Bridging the Modality Reliability Gap in Drug-Target Interaction Prediction via a Confidence-aware Multimodal Fusion Framework

Jie Yang[*1,2], **Junxiong Zhang**[*3], **Kun Qian**[2], **Qingyu Yang**[3], **Weikai Li**[†4], **Zhen Cheng**[†2]

[1]School of Biomedical Engineering, ShanghaiTech University, Shanghai, 201210, China
[2]State Key Laboratory of Drug Research, Molecular Imaging Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China
[3]School of Information Science and Technology, ShanghaiTech University, Shanghai, 201210, China
[4]School of Computer and Artificial Intelligence, Shandong Jianzhu University, Shangdong, 250101, China
{yangjie2024, zhangjx2024, yangqy2023}@shanghaitech.edu.cn, leeweikai@outlook.com, {qiankun1, zcheng}@simm.ac.cn,

Abstract

With the rapid advancement of deep learning, drug target interaction (DTI) prediction has seen substantial performance enhancements. However, existing methodologies face a critical, yet unaddressed challenge, i.e., the **Modality Reliability Gap**. Such a gap arises from the unpredictable variance in the informativeness and reliability of 1D sequence versus 3D structural data across different drug-target pairs, critically limiting model robustness and domain generalization capabilities. To overcome it, we introduce **DrugCMF**, a novel **Drug**-Target interaction prediction method via **C**onfidence-aware **M**ultimodal **F**usion framework designed specifically to bridge the Modality Reliability Gap. Specifically, the DrugCMF employs a four-stage approach: (1) it extracts rich features by utilizing four pre-trained models to obtain token-level embeddings from both 1D sequences and 3D structures. (2) it preserves modality informativeness by independently learning interaction patterns within each modality through a Token-level Interaction module. (3) it explicitly quantifies the reliability gap by employing a novel confidence estimation mechanism to dynamically learn weights for each modality. (4) it bridges the gap by using these confidence scores to guide a learnable cross-modal fusion module, adaptively fusing information from the most trustworthy source. By methodically addressing the Modality Reliability Gap, DrugCMF significantly outperforms SOTA methods.

Code — https://github.com/deku-0621/DrugCMF

Introduction

Drug-Target Interaction (DTI) prediction is crucial in computational drug discovery (Zheng et al. 2020), as it identifies potential binding relationships between small molecules and protein targets. With the advent of deep learning techniques, numerous deep learning-based methods have emerged (Chen et al. 2020; Lu et al. 2022), enabling efficient DTI prediction and facilitating large-scale drug screening in a relatively short time. To better capture the intricacies of molecular binding patterns, researchers are increasingly using multimodal representations that combine 1D sequences and 3D structures (Luo et al. 2024; Lee et al. 2024).

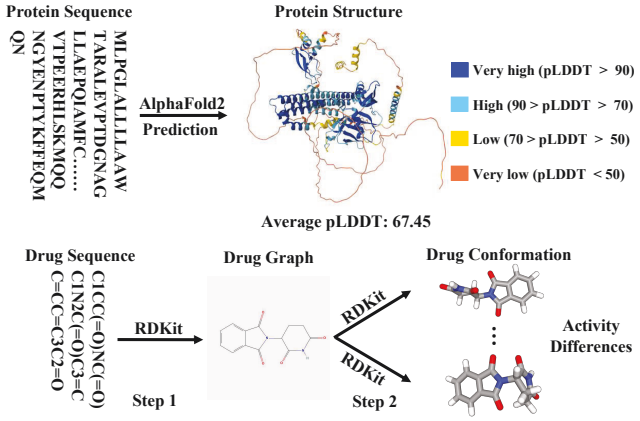

Figure 1: An example of the Modality Reliability Gap. (Top) AlphaFold2 predictions can be uncertain for some protein structures. (Bottom) Pharmacological activity varies among conformational isomers, which cannot be accurately represented by sequences information.

However, we identify a more fundamental challenge that current methods fail to resolve, which we term the ***Modality Reliability Gap***. This gap arises from unpredictable, sample-specific discrepancies in the informativeness and reliability between 1D and 3D modalities. For example, as shown in Figure 1 (Bottom), different conformations of a drug molecular generated via RDKit (Landrum 2013) may lead to entirely distinct binding modes and biological activities, while sequence information fails to accurately capture such differences. Furthermore, as depicted in Figure 1 (Top), protein structures predicted by AlphaFold2 (Jumper et al. 2021) from protein sequences still possess uncertainties in atomic accuracy and local details, consequently impacting the precision of protein binding pockets. Thus, this unpredictable gap necessitates that the model dynamically estimates which modality should be trusted more for each specific drug-target pair.

[*]These authors contributed equally.
[†]Corresponding authors.

Existing approaches struggle to address this challenge. Single modality models (Chen et al. 2020; Nguyen et al. 2021; Gao et al. 2023) attempt to avoid the modality gap by relying solely on one type of information. This inevitably leads to incomplete representations. Although recent multimodal approaches (Luo et al. 2024; Lee et al. 2024) have emerged, they typically adopt static or naive fusion strategies that fail to adapt to the modality reliability gap. These methods fail to balance inter-modal conflicts or effectively exploit valuable information, which restricts their generalization ability for novel drugs and targets.

To tackle this challenge, we propose **DrugCMF**, a confidence-aware multimodal fusion framework for drug–target interaction prediction, which consists of four key components: (i) We leverage four state-of-the-art pretrained models (Ross et al. 2022; Su et al. 2023; Zhou et al. 2023; Heinzinger et al. 2024) to comprehensively encode both one-dimensional sequence and three-dimensional structure information of drugs and proteins. (ii) A token-level bidirectional cross-attention mechanism is introduced to capture fine-grained binding interactions between drugs and targets. (iii) A modality confidence estimation module dynamically assesses the informativeness of each modality across diverse drug-target pairs, enabling dynamic fusion of reliable modalities while mitigating the impact of noisy ones. (iv) Finally, a learnable-query cross-modal fusion module is designed to extract complementary information across modalities, achieving more effective and robust multimodal integration.

For clarification, the contributions of our method can be summarized as follows:

- We have defined and resolved the Modality Reliability Gap in DTI research, a problem driven by the diverse informational value and trustworthiness of each modality.
- Our novel confidence-aware multimodal fusion framework dynamically evaluates and integrates modality information, bridging the modality gap to achieve superior prediction performance.
- Extensive experiments demonstrate that DrugCMF consistently outperforms SOTA baselines in both in-domain and out-of-distribution settings.

