FuseMine: Robust Multi-Modal Compound-Protein Interaction Prediction via Differential Attention Feature Mining

Junlin Xu[1], Zhuang Zhang[2], Zhenghang Gong[2], Jincan Li[3], Pan Zeng[4], Zilong Zhang[5], Xiong Li[6], Shuting Jin[1*], Haowen Chen[7*], Yajie Meng[2*]

[1]Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, Hubei 430065, China
[2]School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, Hubei 430200, China
[3]School of Mathematics and Statistics, Hainan Normal University, Haikou, Hainan 570228, China
[4]School of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China
[5]School of Computer Science and Technology, Hainan University, Haikou, Hainan 570228, China
[6] School of Information and Software Engineering, East China Jiaotong University, Nanchang 330013, China
[7]College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, 410082, China
xjl@hnu.edu.cn, 2315363089@wtu.edu.cn, 2415283013@wtu.edu.cn, ljc682@hainnu.edu.cn, zengpan@cqnu.edu.cn, zhangzilong@hainanu.edu.cn, lixiong@ecjtu.edu.cn, shutingjin@wust.edu.cn, hwchen@hnu.edu.cn, myj@hnu.edu.cn

Abstract

Accurate prediction of compound protein interactions (CPIs) is crucial for drug discovery. However, existing deep learning-based methods suffer from hidden biases and poor cross-domain generalization, leading to spurious correlations and inadequate representation of unseen compound-protein pairs. To address these limitations, we propose FuseMine, a multimodal deep learning framework that jointly leverages molecular structures and biological sequences for reliable CPI prediction. Specifically, FuseMine adopts a dual-representation strategy for each molecule. It employs a convolutional encoder to capture structural features, combined with pretrained large language models for extracting semantic information from sequences. We propose a novel Multimodal Feature Orchestration Aggregation (MFOA) module that enables deep and synergistic fusion between the structural features and the sequential semantics of molecules, effectively capturing the complementary patterns across modalities. Additionally, we design a Reduction Differential Feature Mining (RDFM) module to further enhance the representation of discriminative features, thereby improving the model's generalization capability. Extensive experiments on multiple benchmark datasets demonstrate that our framework consistently outperforms state-of-the-art methods in both intra-domain and cross-domain scenarios. These results highlight the synergistic value of combining structural and sequential data for CPIs.

Code — https://github.com/Biowust/FuseMine.

Introduction

Drug development is a protracted, capital-intensive process with characteristically high failure rates (Berdigaliyev and Aljofan 2020). Each year, only a limited number of first-in-class drugs receive regulatory approval, while potential compound-protein combinations constitute a virtually infinite exploration space. Conventional experimental screening methodologies are inadequate for addressing such vast scales and require substantial financial and temporal investments. Against this backdrop, computational prediction of compound protein interactions (CPIs) has emerged as a promising approach. Through virtual screening technologies, this methodology can rapidly identify promising candidates from extensive molecular libraries, thereby accelerating drug discovery, mitigating developmental risks, and providing insights into molecular mechanisms of action.

The evolution of computational methods has witnessed a paradigm shift from traditional machine learning to deep learning approaches (Zhao et al. 2022a). Traditional machine learning models relied on hand-crafted features, which exhibited limited capability in capturing complex, non-linear interactions despite their partial effectiveness (Nagamine and Sakakibara 2007; Peska, Buza, and Koller 2017; Madhukar et al. 2019). Deep learning has provided a significant breakthrough by enabling automatic feature extraction. Researchers have employed various neural network architectures to model compound-protein interactions. For example, Convolutional Neural Networks (CNNs) have been applied to both SMILES representations and protein sequences to capture local patterns (Öztürk, Özgür, and Ozkirimli 2018; Lee, Keum, and Nam 2019), Graph Neural Networks (GNNs) are commonly used to represent molecular graphs (Nguyen et al. 2021; Bai et al. 2023), and Transformer-based models leverage attention mechanisms to encode contextual relationships in sequences or graphs (Zhao et al. 2022b; Huang et al. 2021). These methods have substantially improved prediction performance through automated feature learning. Recently, pre-trained Large Language Models (LLMs) trained on extensive biological datasets have been introduced to the field, offering novel perspectives for understanding compound-protein interactions through their sophisticated sequence representation capabilities (Xie, Tu, and Xu 2024). These advancements, particularly the integration of LLMs, have significantly improved the semantic understanding of biological sequences and enabled richer molecular representations.

*Corresponding authors

However, despite these promising breakthroughs, CPI prediction effectiveness remains constrained by several key limitations.

A major limitation stems from the scarcity of high-quality labeled data, which restricts the training capacity of deep models and undermines their generalization. Without sufficient data, models often overfit to statistical artifacts rather than capturing biologically meaningful patterns, leading to degraded performance on novel compounds or targets (Chen et al. 2020). To address this challenge, recent research has focused on multi-modal integration, particularly leveraging LLMs to enrich semantic representations. However, most current fusion approaches rely on late-stage strategies, wherein compound and protein modalities are processed separately and combined via simple concatenation (Sun et al. 2024). Such designs often fail to capture deep cross-modal synergies, thereby limiting joint representational power. Developing effective deep fusion mechanisms thus remains a critical open problem.

