonic regression for calibration. The process is as follows:

$$\hat{y}'_i = \text{IsotonicRegression}(\hat{y}_i) \tag{13}$$

$$\min_f \sum_{i=1}^{n} \left(f(\hat{y}'_i) - y_i\right)^2 \tag{14}$$

Here, f is a monotonic function, typically in the form of piecewise constants. The goal of isotonic regression is to adjust the predicted values $\hat{y}'_i$ such that they are as close as possible to the true values y_i, while maintaining the monotonicity constraint, i.e., $\hat{y}'_i \leq \hat{y}'_{i+1}$ for all i's. This ensures that the calibration process does not violate the natural ordering of the predicted values.

This approach integrates protein semantic features, drug semantic features, and protein features from DenseSEnet using a hybrid expert model with a top-k gating mechanism. The model autonomously selects the most relevant features, improving performance and generalization by preventing overfitting. The top-k operation activates only the most informative experts, while isotonic regression refines predictions to ensure accuracy and monotonicity.

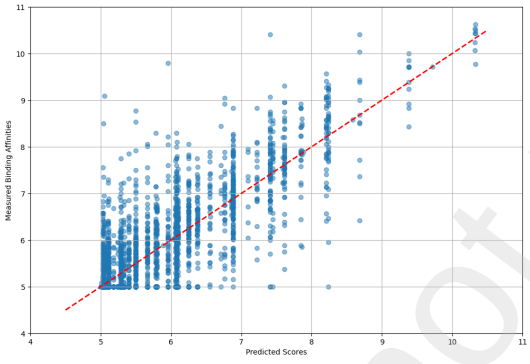


Figure 7: The correlation between the predicted values of MDM-DTA and the true affinities on the Davis dataset.

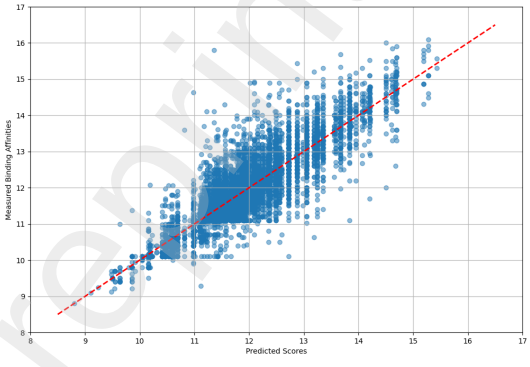


Figure 8: The correlation between the predicted values of **MDM-DTA** and the true affinities on the KIBA dataset.

Table 1
Summary of data sets for baseline comparison in DTA prediction tasks

Dataset	Proteins	Drugs	Binding entries	Train	Test
Davis	442	68	30056	25046	5010
KIBA	229	2111	118254	98545	19709
Metz	170	1423	35259	28207	7052

4. Results

4.1. Data Preparation and Experimental Setup

To comprehensively evaluate the performance of our proposed MDM-DTA, we utilized two widely used benchmark datasets for drug–target affinity prediction: the Davis dataset [35] and the KIBA dataset [36]. To further enrich the evaluation, we additionally included the Metz dataset [37] in our experiments. Table 1 presents detailed statistical information of the three datasets, and their key characteristics for the regression task are summarized as follows.

The Davis dataset, created by *Davis et al.*, contains binding affinity data for 68 compounds and their 442 protein targets. For each compound, the binding affinity is experimentally measured using the dissociation constant (Kd), which reflects the strength of the interaction between the molecule and its target.

$$pK_d = -\log_{10}\left(\frac{K_d}{10^9}\right) \tag{15}$$

Where pK_d is the logarithmic transformation of the dissociation constant K_d, which is a measure of the binding affinity between a drug and its target. The K_d value is expressed in units of Molar (M), and the transformation is scaled by 10^9 to make the value more manageable.

The KIBA dataset, introduced by *Tang et al.*, is an integrated bioactivity matrix that provides affinity data for 2,111 compounds and 229 targets. The KIBA score reflects the interaction between a drug and its target.

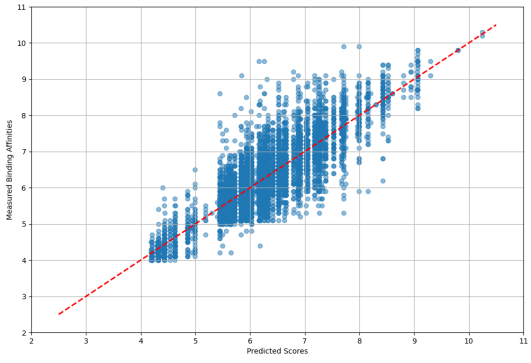


Figure 9: The correlation between the predicted values of **MDM-DTA** and the true affinities on the Metz dataset.