Related Work

Multimodal Fusion

Multimodal learning has emerged as a powerful paradigm for integrating complementary information across diverse data modalities, typically achieving superior performance compared to unimodal approaches (Guo, Wang, and Wang 2019; Xu, Zhu, and Clifton 2023). The core challenge lies in effective fusion strategies (Gadzicki, Khamsehashari, and Zetzsche 2020; Boulahia et al. 2021), which are commonly categorized into early fusion (Tian et al. 2018; Schlarmann et al. 2025), intermediate fusion (Fan et al. 2023; Gao et al. 2024), and late (decision) fusion (Wang et al. 2021; Liu et al. 2021). Early fusion integrates raw data through concatenation or addition (Liu et al. 2023; Schlarmann et al. 2025), but struggles with heterogeneous modalities and inconsistent feature scales. Intermediate fusion, supported by evidence from neuroscience (Macaluso 2006), enables cross-modal interaction at the feature level through complex architectures (Zhang et al. 2024b), while late fusion remains widely adopted for its simplicity and interpretability (Liu et al. 2021; Han et al. 2022a). Recent advances in dynamic fusion mechanisms, like uncertainty-based weighting (Li et al. 2022; Han et al. 2022a; Zhang et al. 2023), adaptively balance modality contributions, and have shown success in fields such as autonomous driving (Chen et al. 2025) and medical diagnosis (Lu et al. 2023).

Confidence and Uncertainty Estimation

Recent advances in uncertainty modeling have witnessed the emergence of various confidence-based approaches (Han et al. 2022a; Zheng et al. 2023; Zou et al. 2023; Zhang et al. 2023; Gao et al. 2024). Bayesian Neural Networks (BNNs) (Denker and LeCun 1990; Mackay 1992; Neal 2012) quantify model confidence through parameter distributions, while their improved variant MC-Dropout (Gal and Ghahramani 2015, 2016) employs stochastic dropout mechanisms for efficient confidence estimation. Within ensemble learning frameworks (Lakshminarayanan, Pritzel, and Blundell 2017; Havasi et al. 2020), predictive disagreement among multiple models serves as a confidence metric. Notably, Evidential Deep Learning (EDL) (Sensoy, Kaplan, and Kandemir 2018) directly models prediction confidence by constructing evidence space, circumventing complex sampling procedures. Recent studies have further explored confidence modeling through Dempster-Shafer theory (DST) (Han et al. 2021, 2022b) and energy scores (Liu et al. 2020), which employ distinct mathematical frameworks to quantify prediction reliability. Confidence-based methods demonstrate unique advantages in uncertainty estimation due to their directness and interpretability (Lu et al. 2024; Hu et al. 2025).

Drug-Target Interaction

With the advancement of deep learning techniques (LeCun, Bengio, and Hinton 2015), researchers have developed various innovative methods. Early representative works, such as DeepDTA (Öztürk, Özgür, and Ozkirimli 2018), employed CNNs to process SMILES strings (Weininger, Weininger, and Weininger 1989) and protein sequences, while DeepConv-DTI (Lee, Keum, and Nam 2019) combined molecular fingerprints with CNNs for prediction. Subsequent studies like GraphDTA (Nguyen et al. 2021) and MGraphDTA (Yang et al. 2022) introduced GNNs to model molecular graph structures, significantly improving feature extraction capabilities. In recent years, attention-based models such as TransformerCPI (Chen et al. 2020) and MolTrans (Huang et al. 2021) have further enhanced prediction accuracy by capturing long-range dependencies. Recent trends (Luo et al. 2024; Lee et al. 2024) indicate a growing use of multimodal information in DTI, but challenges persist in cross-modal interaction modeling, particularly concerning the integration of spatial conformations of drug-target complexes. These gaps highlight the need for reliability-aware models that our work directly addresses.