To address these challenges, we propose FuseMine, a deep learning framework that learns joint representations from multiple data modalities. FuseMine combines a Graph Neural Network for encoding molecular graph structures, a Convolutional Neural Network for extracting local patterns from protein sequences, and pre-trained Large Language Models for incorporating global semantic information derived from large-scale biological data. Building upon this architecture, we design two novel modules to enhance feature integration and improve the model's generalization capability. The Multi-modal Feature Orchestration and Aggregation (MFOA) module systematically orchestrates and aggregates complementary features from structure and sequence. Concurrently, the Reduction Differential Feature Mining (RDFM) module is introduced to mine discriminative and generalizable features, which improves the modeling of fundamental interaction mechanisms and enhances predictive robustness for unseen compounds and targets. Experimental evaluations on public benchmarks demonstrate that FuseMine achieves state-of-the-art performance. Furthermore, the model excels in challenging cross-domain tests, showcasing its superior generalization and affirming its potential for real-world drug discovery applications.

Our contributions are summarized as follows:

- We design the innovative MFOA module, which effectively fuses multi-modal information, overcoming the limitations of vanilla fusion approaches.
- We introduce the RDFM module to enhance model generalization, significantly improving the predictive robustness for unseen compound-protein pairs and cross-domain scenarios.
- We conduct extensive experiments on four benchmark datasets to evaluate the effectiveness and generalization ability of our model.

Related Work

Compound-Protein Interaction Prediction

Machine learning methods have been instrumental in computational drug discovery for decades. Traditional approaches leverage quantitative structure-activity relationships, protein stoichiometry, and molecular docking to model CPIs through classification algorithms including support vector machines, random forests, and artificial neural networks (Barkat, Moussa, and Badr 2021). While these methods demonstrate effectiveness on well-characterized datasets, they often struggle with generalization to novel compound-protein pairs and exhibit limited scalability.

The emergence of deep learning, with its enhanced representation learning capabilities, has significantly advanced CPI prediction. Current methodologies can be broadly grouped by their input modalities and architectural designs.

One prominent paradigm centers on sequence-based methods, which encode compounds as SMILES strings and proteins as amino acid sequences. DeepDTA (Öztürk, Özgür, and Ozkirimli 2018) pioneered this line of work by employing dual 1D-CNNs to extract features from raw sequences. To better capture protein motifs at varying scales, DeepConv-DTI (Lee, Keum, and Nam 2019) advanced this architecture by introducing variable-length convolution windows, which improved the detection of binding sites.

To model long-range dependencies and complex interaction patterns more effectively, another stream of research has focused on attention-based architectures. TransformerCPI (Chen et al. 2020) was the first to adapt the transformer model for this task, utilizing multi-head self-attention on atom-level compound tokens to enhance interpretability without sacrificing performance. Building on this, MolTrans (Huang et al. 2021) combined sub-structural pattern mining with transformer attention, significantly improving performance in challenging cold-start scenarios.

A more recent paradigm shift involves structure-aware methods that incorporate molecular graph representations. GraphDTA (Nguyen et al. 2021) introduced a significant advance by integrating Graph Neural Networks (GNNs) to process compound molecular graphs, while retaining a 1D-CNN for protein sequences. This hybrid approach achieved state-of-the-art results by leveraging rich structural information, including atom types and bond connectivity.

Most recently, the field has begun to leverage pre-trained language models for representation learning. These models, pre-trained on large-scale unlabeled biomedical data, show great promise in capturing complex molecular and biological patterns. However, significant challenges remain in effectively integrating such multimodal information and addressing domain-shift issues.

Large Language Models for Drug Discovery

The application of large language models to molecular and protein data has become a transformative force in computational drug discovery. These models leverage self-supervised pre-training on massive datasets to learn rich, semantically meaningful representations of chemical and biological entities.

Molecular language models have rapidly advanced through several paradigms. Initial work adapted BERT-style architectures for SMILES strings, with models like SMILES-BERT (Wang et al. 2019) and ChemBERTa (Chithrananda, Grand, and Ramsundar 2020) pi-

