# Sequence-Free for Compound–Protein Interaction Prediction

**Hongzhi Zhang[1], Jiameng Chen[1], Kun Li[1], Yida Xiong[1], Xiantao Cai[1*], Wenbin Hu[2,1*], Jia Wu[3]**

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]Shenzhen Research Institute, Wuhan University, Shenzhen, China
[3]Department of Computing, Macquarie University, Sydney, Australia
{zhanghongzhi, jiameng.chen, likun98, yidaxiong, caixiantao, hwb}@whu.edu.cn, Jia.wu@mq.edu.au

## Abstract

The prediction of compound–protein interactions (CPIs) is crucial for drug discovery. Most existing CPI prediction models rely on protein sequence information as input. However, in early-stage drug development, particularly in phenotype-driven studies or compound-response analyses, proteins are often annotated only with functional labels, and their sequences remain undetermined. Consequently, current methods are inapplicable in such scenarios. Furthermore, our experiments find that even when large-scale perturbations were applied to protein sequences, the predictive performance of the existing models did not show a significant decline. It indicates that the high investment in sequencing may not bring corresponding returns. To address the above issues, we propose an inexpensive, protein-sequencing-free framework **BioText-CPI**, based on the **Bio**medical **Text**ual description of protein for **CPI** prediction. Firstly, during the pre-training stage of the model, we use contrastive learning to align protein texts and sequence modalities. Subsequently, we add biological text descriptions of proteins to the existing public CPI dataset to construct a new CPI dataset. Finally, in the CPI prediction stage, the sequence and biomedical text descriptions of proteins can be used as the input for CPI prediction either separately or simultaneously to meet the application requirements of different scenarios. The experiments demonstrate that BioText-CPI achieves comparable effects to the traditional methods when only the biomedical description of protein is input. Moreover, when the two modalities of protein information are input simultaneously, BioText-CPI achieves state-of-the-art performance across multiple scenarios.

**Code** — https://github.com/Hoch-Zhang/BioText-CPI

## Introduction

Targeted drug research is an important research direction in modern pharmaceutical research and development and precision medicine. With the development of deep learning technology, a large number of representation learning-based methods (Peng et al. 2021; Bian et al. 2023; Dalkıran et al. 2023) have been proposed to solve the problem of high-precision CPI prediction, making it possible to screen a large number of compound molecules in a relatively short time.

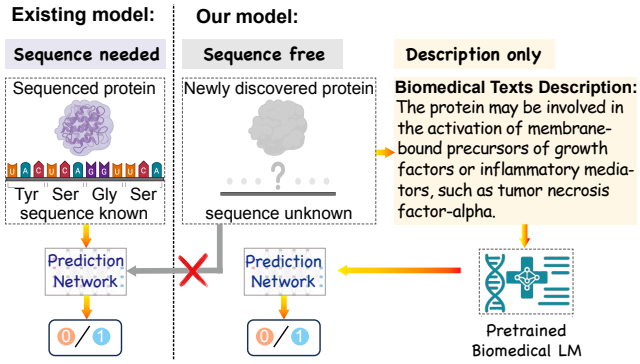



Figure 1: Difference between our model and existing mainstream models. Left: Existing models must rely on the sequence of proteins to complete virtual screening, which cannot meet the needs of proteins with unknown sequences. Right: Our model only needs to know the possible biomedical descriptions of proteins to complete the prediction.

As shown in Figure 1, most existing methods (Torng and Altman 2019; Peng et al. 2024b) usually adopt protein sequence information as the input of the encoder, and predict the probability of their interaction by mapping the protein and compound molecules in the same space.

Despite the significant advancements in protein sequencing technologies, our current knowledge encompasses only a limited portion of the extensive protein diversity present in nature. As a result, in numerous practical drug discovery settings, the sequence information of target proteins is often inaccessible. For example, in certain high-throughput functional screening experiments, particularly those based on phenotypic or compound reaction outcomes, proteins may be annotated as *related to a specific functional category* without their sequences being provided (Hughes et al. 2021; Lehmann et al. 2022). Similarly, in disease-related research, certain proteins may be indirectly recognized as potential therapeutic targets through phenotypic, metabolite or pathway enrichment analyzes, but their sequences have not been conclusively determined (Chu et al. 2021; Qiu et al. 2023). In such early-stage studies, protein sequences are typically not characterized, and only tentative biomedical associations are available. Therefore, developing an effective approach to

*Corresponding authors.

| Cond. | SB | UC | UP | UB |
|---|---|---|---|---|
| Original | 89.9 | 85.2 | 73.1 | 60.1 |
| Pert70 | 88.5(-1.4) | 83.8(-1.4) | 69.4(-3.7) | 56.9(-3.2) |
| Pert80 | 88.8(-1.1) | 84.2(-1.0) | 68.1(-5.0) | 57.0(-3.1) |
| Pert90 | 88.7(-1.2) | 84.2(-1.0) | 68.4(-4.7) | 56.3(-3.8) |
| Pert100 | 88.6(-1.3) | 84.2(-1.0) | 68.8(-4.3) | 57.7(-2.4) |

Table 1: The model effectiveness for CPI in multiple scenarios under the condition of amino acids with different perturbation ratios. The results are the average of five different random seed experiments on BioSNAP (Zitnik et al. 2018) dataset (measured using AUROC). *Abbr.,* Cond.: Conditions; SB: *Seen-Both*; UC: *Unseen-Compound*; UP: *Unseen-Protein*; UB: *Unseen-Both*.

screen compounds based solely on biomedical descriptions related to proteins presents a significant challenge that current methods have yet to address.

