

M^3 ST-DTI: A Multi-Task Learning Model for Drug-Target Interactions Based on Multi-Modal Features and Multi-Stage Alignment

Xiangyu Li
 College of Intelligence
 and Computing
 Tianjin University
 Tianjin, China
 Email: xiangyuli@tju.edu.cn

Ran Su
 College of Intelligence
 and Computing
 Tianjin University
 Tianjin, China
 Email: ran.su@tju.edu.cn

Liangliang Liu*
 College of Information
 and Management Science
 Henan Agricultural University
 Zhengzhou China
 Email: liangliu@henau.edu.cn

Abstract—Accurate prediction of drug-target interactions (DTI) is pivotal in drug discovery. However, existing approaches often fail to capture deep intra-modal feature interactions or achieve effective cross-modal alignment, limiting predictive performance and generalization. To address these challenges, we propose M^3 ST-DTI, a multi-task learning model that enables multi-stage integration and alignment of multi-modal features for DTI prediction. M^3 ST-DTI incorporates three types of features-textual, structural, and functional and enhances intra-modal representations using self-attention mechanisms and a hybrid pooling graph attention module. For early-stage feature alignment and fusion, the model integrates MCA with Gram loss as a structural constraint. In the later stage, a BCA module captures fine-grained interactions between drugs and targets within each modality, while a deep orthogonal fusion module mitigates feature redundancy. Extensive evaluations on benchmark datasets demonstrate that M^3 ST-DTI consistently outperforms state-of-the-art methods across diverse metrics.

Index Terms—Drug-Target Interaction, Multi-modal Feature Fusion, Multi-stage Alignment, Multi-task Learning

I. INTRODUCTION

In the study of modern medicine, accurately predict drug-target interaction (DTI) has become a key foundation of the drug discovery and development process of [1]–[3]. However, as a result of the experiment method of the nature of the labor-intensive, time-consuming and expensive, large-scale still challenging for identification of the new type of DTI [4]–[6]. Therefore, the development of efficient and accurate computational methods for DTI prediction has become a task of strategic importance [7].

In recent years, the method based on machine learning and deep learning in DTI prediction has attracted more and more attention [8]–[10]. These methods often use a type of input data, using machine learning, GNNs or based on sequence model (for example, CNN, RNN, transformer) to extract features and effectively modeling nonlinear relationship [11], [12]. For example, Lee et al. [13] proposed DeepConv-DTI, which uses 1D-CNN to extract protein features and ECFP to encode drug structures, followed by prediction using a fully connected network. Huang et al. [14] introduced

MolTrans to construct interpretable drug-target interaction maps to guide prediction. Zhao et al [15] HyperAttentionDTI was proposed, combined with super attention mechanism to enhance characteristic expression ability. Tsubaki et al. [16] used gnn for drug feature extraction and cnn for protein feature extraction. Cheng et al. [17] proposed IIFDTI, which combines cnn and GATs in an encoder-decoder framework for drug-protein representation learning. Despite their success, single-modal methods face significant challenges, with most existing methods relying on source domain priors (such as Word2Vec [18] or Graph neural networks [19]) to build embedding representations. These methods often fail to generalize to new structures and semantics. This leads to noticeable performance degradation in cross-domain and cold-start scenarios. The advent of bioomics has stimulated DTI research, leading to the development of methods that exploit multimodal data. Hu et al. [20] leverage 1D, 2D, and 3D embeddings of drugs and proteins under the UnitedDTA framework to obtain a unified and discriminative representation of cross-modal alignments. Hua et al [21] proposed MMDG-DTI, which fuses generalized textual and structural features through domain adversarial training (DAT) and contrastive learning to enhance generalization ability. However, Du et al. [22] points out that current DTI models still exploit the richness of multimodal data. The fusion strategy is still a key bottleneck: early fusion often ignores cross-modal interactions while building a comprehensive representation. Late fusion captures complementary signals, but may introduce redundancy and noise, affecting robustness and discrimination.

To address the challenges faced by DTI prediction, we propose M^3 ST-DTI, a multi-modal, multi-stage feature fusion framework based on multi-task deep learning. The model extracts and integrates information from textual, structural and functional modalities. The main contributions of this study are as follows:

- We propose M^3 ST-DTI, a multi-modal, multi-stage fusion, multi-task learning framework to construct textual,

structural, and functional features from standard drug-target interaction datasets, significantly improving prediction accuracy and generalization ability.