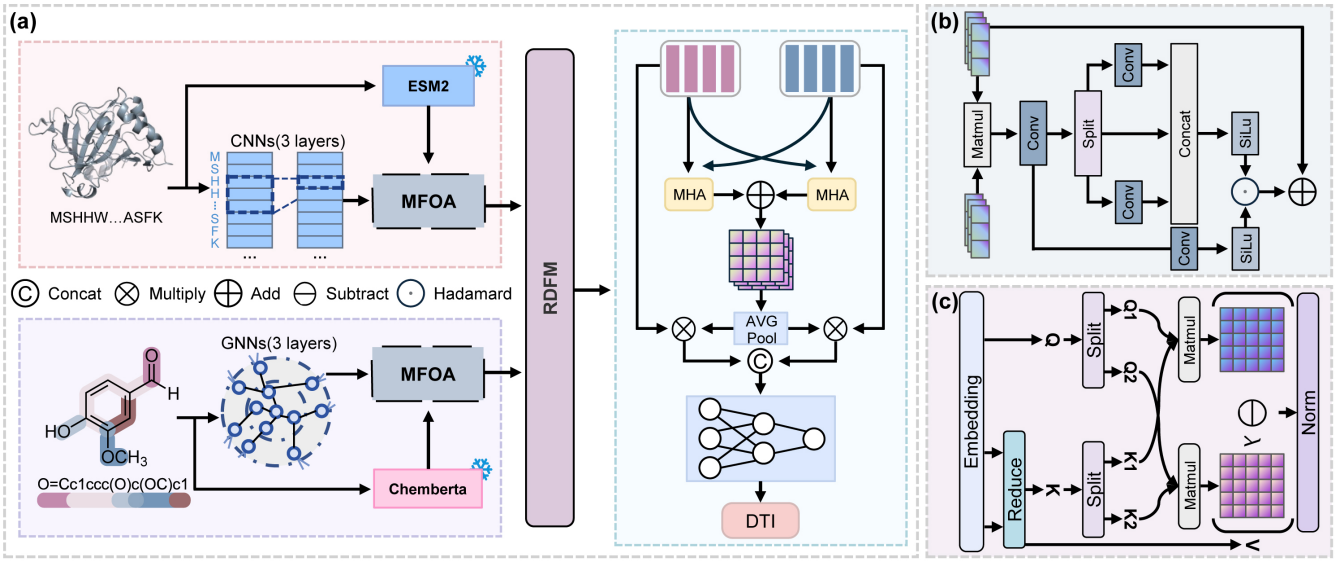

Figure 1: (a). The overall framework of FuseMine. The model leverages two pre-trained language models (ChemBERTa and ESM-2) to encode SMILES and protein sequences, respectively. Structural representations are extracted using a 3-layer GNN for compounds and a 3-layer CNN for proteins. These features are fused via the MFOA module. The resulting embeddings from both molecules are passed to the RDFM module for further refinement. The final representation is used for DTI prediction via a bidirectional cross-attention mechanism (b). The internal structure of the MFOA module. (c). The structure of the RDFM module.

oneering masked language modeling on large molecular datasets. A parallel stream focused on generative models, such as MolGPT (Bagal et al. 2021) and Chemformer (Irwin et al. 2022), for molecular design. More recently, systems like ChemCrow (Bran et al. 2023) have integrated LLMs with chemical tools to create interactive, tool-augmented research platforms.

Protein language models have followed a similar trajectory. After early LSTM-based models like SeqVec (Heinzinger et al. 2019), the field quickly adopted more powerful transformer architectures. Models like ProtT5 (Elnaggar et al. 2021) demonstrated the power of scaling, while the influential ESM series (Rives et al. 2019; Lin et al. 2022) showed that structural information emerges from large-scale sequence pre-training, enabling accurate structure prediction with ESMFold. The most groundbreaking advance in this field came from DeepMind's AlphaFold series (Jumper et al. 2021; Abramson et al. 2024), which predicts highly accurate 3D structural models from amino acid sequences. Concurrently, generative models like ProGen (Madani et al. 2023) have enabled the design of novel, experimentally validated functional proteins.

Method

In this section, we first introduce the overall framework and then provide detailed descriptions of the implementation of each component.

Problem Definition

Given a compound C and a protein P, our objective is to predict whether an interaction exists between them. This constitutes a binary classification problem. We define this task as learning a prediction function $f : (C, P) \rightarrow y$, where $y \in \{0, 1\}$, and $y = 1$ indicates the presence of an interaction.

Overview of FuseMine Framework

Figure 1 illustrates the overall architecture, which takes a compound–protein pair as input and aims to predict their potential interaction. We begin with a dual-encoder module, designed to extract comprehensive representations for both the compound and the protein. It consists of a structure encoder and a sequence-level semantic encoder, which collaboratively capture molecular topology and contextual information from sequences. Subsequently, a Multi-modal Feature Orchestration Aggregation (MFOA) module is applied to aggregate and align the structural and semantic representations. This operation captures the complementary interaction patterns across modalities and produces a unified representation of the molecule. Next, a Reduction Differential Feature Mining (RDFM) module is employed to further refine the fused features by enhancing the discriminative capacity of molecular representations. Finally, the processed features are fed into an interaction prediction module, which estimates the likelihood of interaction between the compound and the protein.

Dual Molecular Encoder

We employ four parallel encoders to extract multimodal feature representations of compounds and proteins. Specifically, we utilize a 3-layer Graph Neural Network (GNN) to process molecular graph structures and a Convolutional

Neural Network (CNN) to process protein sequences, obtaining convolutional feature embeddings H_{ct} and H_{pt}, respectively.

$$\mathbf{H}_{ct}^{(l+1)} = \sigma\Big(\hat{\mathbf{D}}^{-\frac{1}{2}}\,\hat{\mathbf{A}}\,\hat{\mathbf{D}}^{-\frac{1}{2}}\,\mathbf{H}_{ct}^{(l)}\,\mathbf{W}_{c}^{(l)}\Big) \tag{1}$$

$$\mathbf{H}_{pt}^{(l+1)} = \sigma\Big(\mathrm{BN}\big(\mathrm{CNN}(\mathbf{W}_{p}^{(l)}, \mathbf{H}_{pt}^{(l)})\big)\Big) \tag{2}$$

where $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix with self-connections, $\mathbf{I}$ is the identity matrix, $\hat{\mathbf{D}}$ is the diagonal node-degree matrix of $\hat{\mathbf{A}}$, $\mathbf{H}_{ct}^{(l)} \in \mathbb{R}^{n\times d}$ denotes the latent representation of the l-th layer, $\mathbf{W}_{c}^{(l)}$ is the layer specific learnable weight matrix, and $\sigma(\cdot)$ is the ReLU activation function.

Concurrently, we leverage ChemBERTa-MTR (Chithrananda, Grand, and Ramsundar 2020) and ESM-2 (Lin et al. 2022) pre-trained language models to encode molecular SMILES sequences and amino acid sequences, yielding sequence feature representations $\mathbf{H}_{cs}$ and $\mathbf{H}_{ps}$.