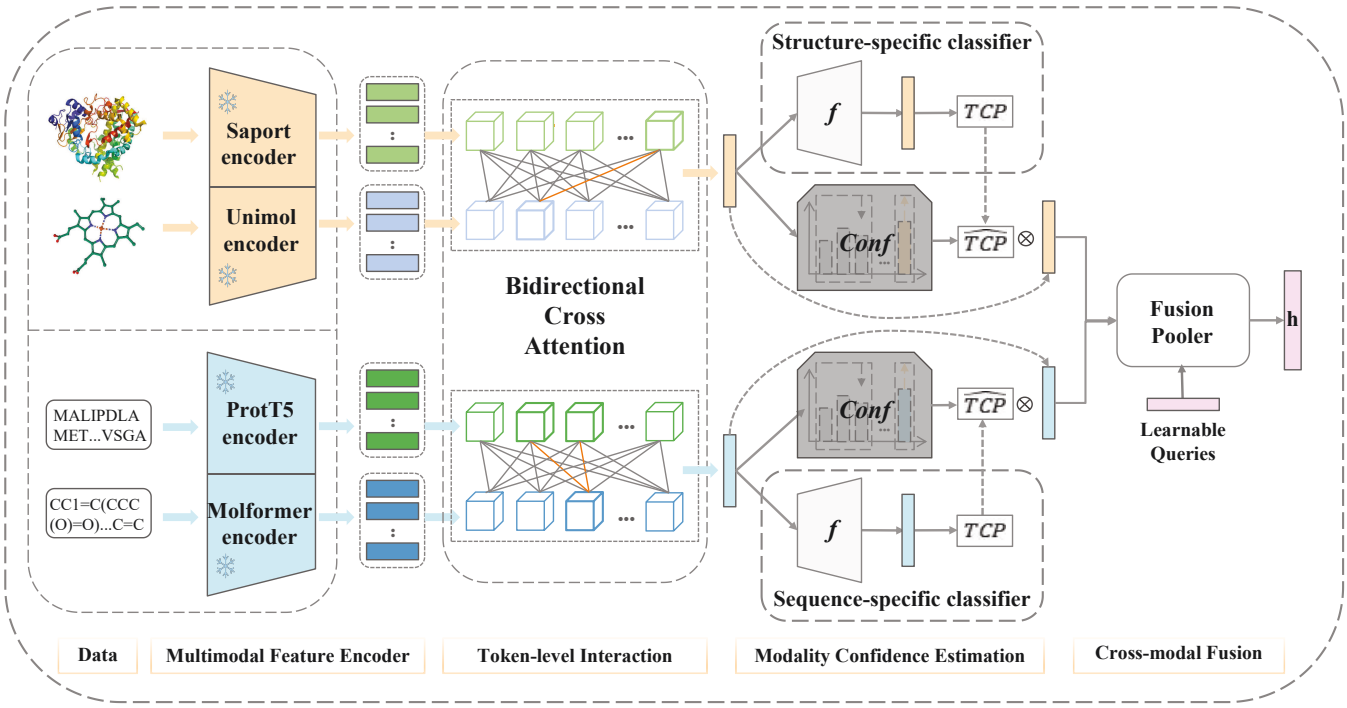

Figure 2: Our DrugCMF comprises four main stages: (i) A multimodal feature encoder utilizes four pre-trained models for fine-grained feature extraction from drug-target pairs. (ii) An intra-modal token-level bidirectional cross-attention module learns interaction patterns within each modality. (iii) A confidence regression network, f_{Conf}, estimates the predictive probability $\widehat{TCP}$ of a modality-specific classifier f for the true label, quantifying modality informativeness. (iv) A learnable cross-modal fusion mechanism effectively integrates information richness across diverse modalities.

Method

As shown in Figure 2, we detail the architecture of **DrugCMF**, our proposed framework designed specifically to address the **Modality Reliability Gap** introduced above. The framework operates in a sophisticated, multi-stage process: (1) extracting modality-specific representations, (2) capturing intra-modal interaction patterns, (3) quantifying the reliability of each modality for the given sample, and (4) adaptively fusing information guided by the learned reliability scores.

Formally, given a drug molecule represented by its **1D SMILES sequence** $\mathcal{D}_{1D} = (d_1, d_2, \ldots, d_{N_d})$ (d_i denotes chemical symbols) and **3D molecular conformation** $\mathcal{D}_{3D}$ together with a target protein represented by its **1D amino acid sequence** $\mathcal{T}_{1D} = (a_1, a_2, \ldots, a_{M_d})$ ($a_i \in \mathbb{A}_{23}$, $\mathbb{A}_{23}$ denotes the set of 23 amino acids) and **3D protein structure** $\mathcal{T}_{3D}$, DrugCMF learns a parameterized mapping function: $f_\theta : (\mathcal{D}_{1D}, \mathcal{D}_{3D}) \times (\mathcal{T}_{1D}, \mathcal{T}_{3D}) \to [0, 1]$ that yields the binding probability $P(y = 1 \mid \mathcal{D}_{1D}, \mathcal{D}_{3D}, \mathcal{T}_{1D}, \mathcal{T}_{3D})$, where $y \in \{0, 1\}$.

Multimodal Feature Encoder

To build a foundation for assessing modality reliability, we first need to extract comprehensive and high-quality features from both 1D and 3D data. Relying on a single modality, as in prior works (Xie, Tu, and Xu 2024; Bai et al. 2023), inherently limits the model's perspective. We employ four SOTA Pre-trained Language Models (PLMs): MolFormer (Ross et al. 2022) and ProtT5 (Heinzinger et al. 2024) encode the 1D sequences, while UniMol (Zhou et al. 2023) and SaProt (Su et al. 2023) embed their 3D counterparts. This strategy yields rich, token-level embeddings for each modality:

$$\begin{aligned}
\mathbf{E}^{1D}_{\text{drug}} &= \text{MolFormer encoder}(\mathcal{D}_{1D}) \in \mathbb{R}^{N_d \times d_{\text{mol}^{1D}}}, \\
\mathbf{E}^{1D}_{\text{prot}} &= \text{ProtT5 encoder}(\mathcal{T}_{1D}) \in \mathbb{R}^{N_p \times d_{\text{prot}^{1D}}}, \\
\mathbf{E}^{3D}_{\text{drug}} &= \text{UniMol encoder}(\mathcal{D}_{3D}) \in \mathbb{R}^{M_d \times d_{\text{mol}^{3D}}}, \\
\mathbf{E}^{3D}_{\text{prot}} &= \text{SaProt encoder}(\mathcal{T}_{3D}) \in \mathbb{R}^{M_p \times d_{\text{prot}^{3D}}}.
\end{aligned} \tag{1}$$

Token-level Interaction

Before assessing the reliability of each modality, it is crucial to allow each to form its own "view" of the drug-target interaction independently. A premature fusion would risk introducing noise from one modality to another. Therefore, we introduce a *Token-level Interaction (TLI)* module, shown in Figure 3, to model fine-grained binding patterns within each modality.

Modality-alignment Projection. Direct concatenation or early fusion of the representations risks information loss and bias amplification (Xie, Tu, and Xu 2024). Therefore, we linearly project all token embeddings into a shared d_{model}-

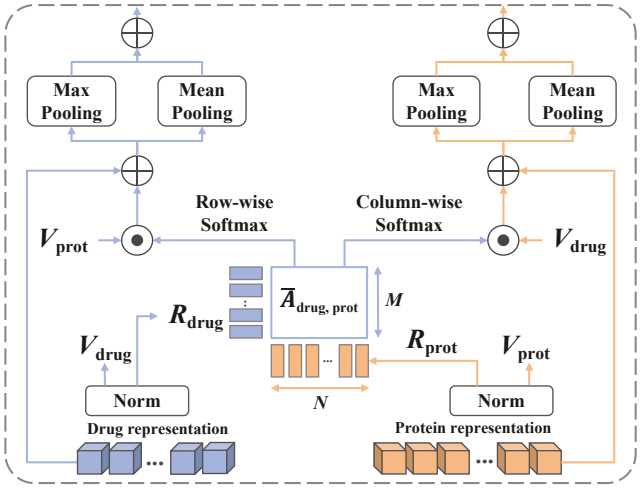

Figure 3: Token-level Interaction details.

dimensional space:

$$\mathbf{M}^k = \mathbf{E}^k_{\text{drug}}\,\mathbf{W}^k_d, \mathbf{P}^k = \mathbf{E}^k_{\text{prot}}\,\mathbf{W}^k_p, k \in \{1\text{D}, 3\text{D}\}. \quad (2)$$

where $\mathbf{W}^k_d \in \mathbb{R}^{d_{\text{mol}^k} \times d_{\text{model}}}$ and $\mathbf{W}^k_p \in \mathbb{R}^{d_{\text{prot}^k} \times d_{\text{model}}}$ are learnable projections.