In addition, we conduct a series of experiments to systematically evaluate the importance of protein sequence information for model performance. The experimental conditions are to randomly replace the amino acids in the protein sequence of the original CPI data to perturb the original data, and then perform CPI prediction on the perturbed data. To conduct a more comprehensive assessment of the influence of protein sequences, we divide the BioSNAP (Zitnik et al. 2018) dataset into four subsets according to the presence of molecules and proteins in the training data, similar to PSC-CPI (Wu et al. 2024). We adopt Deep-ConvDTI (Lee, Keum, and Nam 2019), a widely used classical model for CPI prediction. The results are summarized in Table 1. Here, *Original* denotes evaluation on the unmodified CPI dataset, while *Pert70* refers to the scenario where 70% of the amino acids in the protein sequences are randomly substituted. *Pert80* and *Pert100* are also similar. Notably, even when the majority of amino acids are perturbated, the model's performance does not degrade substantially in four scenarios. Experiments show that even when the protein sequence information is greatly disturbed, the prediction performance of the existing methods does not show a significant decline. Therefore, in the context where the impact of sequence perturbation on performance is limited, investing a large amount of sequencing costs to obtain high-quality sequences may lead to the predicament of high investment but limited returns for existing methods in practical applications.

To address these challenges, we propose an inexpensive and practical protein sequence-free framework, **BioText-CPI**, which leverages **Bio**medical **Text**ual descriptions of proteins for **CPI** prediction. BioText-CPI enables the screening of compound molecules by relying solely on the biomedical description text of proteins, even under conditions where protein sequences are entirely missing. First, during the pretraining phase, we perform contrastive learning between a protein sequence encoder and a biomedical text encoder to align their latent representations. This process allows the textual modality to implicitly absorb and incorporate knowledge from protein sequences. Subsequently, we reconstruct the CPI dataset by retrieving relevant biomedical descriptions from the publicly available UniProt database (Consortium 2019). The format of these text descriptions is consistent with that of the ProtDescribe dataset (Xu et al. 2023). Based on these datasets, BioText-CPI is capable of fine-tuning with either single-modal or multi-modal protein inputs, enabling it to adapt flexibly to a wide range of practical application scenarios. As a result, the pre-trained sequence and text encoders can effectively leverage multimodal information during training, without requiring the explicit provision of both protein representations during inference. Finally, to more fairly evaluate the generalization ability of the model, we divide the test data into four distinct groups based on whether the compounds and proteins were observed during training. Extensive experiments demonstrate that BioText-CPI achieves performance comparable to traditional sequence-based methods when only biomedical descriptions are provided. Furthermore, when both modalities of protein information are provided, BioText-CPI achieves state-of-the-art (SOTA) performance in multiple scenarios.

## Related Work

Recent CPI prediction research has shifted from traditional molecular docking (Lu et al. 2022) to more scalable and efficient representation learning approaches (Li et al. 2025b,c). These models aim to learn low-dimensional embeddings of compounds and proteins, enabling affinity prediction without explicit structural modeling. Early works (Öztürk, Özgür, and Ozkirimli 2018; Öztürk, Ozkirimli, and Özgür 2019) employed Convolutional Neural Networks (CNNs) (LeCun, Bengio et al. 1995) to encode protein sequences and compound SMILES, while methods (Li et al. 2024; Yang, Yang, and Chu 2024; Li et al. 2025a) in recent years introduced Graph Neural Networks (GNNs) (Kipf and Welling 2016; Wu et al. 2022) to better capture molecular graph topology. More recent models integrate attention mechanisms (Zhang et al. 2022; Zhao et al. 2022; Liu et al. 2024) or leverage pretrained encoders (Kang et al. 2022; Li et al. 2025c) to enhance representation quality. Despite their success, these methods share a key limitation: they all rely on protein sequences as model input. However, in many real-world scenarios, such as phenotypic screening or metagenomic annotation, protein sequences may be unavailable. This constraint motivates the development of sequence-free CPI prediction frameworks that can exploit alternative sources of protein information.