- First, early fusion is combined with Gram loss and multi-source cross-attention for semantic modality alignment, and deep orthogonal fusion is used to alleviate later redundancy.
- We use self-attention and hybrid pooling graph attention to handle fine-grained feature patterns, and bidirectional cross-attention and hybrid graph convolution to handle robust drug-target interactions.
- Extensive experiments verify the robustness of $M^3\text{ST}$ -DTI, providing interpretability for precision drug design.

II. MATERIALS AND METHODS

A. Benchmark Datasets

We used the public BindingDB dataset [23] to train and evaluate our framework. The BindingDB dataset comprises 10,665 drugs and 1,413 target proteins, with 32,601 interactions measured by dissociation constant (K_d). Among these, 9,166 interactions are labeled as positive and 23,435 as negative. In total, we obtained 32,601 samples from BindingDB. To ensure a fair and consistent evaluation, we adopted the same data splits (training, validation, and testing) as those used in previous studies [14], [24].

B. Multi-modal Feature Construction

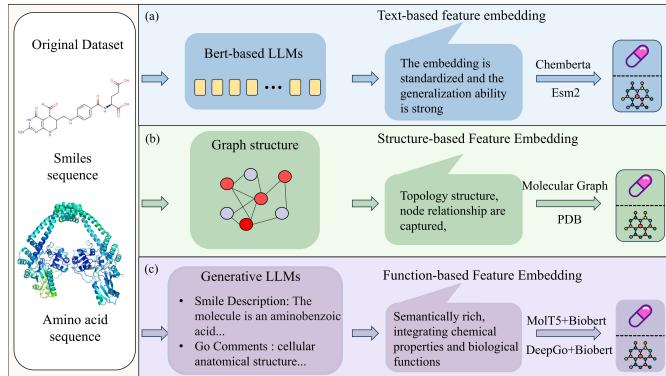


Fig. 1. The process of constructing multi-modal features from a standard DTI dataset.(a) the process of constructing textual features (b) the process of constructing structural features (c) the process of constructing functional featuresn.

The proposed $M^3\text{ST}$ -DTI framework constructs expressive and biologically meaningful representations for both drugs and protein targets by leveraging features from three complementary modalities: text, structure, and function, as shown in Fig. 1. Feature extraction for each modality is performed using pre-trained models tailored to preserve domain knowledge and enhance modality-specific representations.

Based on these multi-modal extracted features, We propose a multi-task learning deep learning framework based on multi-modal data and multi-stage feature fusion ($M^3\text{ST}$ -DTI), The process of $M^3\text{ST}$ -DTI framework as shown in Fig. 2

C. Multi-modal feature extraction stage

This section will introduce the screening process of text, structure and functional features. As shown in Fig. 2(a), each modal feature is processed independently to retain its domain-specific semantics.

1) Text Feature Extraction: As the foundational modality in DTI modeling, textual features describe drug SMILES strings and protein amino acid sequences with high-resolution semantic detail. To capture rich contextual dependencies, we apply a multi-head self-attention mechanism that enables each position in the sequence to attend to all others.

2) Structure Feature Extraction: Structural features are crucial in drug-target interaction prediction, capturing the graph topology of drug molecules and protein spatial conformations. We use graph-based structural modality and employ Graph Attention Networks (GATs) to dynamically weight nodes based on neighborhood relationships and edge attributes. To address limitations of global mean or max pooling, which may miss valuable information, we propose a Mixed Pooling Graph Attention (MPGAT) Module for comprehensive structural feature extraction from multiple perspectives.

3) Function Feature Extraction: Functional features in drug-target interaction prediction provide high-level semantic information regarding the biological processes, molecular functions, and cellular components associated with the molecules in biological systems. These features compensate for the limitations of structural and sequence features in capturing biological semantics. The feature extraction approach for the functional modality employed here is identical to the textual feature extraction method.