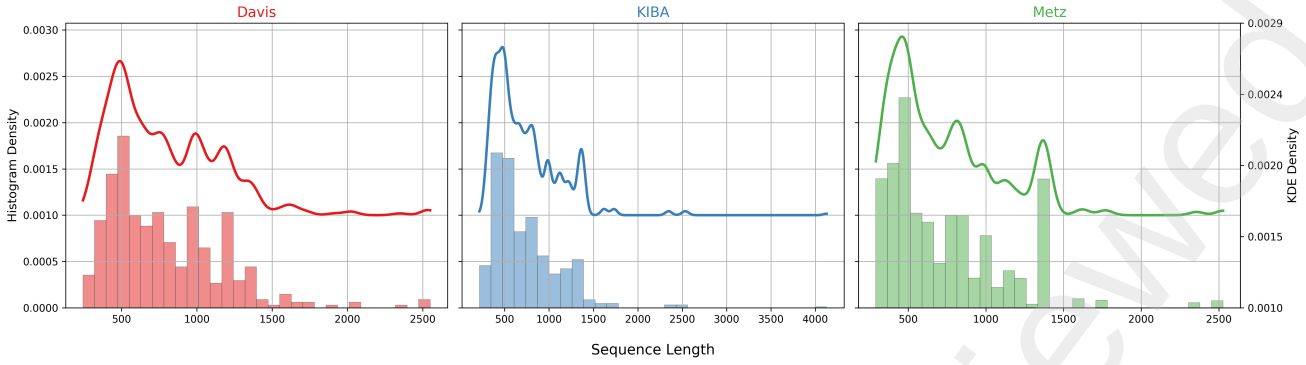


Figure 10: Comparison of protein sequence length distribution among the Davis, KIBA and Metz datasets.

The Metz dataset contains 1,423 compounds and 170 targets, with binding affinity provided in the form of pKi values.

In Figures 7, 8, and 9, the affinity distributions across the three datasets—Davis, KIBA, and Metz—are illustrated and compared. Among them, the Davis dataset exhibits a relatively narrow affinity range, predominantly concentrated between 5 and 7, indicating a bias toward low-affinity interactions. In contrast, the KIBA dataset demonstrates the widest affinity distribution, approximately ranging from 2 to 14, and encompasses the largest sample size. Its data points are densely concentrated between 8 and 12, reflecting a predominance of medium to high-affinity interactions. The Metz dataset shares a similar distribution range with KIBA, but with a comparatively lower sample density, and its affinity values are primarily distributed within the 7 to 11 interval. These differences highlight the distinct statistical characteristics and data coverage of each benchmark dataset.

Based on the visualization analysis of the protein sequence length distribution in Figure 10,the three datasets show significant statistical and biological differences: The Davis dataset presents a highly concentrated single-peak distribution (the main peak is around 500), and more than 99% of the sequence lengths are strictly limited (<1600); Although the Metz dataset has a main peak near 500, it presents a bimodal distribution pattern as a whole (the secondary peak is in the range of 1000-1500). However, the KIBA dataset shows fundamental heterogeneity - although both are univariate distributions, its distribution range shifts significantly to the right (extending to >2500), and there are clearly very long sequences (>2500).

Based on the preceding analysis, each dataset was divided into two parts: one part was designated as the test set, while the remaining data was used for cross-validation during training. Evaluation on these three datasets enables a comprehensive assessment of our model's predictive performance.

To measure performance, we adopt three widely used metrics: Mean Squared Error (MSE), Concordance Index (CI), and the squared correlation coefficient with adjustment, denoted as R_m^2. Their formulations are as follows:

Table 2
Hyperparameters and experimental environment settings

Hyperparameter	Setting
GPU	NVIDIA RTX 3090 24GB
Software Environment	Pytorch,CUDA 11.7
Learning Rate	0.0005
Epoch	600
Batch Size	100
Optimizer	Adam
Dropout Rate	0.2

$$MSE(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{n}(y, \hat{y})^2 \tag{16}$$

$$CI = \frac{1}{Z}\sum_{\delta_i > \delta_j} h(b_i - b_j) \tag{17}$$

$$R_m^2 = R^2 \times \left(1 - \sqrt{R^2 - R_0^2}\right) \tag{18}$$

Where $\hat{y}_i$ represents the predicted value of the i^{th} sample, y_i is the corresponding true value, and N is the total number of drug-target pairs in the dataset. b_i is the predicted value corresponding to the largest δ_i, b_j is the predicted value corresponding to the smaller δ_j, and Z is a normalization constant. The function h typically represents a step function that determines whether the condition $\delta_i > \delta_j$ holds true.

As shown in Table 2, all experiments in this paper were conducted using NVIDIA RTX 3090 GPUs. The experimental framework was implemented in Python 3.9.20 with PyTorch 1.13.1+cu117. We employed the Adam optimizer with a batch size of 100, constrained by GPU memory limitations. The learning rate was set to 0.0005, and a dropout rate of 0.2 was applied. The model was trained for a total of 600 epochs.

To ensure the reliability of the reported results, all evaluations were conducted using five-fold cross-validation, with the final performance metrics averaged over these iterations.