Multi-modal Feature Orchestration Aggregation

To effectively integrate the structural and sequential information, we propose a Multi-modal Feature Orchestration Aggregation (MFOA) module that systematically combines heterogeneous features through a three-stage orchestration process. We illustrate the process using the compound as an example, while an identical procedure is employed for the protein.

Preliminary Feature Fusion. Given compound structural features $\mathbf{H}_{ct}$ from the GNN encoder and compound sequential features $\mathbf{H}_{cs}$ from the ChemBERTa-MTR model, we first perform matrix multiplication to obtain preliminary fused compound features:

$$\mathbf{H}_{c} = \mathrm{Conv}(\mathbf{H}_{ct}\mathbf{H}_{cs}^{T}) \tag{3}$$

where $\mathrm{Conv}(\cdot)$ denotes a convolutional transformation that refines the initial fusion of compound structure and sequence information.

Orchestrated Feature Partitioning. The fused compound features $\mathbf{H}_{c}$ are strategically partitioned into three functionally complementary subsets with a carefully designed 3:4:1 ratio:

$$\mathbf{H}_{c1}, \mathbf{H}_{c2}, \mathbf{H}_{c3} = \mathrm{Split}(\mathbf{H}_{c}) \tag{4}$$

where $\mathbf{H}_{c1} \in \mathbb{R}^{n\times 3d/8}$, $\mathbf{H}_{c2} \in \mathbb{R}^{n\times d/2}$, and $\mathbf{H}_{c3} \in \mathbb{R}^{n\times d/8}$. This asymmetric partitioning strategy enables specialized processing of different aspects of compound features while maintaining information diversity between structural and sequential representations.

Coordinated Feature Aggregation. Each compound feature subset undergoes purposefully designed transformations before final aggregation:

$$\mathbf{G}_{c} = \mathrm{SiLU}(\mathrm{Conv}(\mathbf{H}_{c})) \tag{5}$$

$$\widetilde{\mathbf{H}}_{c} = \mathrm{SiLU}(\mathrm{Concat}(\mathrm{Conv}(\mathbf{H}_{c1}), \mathbf{H}_{c2}, \mathrm{Conv}(\mathbf{H}_{c3}))) \tag{6}$$

$$\mathbf{Z}_{st}^{c} = \mathbf{H}_{ct} + \mathrm{Conv}(\mathbf{G}_{c} \odot \widetilde{\mathbf{H}}_{c}) \tag{7}$$

where $\odot$ denotes element-wise product, $\mathrm{SiLU}(\cdot)$ represents the Swish activation function, and $\mathrm{Concat}(\cdot)$ performs channel-wise concatenation. We retain the structural features $\mathbf{H}_{st}$ as a guidance term during the fusion process to enhance cross-modal interactions. An identical process is applied to protein features, resulting in the fused representation $\mathbf{Z}_{st}^{p}$.

Reduction Differential Feature Mining

We denote the fused representations of the compound and the protein obtained from the MFOA module as $\mathbf{Z}_{st}^{c}$ and $\mathbf{Z}_{st}^{p}$, respectively. These representations are subsequently fed into the Reduction Differential Feature Mining (RDFM) module to enhance the discriminative power by suppressing irrelevant or noisy components. This module further guides the model to identify and extract key discriminative features that are critical for complex molecular interactions, thereby enhancing its capacity for downstream interaction modeling and improving prediction accuracy.

Given the fused feature matrix $\mathbf{Z}_{st}^{c}$ and $\mathbf{Z}_{st}^{p}$, we apply a convolutional projection to reduce the dimensionality of the key and value vectors:

$$\mathbf{K}_{r}^{c} = W_{k}\mathrm{Conv}(\mathbf{Z}_{st}^{\mathrm{c}}), \quad \mathbf{V}_{r}^{c} = W_{v}\mathrm{Conv}(\mathbf{Z}_{st}^{\mathrm{c}}) \tag{8}$$

$$\mathbf{K}_{r}^{p} = W_{k}\mathrm{Conv}(\mathbf{Z}_{st}^{\mathrm{p}}), \quad \mathbf{V}_{r}^{p} = W_{v}\mathrm{Conv}(\mathbf{Z}_{st}^{\mathrm{p}}) \tag{9}$$

This reduction operation decreases computational overhead and improves focus by retaining the most informative components.

We leverage differential attention (Ye et al. 2024) to further filter out redundant or weakly relevant regions within the fused structure-sequence representations. This idea is analogous to differential amplifiers proposed in electrical engineering (Laplante et al. 2018), where the output is computed as the difference between two input signals to eliminate common-mode noise. Differential attention eliminates attention noise by computing the difference between two softmax attention functions. The differential attention is computed as:

$$\mathrm{Diff} = \left[\mathrm{softmax}\left(\frac{\mathbf{Q}_1\mathbf{K}_1^{T}}{\sqrt{d}}\right) - \lambda \cdot \mathrm{softmax}\left(\frac{\mathbf{Q}_2\mathbf{K}_2^{T}}{\sqrt{d}}\right)\right]\mathbf{V} \tag{10}$$

The core design involves decomposing both the query and key matrices into two complementary subspaces: $\mathbf{Q}_1$, $\mathbf{Q}_2 \in \mathbb{R}^{n\times d/2}$ and $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{n\times d/2}$, respectively. A shared value matrix $\mathbf{V} \in \mathbb{R}^{n\times d}$ is retained for downstream attention computation. Here, λ denotes a learnable scalar that modulates the suppression strength of the second attention branch. The first attention term is designed to capture the dominant sequence–structure interaction patterns, while the second term, scaled by λ, highlights components that should be selectively suppressed, thereby enhancing attention sparsity and robustness. The RDFM module produces the final refined features $\mathbf{Z}_{c}$ and $\mathbf{Z}_{p}$, which are then used as inputs to the interaction prediction stage.