D. Multi-level Feature Fusion

1) Early Fusion Stage: As shown in Fig. 2(b), the early fusion strategy first integrates multiple modal features-textual, structural, and functional-of drugs and proteins jointly, producing unified feature representations that span all three modalities. Subsequently, interactions between drug and protein features are modeled. This early fusion approach enhances feature completeness and consistency by leveraging synergistic interactions both within and across modalities. To achieve effective alignment among the three modalities, we introduce a loss function based on the Gramian Representation Alignment Measure Loss (GRAM Loss) [25], which constrains the semantic closeness of different modal features in the embedding space. Set text, structure, and function of the modal characteristics of $x_{txt}, x_{graph}, x_{fun} \in \mathbb{R}^{Seq \times Dim}$, Gram Loss

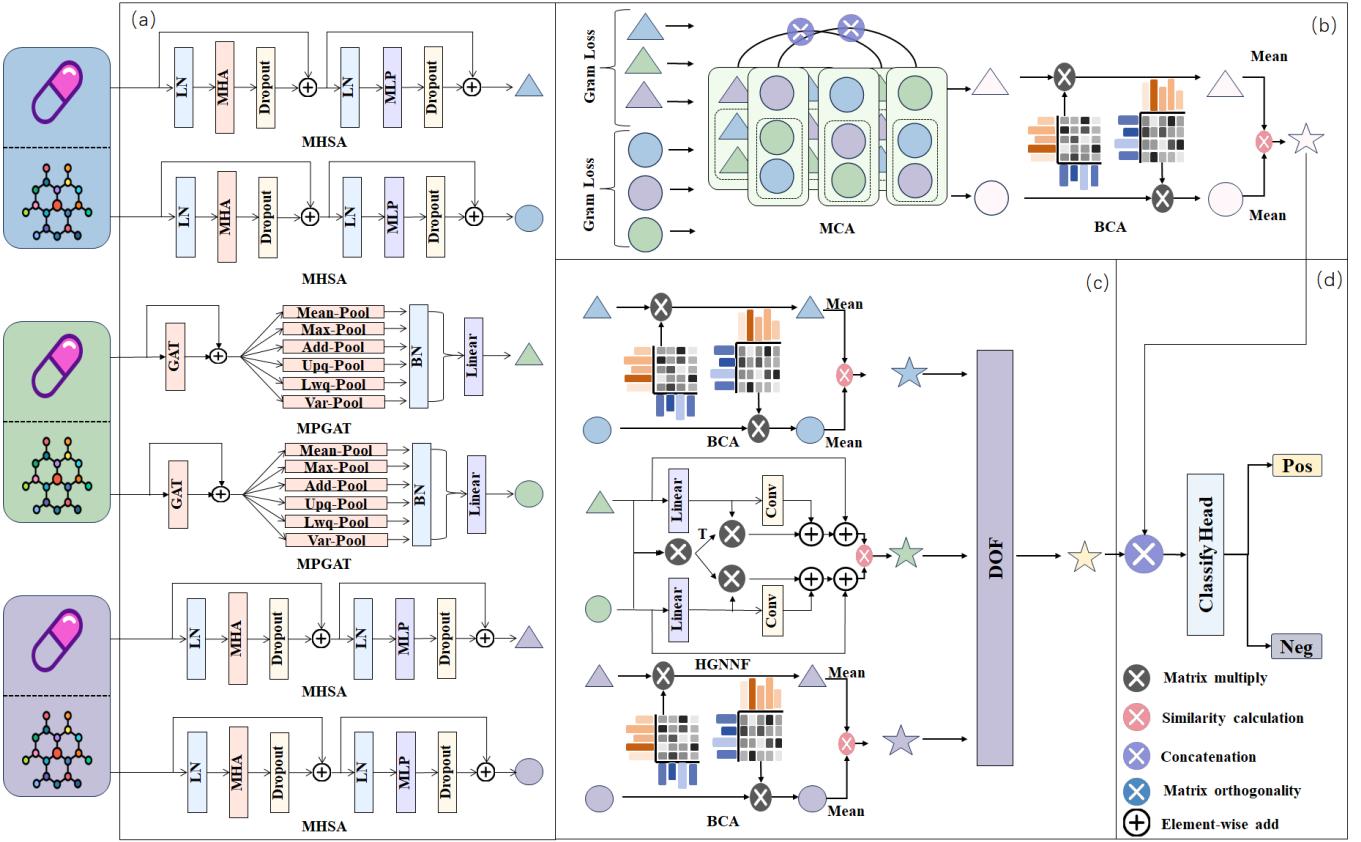


Fig. 2. Architecture flow of M^3 ST-DTI method: (a) multi-modal feature extraction stage (b) early fusion stage (c) late fusion stage (d) interaction prediction stage.