Table 3
Comparisons with baseline models on the Davis dataset

Model	MSE ↓	CI ↑	R_m^2 ↑
KronRLS[38]	0.379	0.871	0.407
SimBoost[39]	0.282	0.872	0.644
DeepDTA[27]	0.261	0.878	0.631
DeepGS[30]	0.252	0.882	0.686
DeepNC[41]	0.233	0.789	0.653
MultiDTA[42]	0.231	0.893	0.694
GraphDTA[10]	0.229	0.893	-
LLMDTA[43]	0.226	0.884	0.717
FusionDTA[28]	0.208	0.749	0.743
MGraphDTA[40]	0.207	0.900	0.710
SMFF-DTA[44]	0.206	0.897	0.733
MAPGraphDTA[45]	0.203	0.900	0.692
DGraphDTA[14]	0.202	0.904	0.700
G-K-BertDTA[17]	0.201	0.912	-
RRGDTA[46]	0.196	0.909	0.749
Ours	**0.1688**	**0.9299**	**0.7865**

Table 4
Comparisons with baseline models on the KIBA dataset

Model	MSE ↓	CI ↑	R_m^2 ↑
KronRLS[38]	0.411	0.782	0.342
SimBoost[39]	0.222	0.836	0.629
FusionDTA[28]	0.208	0.749	0.793
DeepDTA[27]	0.194	0.863	0.673
DeepGS[30]	0.193	0.860	0.684
LLMDTA[43]	0.162	0.872	0.768
MultiDTA[42]	0.156	0.890	0.761
SMFF-DTA[44]	0.151	0.894	0.780
GraphDTA[10]	0.147	0.889	-
DeepNC[41]	0.133	0.897	0.695
MGraphDTA[40]	0.128	0.902	0.801
DGraphDTA[14]	0.126	0.904	0.786
MAPGraphDTA[45]	0.123	0.904	0.813
RRGDTA[46]	0.122	0.905	0.810
G-K-BertDTA[17]	0.121	**0.911**	-
Ours	**0.1196**	0.8984	**0.8228**

Additionally, to maintain experimental consistency and fairness, we trained the baseline models using their optimal hyperparameters and experimental settings as reported in their respective studies. Where available, we directly adopted the published results to mitigate discrepancies arising from implementation variations.

It is worth noting that slight differences between our results and those reported in previous studies may arise due to factors such as randomness in dataset partitioning and variations in hardware performance. However, the adoption of rigorous experimental protocols minimizes these potential sources of variation, ensuring the validity of our comparative analysis.

4.2. Quantitative Evaluation and Comparative Analysis

To validate the effectiveness of the proposed **MDM-DTA** model, we conduct a comprehensive comparison against several baseline methods on standard benchmark datasets. Specifically, we evaluate its performance relative to KronRLS [38], SimBoost [39], DeepDTA [27], DeepGS [30], GraphDTA [10], FusionDTA [28], DGraphDTA [14], G-K-BertDTA [17], MGraphDTA [40], DeepNC [41], Multi -DTA[42], LLMDTA[43], SMFF-DTA[44], MAPGraph-DTA[45] and RRGDTA[46]. Furthermore, for the extended dataset, we compare **MDM-DTA** with GraphDTA [10], MGraphDTA [40], and G-K-BertDTA [17].

According to the findings presented in Table 3, Table 4, and Table 5, both **MDM-DTA** and DeepDTA [27] exhibit substantial improvements over GraphDTA [10], underscoring the advantages of integrating one-dimensional sequence data, two-dimensional topological structures, and semantic features of drug-target pairs. Notably, **MDM-DTA** consistently outperforms GraphDTA [10] across different datasets. For instance, on the Davis dataset, the mean squared error (MSE) decreased by 26.29%, while the concordance index (CI) improved by 4.13%, reaching 0.1688 and 0.9299, respectively. In addition, the R_m^2 value increased to 0.7865, indicating a higher consistency between predicted and true binding affinities.Similarly, on the KIBA dataset, the MSE was reduced by 18.64%, the CI increased by 1.06%, and the R_m^2 value rose to 0.8228, demonstrating enhanced predictive accuracy and robustness.Moreover, when evaluated on the Metz dataset, MSE decreased by 17.8%, accompanied by a 3.2% improvement in CI, further supporting the model's generalization capability. These results suggest that the designed drug and protein feature extraction and fusion modules significantly enhance feature representation, thereby improving the predictive capability of drug-target binding affinity models.The correlation between the predicted values of **MDM-DTA** and the true affinities on the Davis dataset, KIBA dataset, and Metz dataset is shown in the scatter plots in Figures 7, 8, and 9, respectively, illustrating the strong alignment between predictions and true values.

In most cases, **MDM-DTA** achieves superior performance compared to baseline methods. A key factor contributing to this improvement is the ability of our model to effectively extract and integrate drug features, addressing the limitations observed in existing approaches. In particular, methods relying on simple concatenation for drug-target feature fusion fail to capture complex interactions adequately, thereby compromising predictive performance. To overcome this challenge, our model employs a hybrid expert mechanism for drug-target integration, allowing for a more sophisticated fusion of features. Compared to GraphDTA [10], which relies solely on convolutional neural networks (CNNs) [11] for protein feature extraction, our approach demonstrates notable performance gains by leveraging a richer set of molecular and protein characteristics.