Bidirectional Cross-Attention with Residuals. Unidirectional attention (e.g., protein-to-drug only) often amplifies the influence of dominant modalities, while suppressing complementary signals from less prominent modalities. This bias leads to incomplete or skewed interaction modeling. To address this, we adopt a bidirectional cross-attention (BiCA) mechanism inspired by BiXT (Hiller, Ehinger, and Drummond 2024), which enables mutual updates across modalities by enforcing symmetric interaction.

Take 1D as an example. As shown in Figure 3, given $R_{\text{drug}}, V_{\text{drug}} \in \mathbb{R}^{N_d \times d_{\text{model}}}$ and $R_{\text{prot}}, V_{\text{prot}} \in \mathbb{R}^{N_p \times d_{\text{model}}}$, BiCA first computes a shared, symmetric affinity matrix to measure pairwise similarities between drug and protein representations:

$$\text{A}_{\text{drug,prot}} = \frac{R_{\text{drug}} R^\top_{\text{prot}}}{\sqrt{d_{\text{model}}}}, \quad \text{A}_{\text{prot,drug}} = \text{A}^\top_{\text{drug,prot}}, \quad (3)$$

Row-wise and column-wise softmax operations are then applied to generate bidirectional attention maps, allowing both drug and protein features to attend to each other simultaneously. This symmetric design facilitates flexible and reciprocal information exchange:

$$\begin{aligned} \Delta^{\text{attn}}_{\text{drug}} &= \text{softmax}\left(\text{A}_{\text{drug,prot}}\right) V_{\text{prot}}, \\ \Delta^{\text{attn}}_{\text{prot}} &= \text{softmax}\left(\text{A}_{\text{prot,drug}}\right) V_{\text{drug}}, \end{aligned} \quad (4)$$

Finally, to retain the original modality-specific representations while integrating informative cross-modal signals, we apply residual connections:

$$\widetilde{\mathbf{M}}^k = \mathbf{M}^k + \Delta^{\text{attn}\,k}_{\text{drug}}, \quad \widetilde{\mathbf{P}}^k = \mathbf{P}^k + \Delta^{\text{attn}\,k}_{\text{prot}}. \quad (5)$$

And the updated drug and protein features are concatenated as

$$\mathbf{h}^{1\text{D}} = \left[\widetilde{\mathbf{M}}^{1\text{D}}\,;\,\widetilde{\mathbf{P}}^{1\text{D}}\right], \mathbf{h}^{3\text{D}} = \left[\widetilde{\mathbf{M}}^{3\text{D}}\,;\,\widetilde{\mathbf{P}}^{3\text{D}}\right]. \quad (6)$$

Modality Confidence Estimation

This section introduces the central component of our framework: a mechanism to explicitly quantify the Modality Reliability Gap for each drug-target pair. As argued in the introduction, simply fusing $\mathbf{h}^{1\text{D}}$ and $\mathbf{h}^{3\text{D}}$ is suboptimal, as one might be highly informative while the other is noisy or ambiguous. A robust model must dynamically assess which modality to trust more.

We propose that a modality's reliability is directly proportional to its ability to predict the correct outcome, which is fundamental for addressing the Modality Reliability Gap. As we will demonstrate quantitatively in our experiments section, this gap is significant and impacts model performance (see Figure 5). Based on this principle, we use the *True Class Probability* (TCP) (Corbière et al. 2019) as the core measure of modality confidence. TCP directly quantifies the probability that a modality-specific classifier assigns to the ground-truth label. A high TCP value indicates that the modality has captured predictive informativeness aligned with the ground-truth, thus serving as a reliable measure of its confidence.

TCP Classifier. For modality k, let $\mathbf{p}^k(\mathbf{y} \mid \mathbf{h}^k) \in [0,1]^C$ be the softmax-predicted class probabilities. Given the one-hot label $\mathbf{y}$, TCP^k can be computed as

$$TCP^k = \mathbf{y}^\top \mathbf{p}^k(\mathbf{y} \mid \mathbf{h}^k) = \sum_{c=1}^{\text{C}} y_c p^k_c. \quad (7)$$

Accurate TCP computation fundamentally depends on well-calibrated class probability estimations. We implement this through modality-specific classifiers $f^k_{\text{cls}} : \mathbf{h}^k \to \mathbf{y}$ that transform feature representations into class predictions. These independent classifiers undergo joint optimization during training to minimize the divergence between predicted and ground-truth distributions:

$$\mathcal{L}^{\text{TCP}}_{\text{CE}} = -\sum_{k \in \{1\text{D},3\text{D}\}} \mathbf{y} \log \mathbf{p}^k. \quad (8)$$

TCP Confidence Approximation. Since TCP computation requires inaccessible ground truth during inference, we design specialized confidence prediction networks $f^k_{\text{conf}} : \mathbf{h}^k \to TCP^k$ that operate solely on modality features. During training, these predictors minimize the mean squared error (MSE) between estimated and actual TCP values:

$$\widehat{TCP}^k = f^k_{\text{conf}}(\mathbf{h}^k), \quad (9)$$

$$\mathcal{L}^{\text{TCP}}_{\text{MSE}} = \sum_{k \in \{1\text{D},3\text{D}\}} \left(\widehat{TCP}^k - TCP^k\right)^2. \quad (10)$$

The output, $\widehat{TCP}^k$, is our learned, sample-specific confidence score that quantifies the reliability of modality k.

Cross-modal Fusion

Using the dynamic confidence scores $\widehat{TCP}^k$ from the previous step, we can now perform an adaptive fusion to bridge the Modality Reliability Gap. Instead of a static or naive combination, our fusion is explicitly guided by the learned reliability of each modality.