can be represented as:

$$x'_m = \frac{x'_m}{\sqrt{\sum_{j=1}^{\dim} (x'_{m,j})^2}}, \quad m \in \{\text{txt, graph, fun}\}$$

$$G = \begin{pmatrix} x'_{txt} \cdot x'_{txt} & x'_{txt} \cdot x'_{graph} & x'_{txt} \cdot x'_{fun} \\ x'_{graph} \cdot x'_{txt} & x'_{graph} \cdot x'_{graph} & x'_{graph} \cdot x'_{fun} \\ x'_{fun} \cdot x'_{txt} & x'_{fun} \cdot x'_{graph} & x'_{fun} \cdot x'_{fun} \end{pmatrix} \quad (1)$$

$$V = \sqrt{\det(G)} + \varepsilon$$

$$\text{GramLoss} = -\frac{1}{B} \sum_{i=1}^B \log \left(\frac{\exp(-V_i/\tau)}{\sum_{j=1}^k \exp(-V_j/\tau)} \right)$$

Where, $G \in \mathbb{R}^{Seq \times 3 \times 3}$ represents the feature similarity matrix, $\det(\cdot)$ represents the volume calculation function, and $v \in \mathbb{R}^{Seq \times 3 \times 3}$ represents the feature similarity matrix. $\varepsilon = 1 \times e^{-8}$ is a small constant to prevent numerical instability due to a determinant of zero, V represents the volume of the parallel volume formed by the features, and τ is a temperature parameter.

In early fusion stage, we utilize a Multi-source Cross-attention (MCA) module to dynamically capture fine-grained interactions among the textual, structural, and functional modality features. MCA allows one modality (text) to dynamically focus on the relevant information of other modalities

(structure, function), capturing the fine-grained interaction between modalities.

In addition, to ensure balanced representation of drug and protein features, we introduce a Bidirectional Cross-Attention (BCA) module that enables dynamic, two-way interaction between modalities. Unlike prior co-attention methods that risk overemphasizing drug features and weakening protein representations, our BCA module captures complex semantic associations through bidirectional information flow. This promotes semantic consistency and leads to comprehensive, unified embeddings for drug target interaction (DTI) prediction.

2) Late Fusion Stage: As shown in Fig. 2(c), in the M^3 ST-DTI framework, the core idea of the late fusion strategy is to perform deep intra-modal interactions. For each of the three modalities-textual, structural, and functional-we apply a BCA mechanism and a Hybrid Graph Convolutional Fusion (HGCNF) module to drug and protein features, respectively. These modules are designed to uncover hidden associations from multiple perspectives, yielding three sets of high-quality representations for drug-target interactions. Compared to the BCA module, the HGCNF module, tailored for the structural modality, integrates a 1D convolution operation with an interaction matrix to jointly capture local patterns and global interactions. The 1D convolution simulates neighbor-

hood dependencies by aggregating local features through a sliding window, while the interaction matrix models the global relationships between drug and protein features via learned weights. This hybrid architecture preserves the spatial structure of pooled features and simultaneously enhances global semantic alignment across modalities. It can be formulated as:

$$\begin{aligned} f_d &= \text{Linear}(d'_{graph}) \\ f_t &= \text{Linear}(t'_{graph}) \\ f'_d &= \text{GNN}(f_d) + \text{Softmax}(d'_{graph} \cdot (t'_{graph})^T) f_d + f_d \quad (2) \\ f'_t &= \text{GNN}(f_t) + \text{Softmax}(t'_{graph} \cdot (d'_{graph})) f_t + f_t \\ f_{graph} &= \text{Linear}(f'_d \| f'_t) \end{aligned}$$

Where $\text{Linear}(\cdot)$ denotes the Linear function, $\text{GNN}(\cdot)$ denotes the one-dimensional convolution function, $\text{Softmax}(\cdot)$ denotes the Softmax function, and $\|$ denotes the join operation.