Furthermore, across the three benchmark datasets, **MDM-DTA** consistently achieves lower MSE and higher CI scores

Table 5
Comparisons with baseline models on the Metz dataset

Model	MSE ↓	CI ↑
GraphDTA[10]	0.282	0.816
MGraphDTA[40]	0.265	0.822
G-K-BertDTA[17]	0.260	0.8286
Ours	**0.2318**	**0.8421**

than G-K-BertDTA [17], further demonstrating its robustness and superior predictive performance.By incorporating molecular map information and molecular descriptors, our model broadens the spectrum of drug-related features, enabling a more comprehensive characterization of drug-target interactions. Additionally, our hybrid expert mechanism facilitates the integration of molecular and protein semantic features into a unified representation space. In particular, by mapping protein features through DenseSENet [17], the model effectively assigns varying attention weights to different features, maximizing their contributions to the final prediction.

The experimental results highlight the effectiveness of **MDM-DTA** in drug-target affinity prediction tasks, demonstrating its ability to outperform existing state-of-the-art methods through enhanced feature extraction, integration, and fusion strategies.

5. Ablation Study

5.1. Impacts on Different Model Component

As shown in Table 6 and Figure 11, this paper investigates the synergistic effect of the message passing neural network (MPNN) [23], chemical descriptor (Des) [18], and mixture of experts (MoE) model [20] through a series of ablation experiments.The experimental results indicate that retaining only a single component—MPNN (w/o Des&MoE, MSE=0.6639), Des (w/o MPNN&MoE, MSE=0.6597), or MoE (w/o MPNN&Des, MSE=0.4558)—results in a significant increase in prediction error. This performance degradation can be attributed to the absence of complementary information, leading to limitations such as isolated topological modeling, oversimplified feature engineering, or insufficient input for dynamic fusion mechanisms.

In the two-component configurations, the results further highlight the necessity of integrating all three modules. The exclusion of MoE (w/o MoE, MSE=0.6456) reduces adaptability to heterogeneous data distributions and prevents dynamic expert weight allocation. Removing MPNN (w/o MPNN, MSE=0.4277) disrupts the molecular topological relationship, whereas eliminating Des (w/o Des, MSE=0.2921) introduces bias in physicochemical property predictions by neglecting domain-specific prior knowledge.

Ultimately, the complete model (Ours) integrates all three components, achieving the best performance (MSE=

Table 6
Ablation study for model components.

Methods	MPNN	Des	MoE	MSE ↓	CI ↑
w/o Des&MoE	✓	×	×	0.6639	0.7266
w/o MPNN&MoE	×	✓	×	0.6597	0.7198
w/o MPNN&Des	×	×	✓	0.4558	0.7924
w/o MoE	✓	✓	×	**0.6456**	**0.7431**
w/o MPNN	×	✓	✓	**0.4277**	**0.8058**
w/o Des	✓	×	✓	**0.2921**	**0.8695**
Ours	✓	✓	✓	**0.1967**	**0.8983**

0.1967, CI=0.8983). By leveraging MPNN to capture molecular topological structures, Des to incorporate domain-specific characteristics, and MoE to dynamically fuse multi-expert decision-making, the proposed model reduces MSE by 32.7% and improves CI by 3.3% compared to the optimal two-component configuration (w/o Des). These results clearly demonstrate the irreplaceable role and complementary advantages of these three components in the comprehensive modeling of molecular characteristics.

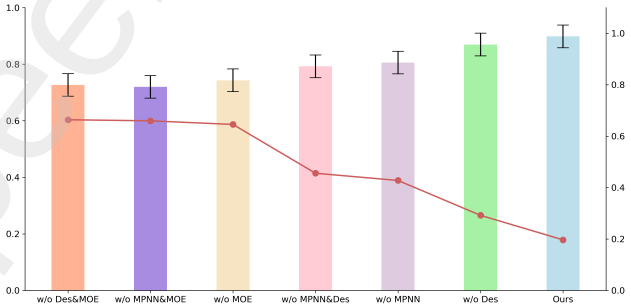


Figure 11: Ablation study on the Davis dataset. Bars represent CI values, and the line indicates MSE values.

5.2. Impact of Expert Activation on Model

To investigate the impact of the number of activated experts on the MoE [20] model, we conducted experiments by varying the number of activated experts and evaluating performance across multiple metrics. As shown in Table 7 and Figure 12, when only one expert was activated, the model exhibited relatively moderate performance, with an RMSE of 0.8105, MSE of 0.6569, Pearson correlation coefficient of 0.4183, and CI of 0.7301.