Compound Protein Interaction Prediction

To capture the intricate interaction patterns between compounds and proteins, FuseMine employs a bidirectional

cross-attention mechanism that generates complementary representations through multiple attention heads. This mechanism systematically aggregates attention patterns to construct comprehensive interaction mappings for each molecular perspective. For each attention head $h \in \{1, 2, \ldots, H\}$, this bidirectional process unfolds as follows: The compound-centric attention pathway captures how specific compound substructures relate to protein binding sites:

$$\text{Attn}_{c\to p}^{(h)} = \text{softmax}\left(\frac{\mathbf{Q}_c^{(h)}(\mathbf{K}_p^{(h)})^T}{\sqrt{d_k}}\right) \tag{11}$$

Conversely, the protein-centric pathway identifies which protein regions are most relevant for compound binding:

$$\text{Attn}_{p\to c}^{(h)} = \text{softmax}\left(\frac{\mathbf{Q}_p^{(h)}(\mathbf{K}_c^{(h)})^T}{\sqrt{d_k}}\right) \tag{12}$$

where $\text{Attn}_{c\to p}^{(h)}$ denotes the attention from compound features to protein features, and $\text{Attn}_{p\to c}^{(h)}$ represents the reverse attention flow, the transformations follow standard multi-head attention with learnable projections $\mathbf{W}_Q^{(h)}, \mathbf{W}_K^{(h)}, \mathbf{W}_V^{(h)}$ We derive global attention descriptors $\mathbf{A}_c$ and $\mathbf{A}_p$ by applying average pooling to the attention matrix along the protein and compound dimensions, respectively.

$$\mathbf{A}_c, \mathbf{A}_p = \text{AvgPool}(\text{Attn}_{c\to p}^{(h)} + \text{Attn}_{p\to c}^{(h)}) \tag{13}$$

Finally, we update the features based on the global descriptors as follows:

$$\mathbf{F}_c = (\mathbf{A}_c\mathbf{Z}_c) + \mathbf{Z}_c \tag{14}$$

$$\mathbf{F}_p = (\mathbf{A}_p\mathbf{Z}_p) + \mathbf{Z}_p \tag{15}$$

This enables the model to adaptively weight interaction-specific features while maintaining a connection to the original molecular embeddings through residual connections. The final prediction emerges from a multi-layer perceptron that transforms the joint representation into an interaction probability.

$$\hat{y} = \text{MLP}(\text{Concat}(\mathbf{F}_c, \mathbf{F}_p)) \tag{16}$$

Given that we treat CPI prediction as a binary classification task, the cross-entropy loss is employed to train our model:

$$\mathcal{L} = -\left[y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right] \tag{17}$$

where y is the binary interaction label, and $\hat{y}$ is the predicted likelihood.

Experiments

Datasets and Splitting Protocols

We evaluate our method on four benchmark datasets: Human, C.elegans (Tsubaki, Tomii, and Sese 2018), BindingDB (Liu et al. 2007), and BioSNAP (Zitnik, Sosic, and Leskovec 2018). To ensure fair comparison, we adopt the identical data partitioning protocol used in the prior works (Bai et al. 2023; Xie, Tu, and Xu 2024).

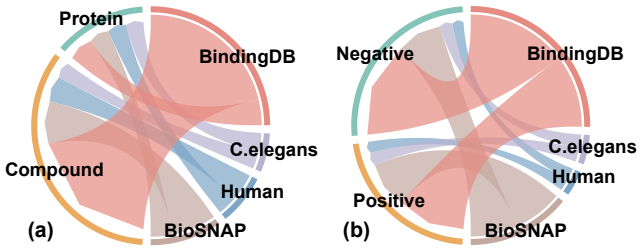

Figure 2: Dataset information: (a) Distribution of compound and proteins across four datasets. (b) Distribution of positive and negative interaction samples in each dataset.

Intra-Domain Random Splitting. For the smaller Human and C.elegans datasets, we randomly divide the data into training, validation, and test sets using an 8:1:1 ratio. For the larger BindingDB and BioSNAP datasets, we adopt a 7:1:2 split.

Intra-Domain Cold-Pair Splitting. On BindingDB and BioSNAP, we further apply a cold-pair protocol: 70% of compounds and proteins are randomly selected, and all associated compound–protein pairs constitute the training set. The remaining 30% of compound–protein pairs are then split 3:7 into validation and test sets, ensuring that every compound and protein in the test set remains unseen during training.

Cross-Domain Splitting. Following DrugBAN, we cluster compounds using ECFP4 fingerprints and proteins using PSC descriptors. Sixty percent of the resulting clusters constitute the source domain, with all associated compound-protein pairs serving as labeled source data. The remaining 40% of clusters form the target domain, ensuring disjoint distributions between the source and target domains.