However, as the dimensionality of features increases and redundant information accumulates after intermodal interactions, traditional linear fusion becomes inadequate for compactly compressing the representation space and may introduce distracting noise. To overcome this limitation, we introduce a Deep Orthogonal Fusion (DOF) module [26] at the late fusion stage (As shown in Fig. 2(c)), aiming to refine the integration of cross-modal representations. The DOF module takes the representations from the three modalities as input and constructs diversity-enhanced features through linear transformation and orthogonal projection, thereby emphasizing modal discrepancies. A masking mechanism is then applied to filter out redundant components and reconstruct the informative ones, ultimately enhancing the discriminability and compactness of the fused representation. The formulation is as follows:

$$\begin{aligned} f_{t,g} &= f_{txt} \| f_{graph} \\ f_{t,f} &= f_{txt} \| f_{fun} \\ f_{g,f} &= f_{graph} \| f_{fun} \\ x_{12} &= \left(f_{t,g} \| \left(f_{t,g} - \frac{f_{t,g} \cdot f_{t,f}}{\|f_{t,g}\|^2} f_{t,f} \right) \right) \\ x_{13} &= \left(f_{t,g} \| \left(f_{t,g} - \frac{f_{t,g} \cdot f_{g,f}}{\|f_{t,g}\|^2} f_{g,f} \right) \right) \quad (3) \\ x_{23} &= \left(f_{t,f} \| \left(f_{t,f} - \frac{f_{t,f} \cdot f_{g,f}}{\|f_{t,f}\|^2} f_{g,f} \right) \right) \\ x_{\text{Red}} &= \text{Linear}(x_{12} \| x_{13} \| x_{23}) \\ \text{mask}_m &= \mathbb{I}(\text{sim} > \text{threshold}) \end{aligned}$$

Where $f_{later} \in \mathbb{R}^{Seq \times Dim}$ is the drug-target interaction feature representation after multi-modal late fusion. $x_{12}, x_{13}, x_{23}, x_{\text{Red}}$ is the redundancy of the orthogonal calculation after said, $\mathbb{I}(\cdot)$ is the indicator function, threshold is a threshold, It is used to filter out the most similar targets, avoid redundancy and retain specific information. mask_m is to calculate the similarity between the redundant representation and the drug-target interaction feature representation of different modalities, and generate a binary mask based on the

similarity score. Like attention mechanism, we use key/value pair (x_{Red}, f_m) and query vector f_m to calculate the cosine similarity:

$$\begin{aligned} \text{sim}_m &= \text{Cos}(tK(x_{\text{Red}}), tQ(f_m)) \\ z_m &= \text{mask}_m \times tV(f_m), m = \text{txt/graph/fun} \quad (4) \\ f_{\text{later}} &= \text{Linear}(z_{\text{txt}} \| z_{\text{graph}} \| z_{\text{fun}} \| x_{\text{Red}}) \end{aligned}$$

Where $tQ(\cdot), tK(\cdot), tV(\cdot)$ denote the linear transformation. Guided by the estimated similarity sim, we filter out features that are relatively similar (high similarity scores) and only select features of a particular modality (low similarity scores) for the final fusion.

3) *Interaction Prediction*: The proposed $M^3\text{ST-DTI}$ method is a multi-task learning framework. We integrate early fusion features and late fusion features through linear connections:

$$f_{\text{output}} = \text{Linear}(f_{\text{early}} \| f_{\text{later}}) \quad (5)$$

We use different loss functions to train 6 branching tasks including: CE_{txt} , CE_{graph} , CE_{fun} , CE_{early} , CE_{later} , $\text{CE}_{\text{output}}$, so, our total classification loss:

$$\begin{aligned} L_c &= \frac{L_{\text{modal}} + L_{\text{stage}} + \text{CE}_{\text{output}}}{3} \\ L_{\text{modal}} &= \frac{\text{CE}_{\text{txt}} + \text{CE}_{\text{graph}} + \text{CE}_{\text{fun}}}{3} \\ L_{\text{stage}} &= \frac{\text{CE}_{\text{early}} + \text{CE}_{\text{later}}}{2} \end{aligned} \quad (6)$$

Our alignment loss Gram Loss includes drug feature alignment loss Gram_d and target feature alignment loss Gram_t . The total alignment loss is as follows:

$$L_g = (\text{Gram}_d + \text{Gram}_t)/2 \quad (7)$$

In the end, we obtain the final loss function as follows:

$$L = (L_c + \lambda L_g)/(1 + \lambda) \quad (8)$$

Where λ is the loss weight.

III. RESULTS

A. Experimental Settings

The initialization parameters of $M^3\text{ST-DTI}$ are configured as follows: the training batch size is set to 32; the Adam optimizer [27] is used with a learning rate of 1e-4 and a weight decay coefficient of 1e-5. The learning rate decays every 10 epochs. The maximum number of training epochs is 50. The hidden layer dimension is set to 512. The number of layers in the graph self-attention module is 3, and the number of layers in the multi-head self-attention module is 1. The loss weight is set to 1.0, and the temperature parameter is 0.1.