With the activation of two experts, the model achieved a substantial improvement in predictive accuracy. Specifically, RMSE decreased to 0.4436, MSE dropped significantly to 0.1967, Pearson correlation coefficient increased to 0.8677, Spearman correlation coefficient rose to 0.7266, and CI improved to 0.8983. These results suggest that utilizing multiple experts enables the model to capture a more diverse range of representations, thereby enhancing overall performance. This improvement can be attributed to the complementary nature of the experts—while one expert may focus on global interaction features, the other may specialize

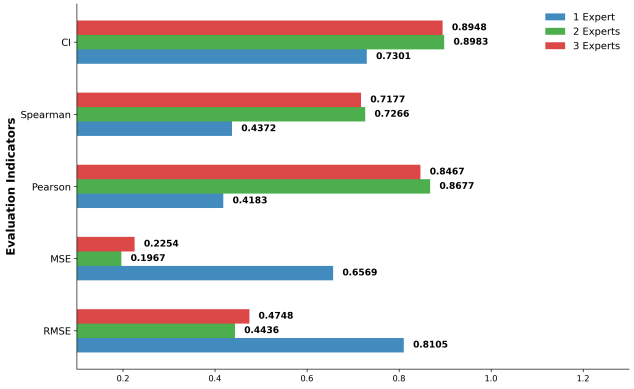


Figure 12: Comparison of the impact of activating different numbers of experts on model performance.

Table 7
Comparison of predictive performance with varying numbers of activated experts in the MoE model.

Number of Expert Activations	RMSE ↓	MSE ↓	Pearson ↑	Spearman ↓	CI ↑
1	0.8105	0.6569	0.4183	0.4372	0.7301
2	**0.4436**	**0.1967**	**0.8677**	**0.7266**	**0.8983**
3	0.4748	0.2254	0.8467	0.7177	0.8948

in domain-specific patterns. Such complementarity strengthens the model's generalization ability and leads to more accurate predictions.

However, when the number of activated experts was further increased to three, a slight decline in performance was observed, with RMSE increasing to 0.4748 and MSE rising to 0.2254 compared to the two-expert configuration. Although the Pearson correlation coefficient remained high at 0.8467 and CI at 0.8948, the performance gains became marginal. This decline may be due to the increased model complexity, which introduces risks such as overfitting. As more experts are activated, their contributions may become diluted, reducing their individual effectiveness in prediction.

These findings indicate that while activating multiple experts enhances model performance, there exists a point of diminishing returns. Beyond two experts, additional complexity does not proportionally improve predictive accuracy and may even hinder performance. Therefore, selecting an optimal number of experts is crucial to balancing model expressiveness and generalization ability.

5.3. Case Study

To evaluate the predictive performance of our model, we randomly selected five distinct drugs (D1–D5) and five different targets, as shown in Figure 13 and 14,and generated all pairwise combinations to predict their binding affinities. The resulting 5×5 affinity matrix is presented in Figure 15 using a heatmap, where color intensity reflects the strength of the predicted affinity. Notably, the combinations D2–T1 and D5–T1 exhibit the highest affinities, with scores of 0.96 and 0.98, respectively. In contrast, D3–T2, D5–T2, and D2–T5 demonstrate much lower affinities, with values as low as 0.08. These findings highlight the model's effectiveness in predicting binding strengths between novel drug–target pairs, which is essential for drug development and therapeutic discovery.

To further assess the model's predictive capability, we randomly selected 20 drug–target pairs from benchmark datasets and predicted their affinities using our model. The predicted values were then compared to the corresponding ground-truth affinities. As shown in Figure 16,positive residuals indicate that the model overestimates the true values, while negative residuals reflect underestimation. Notably, 90% of the samples (18 out of 20) exhibited prediction errors within a ±0.2 range, indicating consistently accurate and stable performance across diverse drug–target interactions. These findings underscore the robustness and generalizability of our model. Overall, our framework holds significant promise for accelerating novel drug discovery and drug repurposing.

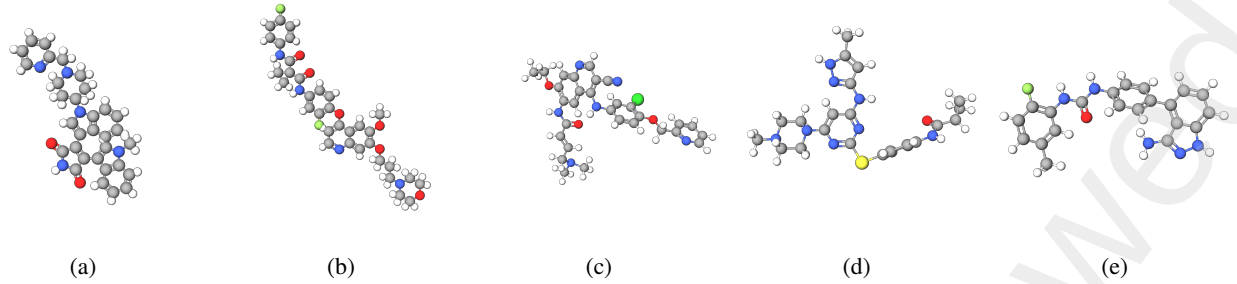
6. Discussion

Our proposed framework integrates MPNNConvNet, molecular description, Knowledge-BERT, DenseSEnet [17], and ESM2 [33], significantly improving the prediction accuracy of drug target binding affinity across different datasets. The combination of these modules allows the model to efficiently capture drug and protein signatures, resulting in a substantial increase in performance.