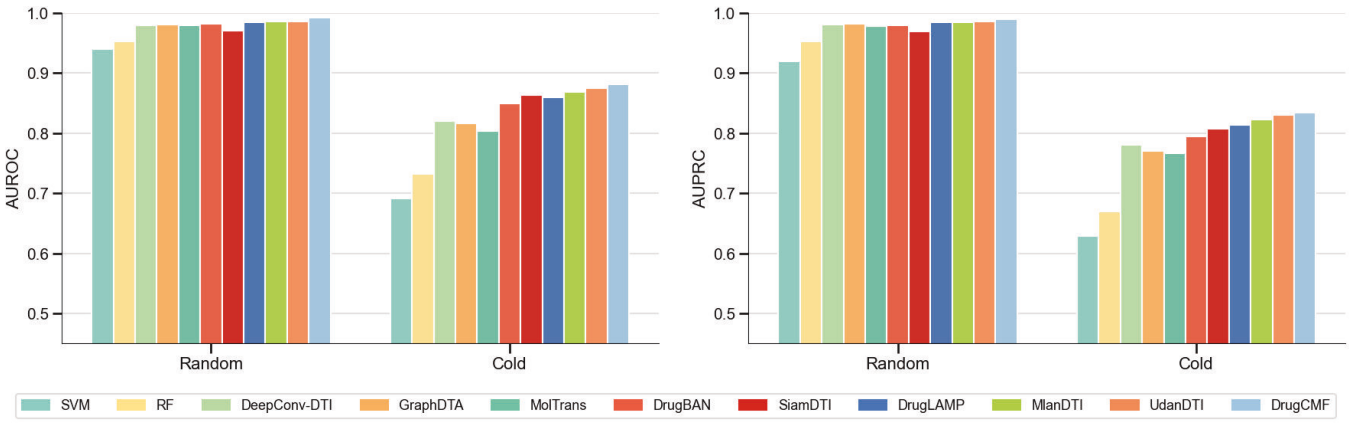

Figure 4: The in-domain performance comparison on Human dataset with random pair and cold pair split. The metrics shown in the figure represent the average values obtained from five experiments conducted with different random seeds.

Cross-Modal Fusion Encoder. Each modality feature is re-scaled by its predicted confidence score $\widehat{\mathrm{TCP}}^k$, such that less reliable modalities contribute proportionally less:

$$\mathbf{h}_{\mathrm{TCP}}^k = \mathbf{h}^k \otimes \widehat{\mathrm{TCP}}^k. \tag{11}$$

The confidence-weighted representations from both modalities are then concatenated $\mathbf{H}_{\mathrm{in}}^\top = \left[\mathbf{h}_{\mathrm{TCP}}^{1\mathrm{D}}; \mathbf{h}_{\mathrm{TCP}}^{3\mathrm{D}}\right]$, and fed into a multi-layer Transformer Encoder. Each layer consists of multi-head self-attention and feedforward network, following (Vaswani et al. 2017):

$$\mathbf{H} = \mathrm{TransformerEncoder}(\mathbf{H}_{\mathrm{in}}). \tag{12}$$

Attention Pooler. Rather than using simple averaging or max-pooling, which may weaken informative but sparse DTI patterns, we adopt learnable query-based attention pooling mechanism, a technique that has proven effective in prominent vision and language models for distilling salient information from long token sequences (Carion et al. 2020; Jaegle et al. 2021). The core idea is that these learnable queries Q act as trainable informativeness detectors. During training, they learn to actively seek out and aggregate the most task-relevant features from the fused representation H.

Specifically, we initialize a set of N_q query vectors with a standard Gaussian distribution:

$$\mathbf{Q} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \quad \sigma = \frac{1}{\sqrt{2 \times d_{\mathrm{model}}}}, \quad \mathbf{I} \in \mathbb{R}^{N_q \times d_{\mathrm{model}}}, \tag{13}$$

and aggregate the fused token representations via multi-head attention:

$$\mathbf{h} = \left[\mathrm{MHA}(\mathbf{Q}, \mathbf{H}, \mathbf{H})\right]_{0,:} \in \mathbb{R}^{2d_{\mathrm{model}}}. \tag{14}$$

where $[\cdot]_{0,:}$ denotes the first row of the resulting matrix. This approach allows the model to dynamically focus on and distill critical interaction features for final prediction.

Label Prediction and Unified Optimization. A separate classifier $f_{\mathrm{label}} : \mathbf{h} \to \mathbf{y}$ processes the fused representation $\mathbf{h}$ to predict the DTI label:

$$\hat{\mathbf{y}} = f_{\mathrm{label}}(\mathbf{h}), \tag{15}$$

$$\mathcal{L}_{\mathrm{CE}}^{\mathrm{Pred}} = -\mathbf{y} \log \hat{\mathbf{y}}. \tag{16}$$

The entire framework is optimized end-to-end with a composite loss:

$$\mathcal{L}_{\mathrm{Total}} = \lambda_{\mathrm{cls}} \mathcal{L}_{\mathrm{CE}}^{\mathrm{TCP}} + \lambda_{\mathrm{conf}} \mathcal{L}_{\mathrm{MSE}}^{\mathrm{TCP}} + \lambda_{\mathrm{label}} \mathcal{L}_{\mathrm{CE}}^{\mathrm{Pred}}. \tag{17}$$

Experiments

Experimental Setup

Datasets. We evaluate our method on three public DTI datasets: BindingDB (Gilson et al. 2016), BioSNAP (Zitnik, Sosic, and Leskovec 2018) and Human (Chen et al. 2020). Following DrugBAN (Bai et al. 2023), we use two different splitting strategies for in-domain and cross-domain scenarios. For in-domain evaluation, all datasets are split randomly into into 7:1:2 ratios for training, validation, and test sets. Furthermore, we conduct a cold-pair split experiment on the smaller Human dataset. For cross-domain evaluation, we adopt a clustering-based pair splitting strategy on the large-scale BindingDB and BioSNAP datasets. This ensured disjoint source and target domain data distributions, enabling more challenging evaluation. We employ AUROC and AUPRC as primary evaluation metrics, selecting models based on optimal validation AUROC.