B. Ablation Experiment

To comprehensively assess the contribution of each module in the proposed M^3 ST-DTI framework, we performed ablation experiments on the BindingDB dataset. Specifically, we examined the roles of four core components: multi-modal feature extraction, early fusion, late fusion, and the interaction-aware prediction head. Ablation settings include three single-modality variants (text-only, graph-only, function-only), two fusion strategies (early-only, late-only), and the full M^3 ST-DTI model. The results of the ablation study are presented in Table I.

TABLE I
PERFORMANCE ON BINDINGDB DATASET (% METRICS)

Model	Accuracy	F1-Score	AUROC	AUPRC
Only txt	81.92	56.56	88.46	56.24
Only graph	80.00	55.37	89.59	55.27
Only fun	81.45	57.10	87.14	55.46
Only early	82.42	58.69	90.25	56.44
Only later	83.05	59.26	90.83	59.09
Ours	83.36	60.38	91.04	60.47

Multimodal fusion provides clear benefits for DTI prediction. On BindingDB, text modality (Only txt) performs the best among single modalities, but there is still a large room for improvement, confirming the necessity of cross-modal integration. The fusion strategy analysis shows that the only early variant improves over all unimodal baselines. However, its performance is lower than that of the full model. In contrast, the only later variant, with BCA and DOF modules, achieved a higher f1 value, confirming the benefit of deep interactive fusion. However, without early alignment, the overall performance is still suboptimal. The full M^3 ST-DTI framework, which integrates multimodal feature extraction, semantic alignment, front-back fusion, and interactive perceptual prediction, achieves the best results on all metrics. These results validate the effectiveness of the two-stage fusion design in capturing coarse-to-fine semantic relationships, highlighting the robustness of the multimodal, multi-stage approach.

C. Comparative Experiment

To comprehensively assess the effectiveness of M^3 ST-DTI, we conducted comparative experiments on the BindingDB dataset against several leading baseline models. As shown in Fig. 3, our model consistently achieved superior performance across multiple evaluation metrics, including Accuracy, F1-score, AUROC, and AUPRC, demonstrating its strength in both representation learning and multi-modal feature alignment.

On the BindingDB dataset, M^3 ST-DTI outperformed strong baselines such as MolTrans, MCANet, HyperAttention, and FMCA, achieving an Accuracy of 83.36%, an F1-score of 60.47%, and an AUROC of 91.04%. Notably, the model exhibited greater robustness in terms of F1-score, indicating enhanced generalization under imbalanced sample distributions. While MolTrans achieved a slightly higher AUPRC,

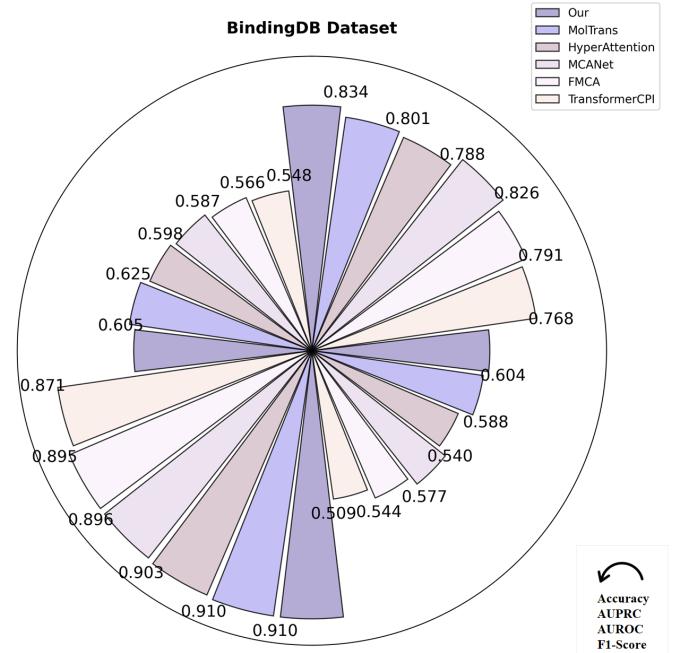


Fig. 3. Radial Plot of the Comparison Experiment Results

our model maintained the most balanced overall performance, effectively identifying positive interactions while preserving a high recall rate.