MPNNConvNet utilizes message passing neural networks (MPNN)[23] to capture the molecular graph structure, enabling the model to learn the inherent topological patterns of drug molecules. This approach is critical to overcoming the challenge of missing node labels, as it shifts the focus from relying on discrete semantic information to emphasizing the structural features of molecules. The molecular description encoder complements this by incorporating additional chemical descriptors, allowing the model to capture a wider range of chemical and structural details within the drug molecule. The synergy between the two drug encoders significantly improves the performance of drug target affinity prediction.

To further enhance the model, Knowledge-Bert infuses domain-specific knowledge by pre-training large-scale biomedical data. By encoding rich physicochemical properties and bioactivity data, Knowledge-BERT improves the model's understanding of complex structure-function relationships between drugs and targets. This wealth of knowledge allows models to more accurately predict drug-target interactions, leading to improved affinity predictions.

For protein coding, DenseSEnet [17] and ESM2 [33] provide powerful feature extraction capabilities. DenseSEnet [17] effectively captures the hierarchical features of proteins through dense linking and feature reuse, minimizing



(a) (b) (c) (d) (e)

Figure 13: Ball-and-stick representations of five randomly selected drug molecules. Atom colors are as follows: Blue for Nitrogen (N), Red for Oxygen (O), Gray for Carbon (C), White for Hydrogen (H), Green for Fluorine (F), and Yellow for Sulfur (S), where applicable.
(a) Structure: `Cn1cc(C2=C(c3cn(C4CCN(Cc5ccccn5)CC4)c4ccccc34)C(=O)NC2=O)c2ccccc21;`
(b) Structure: `COc1cc2c(Oc3ccc(NC(=O)C4(C(=O)Nc5ccc(F)cc5)CC4)cc3F)ccnc2cc1OCCCN1CCOCC1;`
(c) Structure: `CCOc1cc2ncc(C#N)c(Nc3ccc(OCc4ccccn4)c(Cl)c3)c2cc1NC(=O)C=CCN(C)C;`
(d) Structure: `Cc1cc(Nc2cc(N3CCN(C)CC3)nc(Sc3ccc(NC(=O)C4CC4)cc3)n2)n[nH]1;`
(e) Structure: `Cc1ccc(F)c(NC(=O)Nc2ccc(-c3cccc4[nH]nc(N)c34)cc2)c1.`

information loss between layers. This ensures comprehensive sequence coverage and improves the model's accuracy in predicting protein-drug interactions. Meanwhile, the Transformer-based protein sequence coding model ESM2 [33] further refines the protein coding process by identifying complex biological patterns in protein sequences. Together, these models improve protein coding and facilitate effective collaboration with the drug encoder module. By combining MPNNConvNet, molecular description, Knowledge-BERT, DenseSEnet [17], and ESM2 [33], our framework forms a robust, multi-perspective system that takes advantage of each component to achieve high-precision drug target affinity prediction.

However, despite significant improvements in this approach, there are still some limitations. While our approach

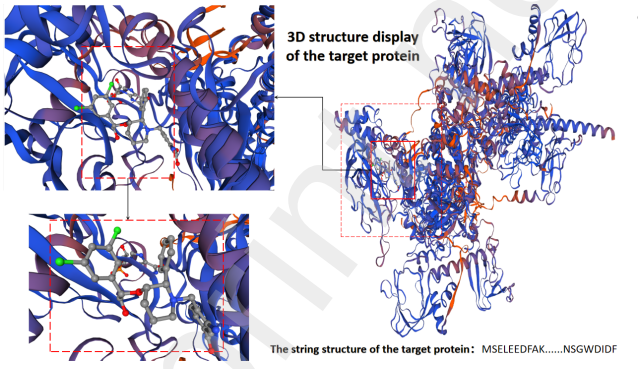


Figure 14: Illustration of the three-dimensional molecular structure of a representative target protein selected from the five randomly chosen targets. The structure highlights key interactions between the protein and its ligand, with specific focus on the binding site. The protein backbone is represented in a ribbon diagram, with different regions colored to distinguish between secondary structural elements such as alpha helices (blue) and beta sheets (orange). Key atoms within the binding site are shown as colored spheres: carbon (green), oxygen (red), and nitrogen (blue).

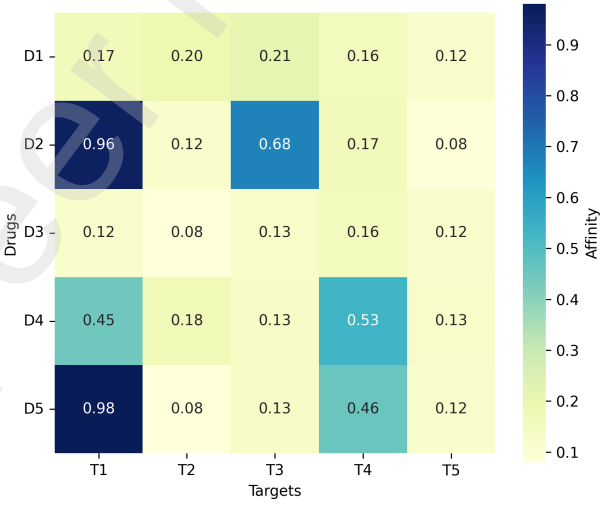


Figure 15: Heatmap visualizing predicted binding affinities between 5 selected drugs(D1–D5) and 5 selected targets (T1–T5). Color depth indicates strength of binding,with darker color denoting higher affinity.

efficiently extracts drug and protein information to minimize data loss, the binding between protein-drug complexes is essentially a three-dimensional physico-chemical process. Therefore, in addition to the one-dimensional sequences and two-dimensional topologies currently used, the three-dimensional structure combining drugs and proteins is a promising direction for future development. In addition, the lack of a clear pattern of affinity between drugs and targets suggests that iterative expansion of the dataset remains critical for future drug-target affinity prediction tasks. Future work should therefore focus on better integrating 3D structural data and improving models to capture more complex and diverse interaction patterns.