Baselines

We evaluate our method against ten representative baselines: SVM (Cortes and Vapnik 1995), RF (Ho 1995), GraphDTA (Nguyen et al. 2021), DeepConv-DTI (Lee, Keum, and Nam 2019), MolTrans (Huang et al. 2021), TransformerCPI (Chen et al. 2020), HyperattentionDTI (Zhao et al. 2022b), DrugBAN (Bai et al. 2023), MlanDTI (Xie, Tu, and Xu 2024) and LAM-DTI (Wei, Wang, and Tang 2025). These baselines encompass both classical machine learning models and state-of-the-art deep learning architectures, providing a comprehensive performance spectrum. All neural baselines were run with the official hyper-parameters released by their authors.

Implementation Details

Our model is implemented using PyTorch Lightning and trained on NVIDIA A800 GPUs. We adopt the AdamW optimizer (Loshchilov and Hutter 2017) with learning rate 5×10^{-5}, batch size 64, and train the model for 100 epochs. Performance is evaluated using AUC-ROC, AUPR, and F1-score. All evaluation metrics are computed on the test set, with results averaged over five independent runs.

Method	human			C.elegans			BindingDB			BioSNAP		
	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1
SVM	0.910	–	**0.967**	0.894	–	0.801	0.939	0.928	0.787	0.862	0.864	0.762
RF	0.940	–	0.878	0.902	–	0.832	0.942	0.921	0.858	0.860	0.886	0.808
GraphDTA	0.960	0.959	0.897	0.974	0.975	0.919	0.951	0.934	0.867	0.887	0.890	0.789
DeepConvDTI	0.967	0.964	0.922	0.983	0.985	0.944	0.945	0.925	0.859	0.886	0.890	0.797
MolTrans	0.974	0.976	0.944	0.982	0.985	0.966	0.952	0.936	0.865	0.895	0.897	0.824
TransformerCPI	0.973	0.975	0.920	0.988	0.986	0.952	0.943	0.925	0.855	0.889	0.893	0.798
HyperAttDTI	0.984	0.984	0.946	0.989	0.990	0.958	0.959	0.948	0.887	0.901	0.902	0.838
DrugBAN	0.981	0.983	0.940	0.986	0.988	0.949	0.959	0.947	0.881	0.903	0.902	0.832
MlanDTI	0.988	**0.990**	0.961	0.990	0.992	0.962	0.945	0.926	0.857	0.909	0.912	0.841
LAM-DTI	0.856	0.876	0.724	0.908	0.916	0.809	0.878	0.837	0.765	0.893	0.897	0.812
FuseMine	**0.989**	0.989	0.957	**0.996**	**0.994**	**0.984**	**0.968**	**0.957**	**0.900**	**0.918**	**0.922**	**0.843**

Table 1: The results of the proposed model and baselines on four datasets. The best results are indicated by bold. "–" means no result for this metric.

Intra-domain Random Split Experiments

Table 1 presents a comparative analysis of FuseMine against several baseline models across four datasets. On the smaller datasets (Human and C.elegans), FuseMine achieves modest improvements across various metrics. Specifically, FuseMine attains an AUC of 0.989 on the Human dataset, representing only a marginal improvement over the second-best model MlanDTI (AUC = 0.988). Similarly, on the C.elegans dataset, FuseMine achieves an AUC of 0.996, slightly outperforming MlanDTI (AUC = 0.990). These limited performance gains can be attributed to the smaller scale of these datasets, which constrains the model's potential for improvement. On the larger datasets (BindingDB and BioSNAP), FuseMine demonstrates superior generalization performance with more substantial improvements. These enhanced results demonstrate the stability and scalability of FuseMine when handling larger and more complex datasets. These findings indicate that while FuseMine exhibits certain advantages, the extent of improvement varies with dataset size. However, despite these positive outcomes from random splitting, it is important to acknowledge potential data leakage, as compounds or proteins may appear in both training and test sets, potentially leading to overly optimistic performance estimates.

Methods	BindingDB			BioSNAP		
	AUC	AUPR	F1	AUC	AUPR	F1
Moltrans	0.595	0.522	0.511	0.672	0.697	0.437
TransformerCPI	0.656	0.594	0.566	0.680	0.708	0.523
HyperAttDTI	0.661	0.598	0.582	0.732	0.760	0.539
DrugBAN	0.655	0.600	0.542	0.651	0.667	0.449
$\text{DrugBAN}_{\text{CDAN}}$	NA	NA	NA	NA	NA	NA
MlanDTI	0.671	0.594	0.601	0.782	0.801	**0.653**
MlanDTI (with PL)	NA	NA	NA	NA	NA	NA
LAM-DTI	0.682	0.568	0.585	0.748	0.758	0.646
FuseMine	**0.745**	**0.665**	**0.606**	**0.798**	**0.814**	0.611

Table 2: Cold-Pair Split Results.

Methods	BindingDB			BioSNAP		
	AUC	AUPR	F1	AUC	AUPR	F1
Moltrans	0.537	0.476	0.389	0.632	0.635	0.401
TransformerCPI	0.568	0.450	0.410	0.656	0.693	0.432
HyperAttDTI	0.545	0.462	0.376	0.654	0.685	0.395
DrugBAN	0.578	0.471	0.484	0.608	0.606	0.438
$\text{DrugBAN}_{\text{CDAN}}$	0.616	0.512	0.426	0.673	0.706	0.542
MlanDTI	0.657	0.537	0.489	0.728	0.759	0.604
MlanDTI (with PL)	**0.687**	0.579	0.564	0.749	0.770	**0.629**
LAM-DTI	0.531	0.453	0.547	0.649	0.720	0.494
FuseMine	0.657	**0.592**	**0.623**	**0.765**	**0.792**	0.591

Table 3: Cross-domain Split Results.