D. Visual Analysis

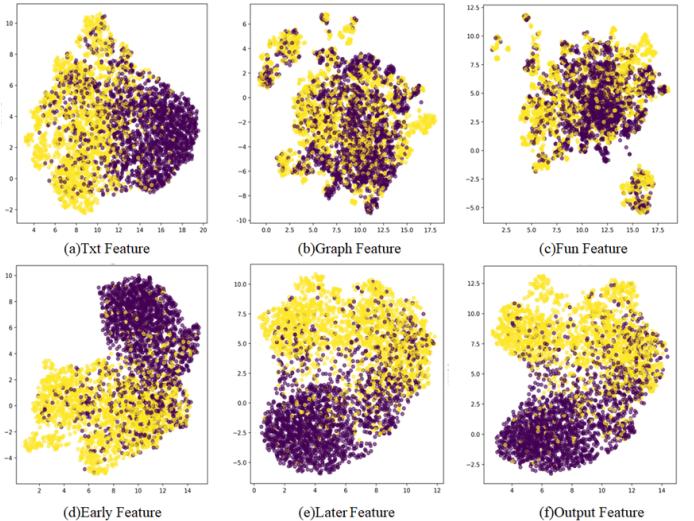


Fig. 4. UMAP Visualization of feature distributions

To gain a deeper understanding of the model performance in multi-modal feature extraction and fusion, the dimensionality reduction visualization of features from different modalities and fusion stages is performed based on umap. As shown in Fig. 4, in the early fusion stage (d), under the joint influence of MCA and Gram alignment loss, the embedding space

becomes more structured with obvious intra-class compactness and inter-class discreteness. In the late fusion (e), the addition of BCA and DOF modules further sharpens the class boundaries, and the concentration between negative samples increases. The final output feature (f) exhibits an optimal clustering structure where the boundary between positive and negative classes becomes clear and class overlap is greatly reduced. These results show that the model progressively optimizes the feature representation throughout the fusion stage. Our multi-stage fusion strategy effectively captures complex molecular-protein interaction patterns, continuously enhancing the discriminability from raw single-modal input to the final joint representation.

IV. CONCLUSION

This study introduces M^3 ST-DTI, a novel framework that integrates multi-modal feature representation with multistage alignment strategies, achieving marked advances in drug-target interaction (DTI) prediction. Experimental results on the BindingDB dataset strongly validate the superior performance of M^3 ST-DTI. Moreover, UMAP-based visualizations reveal a layer-wise progression in feature optimization from unimodal inputs to the final representations, showing increasingly distinct boundaries between positive and negative samples in the embedding space and significantly enhanced class separability. Future work will extend M^3 ST-DTI's functionality and explore its broader biomedical applications to enhance drug development and therapeutic research.