7. Conclusion

In this paper, we introduce **MDM-DTA**, a novel model designed to overcome the limitations of existing DTA prediction approaches. Our framework integrates multi-modal drug representations into a unified latent space, effectively capturing the complex interactions between drugs and targets. Additionally, we employ a hybrid expert mechanism to facilitate the fusion of semantic information, thereby improving predictive accuracy. To further enhance the prediction process, we introduce a correction method that addresses inconsistencies in the ordering of predicted values. Experimental results demonstrate that **MDM-DTA** significantly improves the accuracy of DTA prediction. In the future, we will explore the integration of multi-modal pretraining, dynamic expert networks, and 3D geometric deep learning to improve model interpretability, computational efficiency, and applicability in multi-pharmacological predictions.

Acknowledgement

This work is supported by Shanghai Municipal Natural Science Foundation (23ZR1425400).

Declaration of competing interest

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

Ethics Statement

This study did not involve any human participants, animal experiments, or personally identifiable information. All datasets used are publicly available from open-access sources and have been appropriately anonymized. Therefore, no ethical approval was required. The authors declare that there are no conflicts of interest.

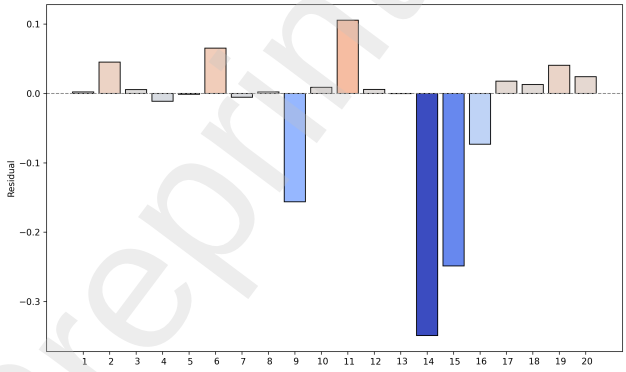


Figure 16: Demonstrating prediction accuracy for 20 drug-target pairs, showing minor deviations between predicted affinities and true affinities.The tightly clustered errors (mostly within ± 0.2 range)

References

[1] Joseph A DiMasi. Assessing pharmaceutical research and development costs. *JAMA internal medicine*, 178(4):587–587, 2018.

[2] Olivier J Wouters, Martin McKee, and Jeroen Luyten. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Jama*, 323(9):844–853, 2020.

[3] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. Innovation in the pharmaceutical industry: new estimates of r&d costs. *Journal of health economics*, 47:20–33, 2016.

[4] Asher Mullard. New drugs cost us $2.6 billion to develop. *Nature reviews drug discovery*, 13(12), 2014.

[5] John Arrowsmith. Phase ii failures: 2008-2010. *Nature reviews Drug discovery*, 10(5), 2011.

[6] Alan Talevi and Carolina L Bellera. Challenges and opportunities with drug repurposing: finding strategies to find alternative uses of therapeutics. *Expert Opinion on Drug Discovery*, 15(4):397–401, 2020.

[7] Yunan Luo, Xinbin Zhao, Jingtian Zhou, Jinglin Yang, Yanqing Zhang, Wenhua Kuang, Jian Peng, Ligong Chen, and Jianyang Zeng. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature communications*, 8(1):573, 2017.

[8] Yang Li, Guanyu Qiao, Keqi Wang, and Guohua Wang. Drug–target interaction predication via multi-channel graph neural networks. *Briefings in Bioinformatics*, 23(1):bbab346, 2022.

[9] Yang Yue and Shan He. Dti-hene: a novel method for drug-target interaction prediction based on heterogeneous network embedding. *BMC bioinformatics*, 22:1–20, 2021.

[10] 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.

[11] Yunhai Li, Pengpai Li, Duanchen Sun, and Zhi-Ping Liu. Predicting drug-target affinity using protein pocket and graph convolution network. In *International Symposium on Bioinformatics Research and Applications*, pages 1–12. Springer, 2024.

[12] Jooyong Shim, Zhen-Yu Hong, Insuk Sohn, and Changha Hwang. Prediction of drug–target binding affinity using similarity-based convolutional neural network. *Scientific Reports*, 11(1):4416, 2021.

[13] Lifan Chen, Xiaoqin Tan, Dingyan Wang, Feisheng Zhong, Xiaohong Liu, Tianbiao Yang, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36(16):4406–4414, 2020.