Intra-domain Cold Pair Split Experiments

In cold-start settings, training and testing samples are completely isolated at both compound and protein levels, providing a more realistic assessment of the model's extrapolation ability to unknown entities. Table 2 shows that FuseMine achieves top-ranking performance on both BindingDB and BioSNAP datasets, obtaining AUC scores of 0.745 and 0.798, respectively. These scores represent improvements of 11.0% and 2.0% over the second-best method. Additionally, AUPR values improved by 10.8% and 1.6%, respectively. These performance gains demonstrate FuseMine's capability to capture discriminative features across diverse chemical and sequence spaces. However, based on fixed-threshold F1 scores, FuseMine only marginally outperforms MlanDTI on BindingDB, achieving 0.606 compared with MlanDTI's 0.601, while underperforming on BioSNAP with 0.611 against MlanDTI's 0.653. This indicates that although FuseMine optimizes overall ranking performance, class imbalance sensitivity still affects decision boundary determination. Overall, FuseMine's robust ranking performance under the demanding cold-start scenario confirms its practical potential as a screening tool for large-scale compound-protein prediction.

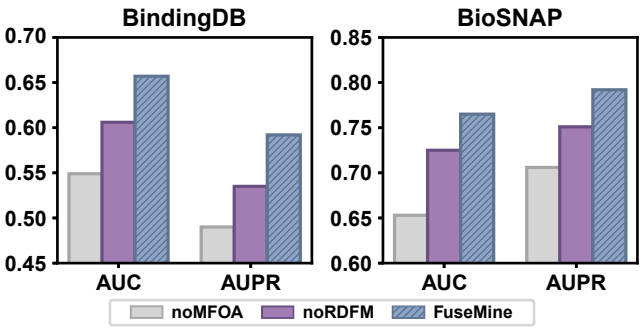

Figure 3: Ablation study on BindingDB and BioSNAP dataset.

Cross-Domain Experiments

The cross-domain evaluation results underscore the challenge of achieving generalization in compound-protein interaction prediction, as the performance of various methods generally deteriorates in cross-domain settings compared to cold-start scenarios. Table 3 shows that AUC values of most baseline methods drop significantly on BindingDB, with several methods achieving AUC scores below 0.58 and approaching random performance, whereas results on BioSNAP are relatively better. Notably, technical enhancement strategies exhibit clear effectiveness: MlanDTI, which employs pseudo-labeling techniques, substantially improves performance on both datasets, raising AUC from 0.657 to 0.687 on BindingDB and from 0.728 to 0.749 on BioSNAP. Similarly, DrugBAN$_{\text{CDAN}}$ notably outperforms the original DrugBAN, particularly on BioSNAP, achieving a substantial increase in AUC from 0.608 to 0.673, confirming the efficacy of adversarial domain adaptation strategies. Our proposed FuseMine model performs exceptionally well in this challenging scenario, achieving an AUC of 0.657 and an AUPR of 0.592 on BindingDB. Although its AUC is comparable to that of the best baseline, FuseMine's F1 score significantly surpasses those of other methods at 0.623. On BioSNAP, FuseMine attains the highest AUC and AUPR among all methods, at 0.762 and 0.794, respectively. These results demonstrate that the innovative technical design of FuseMine effectively mitigates inherent difficulties in cross-domain generalization, providing more reliable predictive performance in realistic lead-compound transfer scenarios and clearly validating our approach's strengths in addressing critical challenges in drug discovery.

Ablation Studies

To quantitatively evaluate the individual contributions of each proposed component within the FuseMine framework, we conduct comprehensive ablation experiments on two benchmark datasets: BindingDB and BioSNAP. Specifically, we assess the performance impact of removing the Multi-modal Feature Orchestration Aggregation (MFOA) and the Reduction Differential Feature Mining (RDFM) modules, respectively, denoted as noMFOA and noRDFM in the figure.

As illustrated in Figure 3, both components are critical to the model's effectiveness. The noMFOA variant results in a substantial decline in performance across both AUC and AUPR metrics, highlighting the importance of MFOA in facilitating effective multi-modal feature alignment and interaction. Similarly, the noRDFM variant causes noticeable performance degradation, underscoring the role of RDFM in enhancing the discriminative capacity of fused features. Overall, the complete FuseMine model consistently achieves superior results compared to its ablated variants, thereby validating the complementary benefits of integrating both MFOA and RDFM modules.

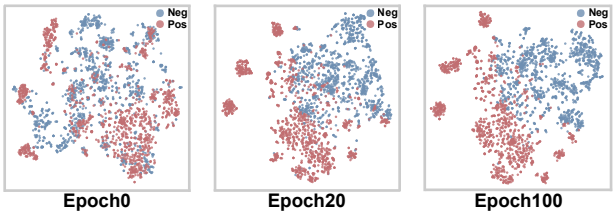

Figure 4: Visualization of the fused embeddings of compound-protein pairs under different epochs.