Compared Methods. To investigate the improvement effect of our proposed method, we conduct a comprehensive comparison between DrugCMF and ten baseline methods. These models include traditional machine learning methods: SVM (Cortes and Vapnik 1995), RF (Ho 1995); advanced deep learning methods: DeepConv-DTI (Lee, Keum, and Nam 2019), GraphDTA (Nguyen et al. 2021), MolTrans (Huang et al. 2021), DrugBAN (Bai et al. 2023), and SiamDTI (Zhang et al. 2024a); and methods integrating pre-trained large language models: DrugLAMP (Luo et al. 2024), MlanDTI (Xie, Tu, and Xu 2024), and UdanDTI

Method	References	BindingDB			BioSNAP		
		AUROC ↑	AUPRC ↑	Accuracy ↑	AUROC ↑	AUPRC ↑	Accuracy ↑
SVM	Common	.939±.001	.928±.002	.825±.004	.862±.007	.864±.004	.777±.011
RF	Common	.942±.011	.921±.016	.880±.012	.860±.005	.886±.005	.804±.005
DeepConv-DTI	PLoS CB'19	.945±.002	.925±.005	.882±.007	.886±.006	.890±.006	.805±.009
GraphDTA	Bioinf.'21	.951±.002	.934±.002	.888±.005	.887±.008	.890±.007	.800±.007
MolTrans	Bioinf.'21	.952±.002	.936±.001	.887±.006	.895±.004	.897±.005	.825±.010
DrugBAN	NatMI'23	.960±.001	.948±.002	.904±.004	.903±.005	.902±.004	.834±.008
SiamDTI	arXiv'24	.961±.002	.945±.002	.890±.006	.912±.005	.910±.003	.855±.004
DrugLAMP	Bioinf'24	.923±.003	.927±.002	.857±.005	.917±.004	.922±.004	.851±.009
MlanDTI	AAAI'24	.919±.008	.896±.012	.828±.006	.903±.001	.908±.002	.823±.003
UdanDTI	TCBB'25	.965±.001	.955±.001	.911±.004	.941±.003	.942±.004	.876±.006
DrugCMF	Ours	**.968±.007**	**.960±.009**	**.915±.004**	**.948±.004**	**.950±.003**	**.884±.002**

Table 1: In-domain performance comparison of our model and baslines on BindingDB and BioSNAP (5 random runs). Best results are indicated by **bold**; second-best are underlined.

Method	References	DA Used	BindingDB		BioSNAP	
			AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
SVM	Common	✗	.490±.015	.460±.001	.602±.005	.528±.005
RF	Common	✗	.493±.021	.468±.023	.590±.015	.568±.018
DeepConv-DTI	PLoS CB'19	✗	.527±.038	.499±.035	.645±.022	.642±.032
GraphDTA	Bioinf.'21	✗	.536±.015	.496±.029	.618±.005	.618±.008
MolTrans	Bioinf.'21	✗	.554±.024	.511±.025	.621±.015	.608±.022
SiamDTI	arXiv'24	✗	.627±.027	.571±.024	.718±.055	.725±.054
DrugLAMP	Bioinf'24	✗	.650±.012	.588±.010	.739±.021	.768±.019
DrugBAN	NatMI'23	✓	.604±.027	.570±.047	.685±.044	.713±.041
MlanDTI	AAAI'24	✓	.666±.021	.596±.033	.722±.016	.751±.013
UdanDTI	TCBB'25	✓	**.713±.017**	**.671±.019**	.805±.011	.825±.008
DrugCMF	Ours	✗	.702±.011	.636±.013	**.822±.013**	**.830±.024**

Table 2: Cross-domain performance of our model and baselines on BindingDB and BioSNAP (5 random runs). Best results are indicated by **bold**; second-best are underlined. "DA Used" indicates whether domain adaptation was employed.

(Zhang, Ma, and Chen 2025). To respect the original authors, we directly adopt the reported metrics for baseline methods, except for MlanDTI (all datasets) and DrugLAMP (BindingDB dataset), which required re-evaluation due to differences in dataset partitioning and experiment setting. Our method is implemented in PyTorch using the AdamW optimizer with the learning rate of 0.0001.

Performance Evaluation

In-domain Performance. The performance of DrugCMF under random split settings is shown in Table 1 (comparative results on BindingDB and BioSNAP datasets), DrugCMF significantly outperform all baseline models across all metrics on both datasets. Notably, other PLM-based methods (DrugLAMP and MlanDTI) demonstrated particularly poor performance on these datasets, especially on BindingDB. We hypothesize that this may stem from their failure to effectively learn drug-target interaction patterns, further highlighting the superiority of our approach. Figure 4 (left) displays the in-domain evaluation results in the Human dataset. In random splits, our model achieved optimal AUROC and AUPRC scores. However, as pointed out in ref (Chen et al. 2020), this dataset carries potential ligand bias that may compromise real-world reliability. Therefore, we implement a cold-pair split strategy to mitigate the dependence of the model on known molecule features. The results in Figure 4 (right) show that while all models experienced significant performance degradation, DrugCMF maintained the best performance in all metrics.

Cross-domain Performance. In-domain classification with random splits fail to accurately reflect real-world application scenarios. Thus, we further evaluate performance on the more challenging cross-domain DTI prediction. Table 2 presents the performance evaluation under the cross-domain setting, showing a significant decline in performance across all models compared to the in-domain scenario. Nevertheless, the DrugCMF model demonstrates exceptional cross-domain prediction capabilities, outperforming the majority of SOTA models overall, particularly

without employing domain adaptation methods. On the BioSNAP dataset, it achieves the best metrics, with AUROC and AUPRC surpassing UdanDTI by 1.7% and 0.5%, respectively, highlighting its robustness in predicting novel drug-target interactions. While suboptimal on BindingDB dataset, DrugCMF still surpasses DrugBAN (with adversarial domain training) and MlanDTI (using semi-supervised pseudo-labeling). Specifically, DrugCMF's AUROC is 9.8% and 3.6% higher than DrugBAN and MlanDTI, respectively, validating its unique advantages in cross-domain prediction.

Method	Parameters ↓	FLOPs ↓
DrugLAMP	14.03M	11.26G
MlanDTI	2.87M	1.46G
UdanDTI	2.51M	1.59G
DrugBAN	1.07M	1.00G
DrugCMF	3.20M	0.86G

Table 3: Comparison of parameters and FLOPs.

Computational Efficiency

To validate practicality, we analyzed DrugCMF's computational cost in Table3. With 3.20M parameters and 0.86G FLOPs, DrugCMF reduces FLOPs by 14% versus the optimal baseline (DrugBAN, 1.00G FLOPs). This demonstrates that its performance stems from sophisticated fusion mechanisms rather than sheer scale, making it ideal for resource-sensitive large-scale drug screening.

Ablation Studies

We conduct cross-domain ablation studies on the BindingDB and BioSNAP datasets to analyze the effectiveness of each module, with detailed results in Table 4.