[14] Mingjian Jiang, Zhen Li, Shugang Zhang, Shuang Wang, Xiaofeng Wang, Qing Yuan, and Zhiqiang Wei. Drug–target affinity prediction using graph neural network and contact maps. *RSC advances*, 10(35):20701–20712, 2020.

[15] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[16] Rohit Singh, Samuel Sledzieski, Bryan Bryson, Lenore Cowen, and Bonnie Berger. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proceedings of the National Academy of Sciences*, 120(24):e2220778120, 2023.

[17] Xihe Qiu, Haoyu Wang, Xiaoyu Tan, and Zhijun Fang. Gk bertdta: a graph representation learning and semantic embedding-based framework for drug-target affinity prediction. *Computers in Biology and Medicine*, 173:108376, 2024.

[18] Li Fu, Shaohua Shi, Jiacai Yi, Ningning Wang, Yuanhang He, Zhenxing Wu, Jinfu Peng, Youchao Deng, Wenxuan Wang, Chengkun Wu, et al. Admetlab 3.0: an updated comprehensive online admet prediction platform enhanced with broader coverage, improved performance, api functionality and decision support. *Nucleic Acids Research*, page gkae236, 2024.

[19] 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.

[20] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[21] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[22] Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Smooth isotonic regression: a new method to calibrate predictive models. *AMIA Summits on Translational Science Proceedings*, 2011:16, 2011.

[23] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[24] Jingru Wang, Yihang Xiao, Xuequn Shang, and Jiajie Peng. Predicting drug–target binding affinity with cross-scale graph contrastive learning. *Briefings in Bioinformatics*, 25(1):bbad516, 2024.

[25] Maryam Bagherian, Elyas Sabeti, Kai Wang, Maureen A Sartor, Zaneta Nikolovska-Coleska, and Kayvan Najarian. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Briefings in bioinformatics*, 22(1):247–269, 2021.

[26] Hongjian Li, Kwong-Sak Leung, Man-Hon Wong, and Pedro J Ballester. Low-quality structural and interaction data improves binding affinity prediction via random forest. *Molecules*, 20(6):10947–10962, 2015.

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

[28] Weining Yuan, Guanxing Chen, and Calvin Yu-Chian Chen. Fusiondta: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. *Briefings in Bioinformatics*, 23(1):bbab506, 2022.

[29] 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):8326, 2023.

[30] Xuan Lin, Kaiqi Zhao, Tong Xiao, Zhe Quan, Zhi-Jie Wang, and Philip S Yu. Deepgs: Deep representation learning of graphs and sequences for drug-target binding affinity prediction. In *ECAI 2020*, pages 1301–1308. IOS Press, 2020.

[31] Mingjian Jiang, Shuang Wang, Shugang Zhang, Wei Zhou, Yuanyuan Zhang, and Zhen Li. Sequence-based drug-target affinity prediction using weighted graph neural networks. *BMC genomics*, 23(1):449, 2022.

[32] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[33] Brian Hie, Salvatore Candido, Zeming Lin, Ori Kabeli, Roshan Rao, Nikita Smetanin, Tom Sercu, and Alexander Rives. A high-level programming language for generative protein design. *BioRxiv*, pages 2022–12, 2022.

[34] Greg Landrum. Rdkit: Open-source cheminformatics. 2006. *Google Scholar*, 2006.

[35] 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.

[36] 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.

[37] 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.

[38] 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, 2015.

[39] Tong He, Marten Heidemeyer, Fuqiang Ban, Artem Cherkasov, and Martin Ester. Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal of cheminformatics*, 9:1–14, 2017.

[40] 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.

[41] Huu Ngoc Tran Tran, J Joshua Thomas, and Nurul Hashimah Ahamed Hassain Malim. Deepnc: a framework for drug-target interaction prediction with graph neural networks. *PeerJ*, 10:e13163, 2022.

[42] Jiejin Deng, Yijia Zhang, Yaohua Pan, Xiaobo Li, and Mingyu Lu. Multidta: drug-target binding affinity prediction via representation learning and graph convolutional neural networks. *International Journal of Machine Learning and Cybernetics*, 15(7):2709–2718, 2024.

[43] 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.

[44] Xun Wang, Zhijun Xia, Runqiu Feng, Tongyu Han, Hanyu Wang, Wenqian Yu, and Xingguang Wang. Smff-dta: using a sequential multi-feature fusion method with multiple attention mechanisms to predict drug-target binding affinity. *BMC biology*, 23(1):1–11, 2025.

[45] Shuo Hu, Jing Hu, Xiaolong Zhang, Shuting Jin, and Xin Xu. Drug target affinity prediction based on multi-scale gated power graph and multi-head linear attention mechanism. *PloS one*, 20(2):e0315718, 2025.

[46] Zhiqin Zhu, Yan Ding, Guanqiu Qi, Baisen Cong, Yuanyuan Li, Litao Bai, and Xinbo Gao. Drug–target affinity prediction using rotary encoding and information retention mechanisms. *Engineering Applications of Artificial Intelligence*, 147:110239, 2025.