REFERENCES

- [1] X. Chen, C. C. Yan, X. Zhang, X. Zhang, F. Dai, J. Yin, and Y. Zhang, "Drug–target interaction prediction: databases, web servers and computational models," *Briefings in bioinformatics*, vol. 17, no. 4, pp. 696–712, 2016.
- [2] M. Wen, Z. Zhang, S. Niu, H. Sha, R. Yang, Y. Yun, and H. Lu, "Deep-learning-based drug–target interaction prediction," *Journal of proteome research*, vol. 16, no. 4, pp. 1401–1409, 2017.
- [3] R. Chen, X. Liu, S. Jin, J. Lin, and J. Liu, "Machine learning for drug–target interaction prediction," *Molecules*, vol. 23, no. 9, p. 2208, 2018.
- [4] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, "The price of innovation: new estimates of drug development costs," *Journal of health economics*, vol. 22, no. 2, pp. 151–185, 2003.
- [5] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, "How to improve r&d productivity: the pharmaceutical industry's grand challenge," *Nature reviews Drug discovery*, vol. 9, no. 3, pp. 203–214, 2010.
- [6] B. Xu, "Redefining the future of cancer care: Intelligent oncology unveiled," *Intelligent Oncology*, vol. 1, no. 1, pp. 31–33, 2025.
- [7] P. Zhang, Z. Wei, C. Che, and B. Jin, "Deepmgt-dti: Transformer network incorporating multilayer graph information for drug–target interaction prediction," *Computers in biology and medicine*, vol. 142, p. 105214, 2022.
- [8] Z. Yu, C. Xin, Y. Yu, J. Xia, and L. Han, "Ai dermatology: Reviewing the frontiers of skin cancer detection technologies," *Intelligent Oncology*, vol. 1, pp. 89–104, 2025.
- [9] K. Wang and M. Li, "Fusion-based deep learning architecture for detecting drug–target binding affinity using target and drug sequence and structure," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 12, pp. 6112–6120, 2023.
- [10] L. Liu, Z. Liu, J. Xie, H. Qiao, and J. Chang, "Transformers in pathological image analysis: A survey," *Engineering Applications of Artificial Intelligence*, vol. 156, p. 111114, 2025.
- [11] L. Liu, J. Cheng, Q. Quan, F.-X. Wu, Y.-P. Wang, and J. Wang, "A survey on u-shaped networks in medical image segmentations," *Neurocomputing*, vol. 409, pp. 244–258, 2020.
- [12] J. Gu, J. Zhang, S. Zeng, W. Zhang, R. Xia, X. Wang, Q. Zhou, S. Guo, H. Wang, and Z.-S. Chen, "Artificial intelligence in tumor drug resistance: Mechanisms and treatment prospects," *Intelligent Oncology*, vol. 1, pp. 73–88, 2025.
- [13] I. Lee, J. Keum, and H. Nam, "Deepconv-dti: Prediction of drug–target interactions via deep learning with convolution on protein sequences," *PLoS computational biology*, vol. 15, no. 6, p. e1007129, 2019.
- [14] K. Huang, C. Xiao, L. M. Glass, and J. Sun, "Moltrans: molecular interaction transformer for drug–target interaction prediction," *Bioinformatics*, vol. 37, no. 6, pp. 830–836, 2021.
- [15] Q. Zhao, H. Zhao, K. Zheng, and J. Wang, "Hyperattentiondti: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism," *Bioinformatics*, vol. 38, no. 3, pp. 655–662, 2022.
- [16] M. Tsubaki, K. Tomii, and J. Sese, "Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences," *Bioinformatics*, vol. 35, no. 2, pp. 309–318, 2019.
- [17] Z. Cheng, Q. Zhao, Y. Li, and J. Wang, "Tifdti: predicting drug–target interactions through interactive and independent features based on attention mechanism," *Bioinformatics*, vol. 38, no. 17, pp. 4153–4161, 2022.
- [18] L. Chen, X. Tan, D. Wang, F. Zhong, X. Liu, T. Yang, X. Luo, K. Chen, H. Jiang, and M. Zheng, "Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments," *Bioinformatics*, vol. 36, no. 16, pp. 4406–4414, 2020.
- [19] X.-b. Ye, Q. Guan, W. Luo, L. Fang, Z.-R. Lai, and J. Wang, "Molecular substructure graph attention network for molecular property identification in drug discovery," *Pattern Recognition*, vol. 128, p. 108659, 2022.
- [20] J. Hu, Y. Liu, X. Zeng, Q. Zou, R. Su, and L. Wei, "Multi-modal deep representation learning accurately identifies and interprets drug–target interactions," *IEEE Journal of Biomedical and Health Informatics*, 2025.
- [21] Y. Hua, Z. Feng, X. Song, X.-J. Wu, and J. Kittler, "Mmdg-dti: Drug–target interaction prediction via multimodal feature fusion and domain generalization," *Pattern Recognition*, vol. 157, p. 110887, 2025.
- [22] B.-X. Du, Y. Qin, Y.-F. Jiang, Y. Xu, S.-M. Yiu, H. Yu, and J.-Y. Shi, "Compound–protein interaction prediction by deep learning: databases, descriptors and models," *Drug discovery today*, vol. 27, no. 5, pp. 1350–1366, 2022.
- [23] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D198–D201, 2007.
- [24] H. Kang, S. Goo, H. Lee, J.-w. Chae, H.-y. Yun, and S. Jung, "Fine-tuning of bert model to accurately predict drug–target interactions," *Pharmaceutics*, vol. 14, no. 8, p. 1710, 2022.
- [25] G. Cicchetti, E. Grassucci, L. Sigillo, and D. Comminello, "Gramian multimodal representation learning and alignment," *arXiv preprint arXiv:2412.11959*, 2024.
- [26] R. Su, J. Xiao, H. Cui, P. Xuan, X. Feng, L. Wei, and Q. Jin, "Mskinet: Towards modality-specific knowledge interaction for glioma survival prediction," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024, pp. 2438–2445.
- [27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.