Ablation	BindingDB		BioSNAP	
	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
DrugCMF	0.702	0.636	0.822	0.830
1D Only	0.662	0.599	0.784	0.796
3D Only	0.687	0.611	0.793	0.814
w/o TLI	0.685	0.618	0.805	0.819
w/o MCE	0.682	0.613	0.809	0.811
w/o CMF	0.690	0.622	0.813	0.822

Table 4: Ablation study on BindingDB and BioSNAP datasets under cross-domain scenario.

Effectiveness of Multimodal Information. Experiments using only 1D or 3D data show a significant performance drop compared to DrugCMF. This confirms that both modalities contain complementary information and emphasizes the importance of a fusion framework capable of bridging the gap between them, rather than relying on one alone.

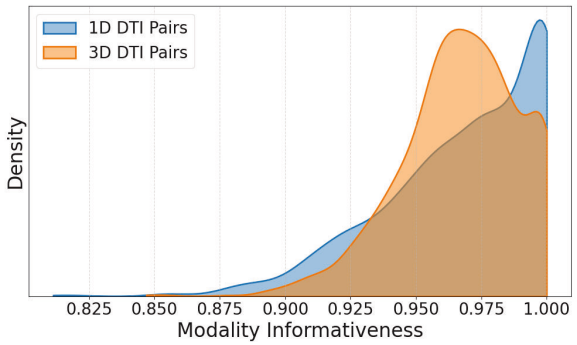

Figure 5: Density of Modality Confidence Distribution

Effectiveness of Token-level Interaction. Removing the Token-level Interaction (TLI) forces the model to fuse coarse feature representations. The result performance drop validates our claim that preserving modality-specific information through dedicated intra-modal interaction is a critical necessity for effective fusion.

Effectiveness of Modality Confidence Estimation. Removing the Modality Confidence Estimation (MCE) deprives the framework of an explicit reliability information, causing the model to return to naive fusion. This results in a significant performance decline, directly confirming that quantifying the reliability gap is essential for robustness.

Effectiveness of Cross-modal Fusion. Removing the Cross-modal Fusion (CMF) also degrades performance. This shows that even with a reliability score, a sophisticated, attention-based mechanism is needed to effectively use that score to dynamically bridge the gap and integrate features.

Visualization of the Modality Confidence

To empirically demonstrate the Modality Reliability Gap, we visualized predicted modality confidence distributions on Human test set. Using Kernel Density Estimation, Figure 5 plots the confidence distributions for the modalities. Results are clear: neither modality is universally reliable. While the 3D modality exhibits strong discriminative characteristics for most samples, some samples show a relatively low informational contribution from this modality. This visualization quantitatively validates our core hypothesis: a robust multi-modal DTI model should dynamically assess modality reliability on a per-sample basis to bridge this gap.

Conclusion

In this work, we tackle a key challenge in multimodal DTI prediction: Modality Reliability Gap. Our solution, DrugCMF, learns per-modality interaction features, estimates confidence scores for each, and then intelligently guides the fusion process. Experiments confirm that directly leveraging modality reliability yields significant performance gains over SOTA methods in both in-domain and cross-domain scenarios. We believe that DrugCMF represents a practical and important step for more reliable computational drug discovery.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (U2267221, 82530069, 82561160164, and 62306051), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB0830300), Gansu Science and Technology Major Project (23ZDFA014), State Key Laboratory of Drug Research (No. SIMM0120231004), and the Taishan Scholars Foundation of Shandong Province (tsqn202507225). Weikai Li and Zhen Cheng are co-corresponding authors.

References

Bai, P.; Miljković, F.; John, B.; and Lu, H. 2023. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence*, 5(2): 126–136.

Boulahia, S. Y.; Amamra, A.; Madi, M. R.; and Daikh, S. 2021. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6): 121.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, L.; Tan, X.; Wang, D.; Zhong, F.; Liu, X.; Yang, T.; Luo, X.; Chen, K.; Jiang, H.; and Zheng, M. 2020. TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36(16): 4406–4414.

Chen, L.; Wang, J.; Mortlock, T.; Khargonekar, P.; and Al Faruque, M. A. 2025. Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22306–22316.

Corbière, C.; Thome, N.; Bar-Hen, A.; Cord, M.; and Pérez, P. 2019. Addressing failure prediction by learning model confidence. *Advances in neural information processing systems*, 32.

Cortes, C.; and Vapnik, V. 1995. Support-vector networks. *Machine learning*, 20: 273–297.

Denker, J.; and LeCun, Y. 1990. Transforming neural-net output levels to probability distributions. *Advances in neural information processing systems*, 3.

Fan, Y.; Xu, W.; Wang, H.; Wang, J.; and Guo, S. 2023. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20029–20038.

Gadzicki, K.; Khamsehashari, R.; and Zetzsche, C. 2020. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd international conference on information fusion (FUSION)*, 1–6. IEEE.

Gal, Y.; and Ghahramani, Z. 2015. Bayesian convolutional neural networks with Bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Gao, B.; Qiang, B.; Tan, H.; Jia, Y.; Ren, M.; Lu, M.; Liu, J.; Ma, W.-Y.; and Lan, Y. 2023. Drugclip: Contrastive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems*, 36: 44595–44614.

Gao, Z.; Jiang, X.; Xu, X.; Shen, F.; Li, Y.; and Shen, H. T. 2024. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26876–26885.

Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; and Chong, J. 2016. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, 44(D1): D1045–D1053.

Guo, W.; Wang, J.; and Wang, S. 2019. Deep multimodal representation learning: A survey. *Ieee Access*, 7: 63373–63394.

Han, Z.; Yang, F.; Huang, J.; Zhang, C.; and Yao, J. 2022a. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20707–20717.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2021. Trusted Multi-View Classification. In *International Conference on Learning Representations*.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2022b. Trusted Multi-View Classification with Dynamic Evidential Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Havasi, M.; Jenatton, R.; Fort, S.; Liu, J. Z.; Snoek, J.; Lakshminarayanan, B.; Dai, A. M.; and Tran, D. 2020. Training independent subnetworks for robust prediction. *arXiv preprint arXiv:2010.06610*.

Heinzinger, M.; Weissenow, K.; Sanchez, J. G.; Henkel, A.; Mirdita, M.; Steinegger, M.; and Rost, B. 2024. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4): lqae150.