Embedding Visualization

We visualize the fused embeddings of compound–protein pairs on the Human dataset using t-SNE at epochs 0, 20, and 100, as illustrated in Figure 4. Embeddings of positive and negative samples evolve from being highly entangled to forming compact, well-separated clusters, indicating progressively enhanced discriminative power. This behavior demonstrates that MFOA effectively fuses structural and sequential modalities, while RDFM further refines the fused features by mining and emphasizing informative interaction patterns, thereby yielding more separable representations for positive and negative samples.

Conclusion

In this work, we propose FuseMine, a multi-modal deep-fusion framework that integrates graph neural networks, convolutional neural networks, and pre-trained language models to predict compound–protein interactions. FuseMine introduces the Multi-modal Feature Orchestration and Aggregation (MFOA) and Reduction Differential Feature Mining (RDFM) modules, which together enable deep cross-modal fusion and enhance the model's ability to capture discriminative features. We evaluate FuseMine under both intra-domain and cross-domain settings. Results on four benchmark datasets demonstrate that it outperforms state-of-the-art baselines, achieving substantial improvements in both AUC and AUPR metrics. Ablation studies further validate the individual contributions of each component, confirming the necessity of the proposed modules. These results highlight the synergistic value of combining structural and sequential data for compound–protein interactions.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. 62302156, 62402351, and 62402349), the Natural Science Foundation of Hunan Province (Grant Nos. 2023JJ40180), the Natural Science Foundation of Hubei Province (Grant Nos. 2024AFB127 and 2024AFB275), and Wuhan Textile University Foundation (Grant Nos. 20230612 and 2024309).

References

Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; et al. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016): 493–500.

Bagal, V.; Aggarwal, R.; Vinod, P.; and Priyakumar, U. D. 2021. MolGPT: molecular generation using a transformer-decoder model. *Journal of chemical information and modeling*, 62(9): 2064–2076.

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.

Barkat, M. R.; Moussa, S. M.; and Badr, N. L. 2021. Drug-target interaction prediction using machine learning. In *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, 480–485. IEEE.

Berdigaliyev, N.; and Aljofan, M. 2020. An overview of drug discovery and development. *Future medicinal chemistry*, 12(10): 939–947.

Bran, A. M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; and Schwaller, P. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

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.

Chithrananda, S.; Grand, G.; and Ramsundar, B. 2020. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Cortes, C.; and Vapnik, V. 1995. Support-vector networks. *Machine learning*, 20(3): 273–297.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. 2021. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10): 7112–7127.

Heinzinger, M.; Elnaggar, A.; Wang, Y.; Dallago, C.; Nechaev, D.; Matthes, F.; and Rost, B. 2019. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20(1): 723.

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

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.

Irwin, R.; Dimitriadis, S.; He, J.; and Bjerrum, E. J. 2022. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1): 015022.

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.

Laplante, P. A.; Cravey, R.; Dunleavy, L. P.; Antonakos, J. L.; LeRoy, R.; East, J.; Buris, N. E.; Conant, C. J.; Fryda, L.; Boyd, R. W.; et al. 2018. *Comprehensive dictionary of electrical engineering*. CRC Press.

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.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*.

Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; and Gilson, M. K. 2007. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1): D198–D201.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Madani, A.; Krause, B.; Greene, E. R.; Subramanian, S.; Mohr, B. P.; Holton, J. M.; Olmos Jr, J. L.; Xiong, C.; Sun, Z. Z.; Socher, R.; et al. 2023. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41(8): 1099–1106.

Madhukar, N. S.; Khade, P. K.; Huang, L.; Gayvert, K.; Galletti, G.; Stogniew, M.; Allen, J. E.; Giannakakou, P.; and Elemento, O. 2019. A Bayesian machine learning approach for drug target identification using diverse data types. *Nature communications*, 10(1): 5221.

Nagamine, N.; and Sakakibara, Y. 2007. Statistical prediction of protein–chemical interactions based on chemical structure and mass spectrometry data. *Bioinformatics*, 23(15): 2004–2012.

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.

Peska, L.; Buza, K.; and Koller, J. 2017. Drug-target interaction prediction: a Bayesian ranking approach. *Computer methods and programs in biomedicine*, 152: 15–21.

Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; and Fergus, R. 2019. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *PNAS*.

Sun, Y.; Li, Y. Y.; Leung, C. K.; and Hu, P. 2024. iNGNN-DTI: prediction of drug–target interaction with interpretable nested graph neural network and pretrained molecule models. *Bioinformatics*, 40(3): btae135.

Tsubaki, M.; Tomii, K.; and Sese, J. 2018. Compound-protein Interaction Prediction with End-to-end Learning of Neural Networks for Graphs and Sequences. *Bioinformatics*.

Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; and Huang, J. 2019. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, 429–436. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366663.

Wei, Z.; Wang, Z.; and Tang, C. 2025. Dynamic Prediction of Drug–Target Interactions via Cross-Modal Feature Mapping with Learnable Association Information. *Journal of Chemical Information and Modeling*, 65(8): 3915–3927.

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.

Ye, T.; Dong, L.; Xia, Y.; Sun, Y.; Zhu, Y.; Huang, G.; and Wei, F. 2024. Differential transformer. *arXiv preprint arXiv:2410.05258*.

Zhao, L.; Zhu, Y.; Wang, J.; Wen, N.; Wang, C.; and Cheng, L. 2022a. A brief review of protein–ligand interaction prediction. *Computational and Structural Biotechnology Journal*, 20: 2831–2838.

Zhao, Q.; Zhao, H.; Zheng, K.; and Wang, J. 2022b. HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics*, 38(3): 655–662.

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).