Hiller, M.; Ehinger, K. A.; and Drummond, T. 2024. Perceiving Longer Sequences With Bi-Directional Cross-Attention Transformers. *arXiv preprint arXiv:2402.12138*.

Ho, T. K. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, 278–282. IEEE.

Hu, X.; Duan, Z.; Chen, B.; and Zhou, M. 2025. Enhancing Uncertainty Estimation and Interpretability via Bayesian Non-negative Decision Layer. *arXiv preprint arXiv:2505.22199*.

Huang, K.; Xiao, C.; Glass, L. M.; and Sun, J. 2021. MolTrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6): 830–836.

Jaegle, A.; Gimeno, F.; Brock, A.; Vinyals, O.; Zisserman, A.; and Carreira, J. 2021. Perceiver: General perception with iterative attention. In *International conference on machine learning*, 4651–4664. PMLR.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873): 583–589.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Landrum, G. 2013. Rdkit documentation. *Release*, 1(1-79): 4.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.

Lee, I.; Keum, J.; and Nam, H. 2019. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15(6): e1007129.

Lee, J.; Jun, D. W.; Song, I.; and Kim, Y. 2024. DLM-DTI: a dual language model for the prediction of drug-target interaction with hint-based learning. *Journal of Cheminformatics*, 16(1): 14.

Li, B.; Han, Z.; Li, H.; Fu, H.; and Zhang, C. 2022. Trustworthy long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6970–6979.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33: 21464–21475.

Liu, X.; Liu, L.; Liao, Q.; Wang, S.; Zhang, Y.; Tu, W.; Tang, C.; Liu, J.; and Zhu, E. 2021. One pass late fusion multi-view clustering. In *International conference on machine learning*, 6850–6859. PMLR.

Liu, Z.; Xiong, C.; Lv, Y.; Liu, Z.; and Yu, G. 2023. Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval. In *Proceedings of ICLR*.

Lu, W.; Wu, Q.; Zhang, J.; Rao, J.; Li, C.; and Zheng, S. 2022. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in neural information processing systems*, 35: 7236–7249.

Lu, Y.; Chen, T.; Hao, N.; Van Rechem, C.; Chen, J.; and Fu, T. 2024. Uncertainty quantification and interpretability for clinical trial approval prediction. *Health Data Science*, 4: 0126.

Lu, Y.; Peng, R.; Dong, L.; Xia, K.; Wu, R.; Xu, S.; and Wang, J. 2023. Multiomics dynamic learning enables personalized diagnosis and prognosis for pancancer and cancer subtypes. *Briefings in bioinformatics*, 24(6): bbad378.

Luo, Z.; Wu, W.; Sun, Q.; and Wang, J. 2024. Accurate and transferable drug–target interaction prediction with DrugLAMP. *Bioinformatics*, 40(12): btae693.

Macaluso, E. 2006. Multisensory processing in sensory-specific cortical areas. *The neuroscientist*, 12(4): 327–338.

Mackay, D. J. C. 1992. *Bayesian methods for adaptive models*. California Institute of Technology.

Neal, R. M. 2012. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Nguyen, T.; Le, H.; Quinn, T. P.; Nguyen, T.; Le, T. D.; and Venkatesh, S. 2021. GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8): 1140–1147.

Öztürk, H.; Özgür, A.; and Ozkirimli, E. 2018. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17): i821–i829.

Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; and Das, P. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12): 1256–1264.

Schlarmann, C.; Croce, F.; Flammarion, N.; and Hein, M. 2025. FuseLIP: Multimodal Embeddings via Early Fusion of Discrete Tokens. *arXiv preprint arXiv:2506.03096*.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Su, J.; Han, C.; Zhou, Y.; Shan, J.; Zhou, X.; and Yuan, F. 2023. Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv*, 2023–10.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, 247–263.

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

Wang, T.; Shao, W.; Huang, Z.; Tang, H.; Zhang, J.; Ding, Z.; and Huang, K. 2021. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature communications*, 12(1): 3445.

Weininger, D.; Weininger, A.; and Weininger, J. L. 1989. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of chemical information and computer sciences*, 29(2): 97–101.

Xie, Z.; Tu, S.; and Xu, L. 2024. Multilevel attention network with semi-supervised domain adaptation for drug-target prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 329–337.

Xu, P.; Zhu, X.; and Clifton, D. A. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12113–12132.

Yang, Z.; Zhong, W.; Zhao, L.; and Chen, C. Y.-C. 2022. MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical science*, 13(3): 816–833.

Zhang, H.; Gong, X.; Pan, S.; Wu, J.; Du, B.; and Hu, W. 2024a. A Cross-Field Fusion Strategy for Drug-Target Interaction Prediction. *arXiv preprint arXiv:2405.14545*.

Zhang, K.; Luan, Y.; Hu, H.; Lee, K.; Qiao, S.; Chen, W.; Su, Y.; and Chang, M.-W. 2024b. MagicLens: self-supervised image retrieval with open-ended instructions. In *Proceedings of the 41st International Conference on Machine Learning*, 59403–59420.

Zhang, P.; Ma, J.; and Chen, T. 2025. Escaping the drug-bias trap: using debiasing design to improve interpretability and generalization of drug-target interaction prediction. *IEEE Transactions on Computational Biology and Bioinformatics*.

Zhang, Q.; Wu, H.; Zhang, C.; Hu, Q.; Fu, H.; Zhou, J. T.; and Peng, X. 2023. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, 41753–41769. PMLR.

Zheng, S.; Li, Y.; Chen, S.; Xu, J.; and Yang, Y. 2020. Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence*, 2(2): 134–140.

Zheng, X.; Tang, C.; Wan, Z.; Hu, C.; and Zhang, W. 2023. Multi-level confidence learning for trustworthy multimodal classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11381–11389.

Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; and Ke, G. 2023. Uni-mol: A universal 3d molecular representation learning framework.

Zitnik, M.; Sosic, R.; and Leskovec, J. 2018. BioSNAP Datasets: Stanford biomedical network dataset collection. *Note: http://snap. stanford. edu/biodata Cited by*, 5(1).

Zou, X.; Tang, C.; Zheng, X.; Li, Z.; He, X.; An, S.; and Liu, X. 2023. Dpnet: Dynamic poly-attention network for trustworthy multi-modal classification. In *Proceedings of the 31st ACM international conference on multimedia*, 3550–3559.