## Methods

### Problem Definition

A chemical compound can be represented as $\mathcal{C} = (\mathcal{S}_C, \mathcal{G}_C)$, where $\mathcal{S}_C = (s_1, s_2, ..., s_n)$ denotes its SMILES sequence with $n$ SMILES symbols, and $\mathcal{G}_C = (\mathcal{V}_C, \mathcal{E}_C)$ denotes a molecular graph. Each node $a_i \in \mathcal{V}_C$ corresponds to an atom in the compound, and each edge $e_{i,j} \in \mathcal{E}_C$ denotes a chemical bond between atoms $a_i$ and $a_j$. A protein consisting of $N_P$ amino acid residues can be represented by its sequence $S = (r_1, r_2, \ldots, r_{N_P})$, where each residue $r_i$ is one of the 20 standard amino acids. The biological textual description

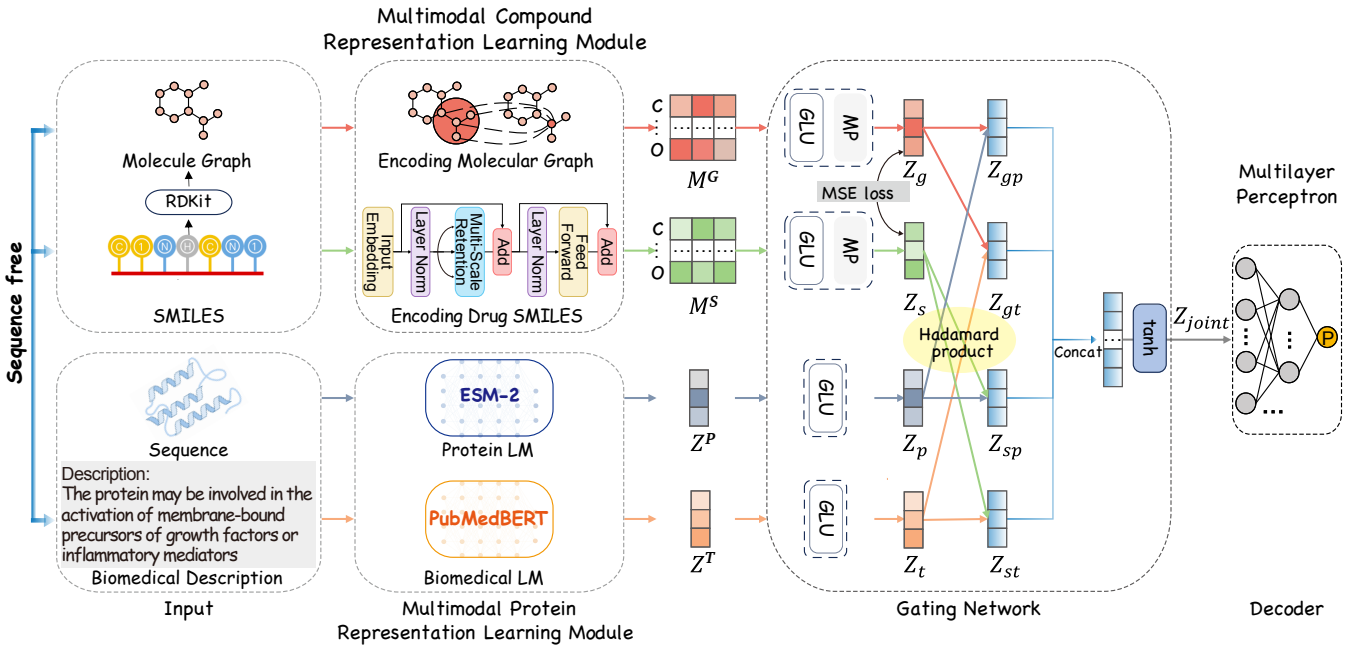



Figure 2: Graphical illustration of BioText-CPI framework. The illustrated framework supports multimodal protein inputs. For sequence-free settings, the sequence encoder branch is removed, and prediction is performed using only the available modality.

of a protein is represented as $T = (t_1, t_2, \ldots, t_{N_t})$, where $N_t$ indicates the number of tokens in the text description. Together, the sequence and biomedical text form a multimodal representation of the protein, denoted $\mathcal{P} = (S, T)$. Given a set of $N$ proteins $\{\mathcal{P}^{(i)} = (S^{(i)}, T^{(i)})\}_{i=1}^{N}$ and $\{\mathcal{C}^{(j)} = (\mathcal{S}_C^{(j)}, \mathcal{G}_C^{(j)})\}_{j=1}^{M}$ compounds, the objective of the CPI prediction is to learn two functions that capture the interaction between compounds and proteins, ultimately predicting an overall interaction strength score, which is modeled as a non-negative real value $\mathbb{R}_{\geq 0}$.

## BioText-CPI Framework

The BioText-CPI prediction framework comprises three core components: (1) a multimodal compound representation learning module, which extracts compound representations from SMILES sequence and molecular graphs; (2) a multimodal protein representation learning module, which derives protein representations from sequences, biomedical texts, or their combination; (3) a gating network module for pairwise interaction pattern prediction, which learns joint representations of compounds and proteins. The proposed BioText-CPI framework is provided in Figure 2. The illustrated framework supports multimodal protein inputs. In cases where only a single protein modality is available, the corresponding encoder branch is omitted. In the following, we provide a detailed explanation of the three components.

## Multimodal Compound Representation Learning Module

The compound representation learning module consists of two parts: an SMILES encoder and a molecular graph encoder. The SMILES encoder captures the sequential features from the linear string representation of the molecule, while the graph encoder extracts structural information by modeling atoms as nodes and bonds as edges. These two encoders work together to generate a more informative representation of the compound.

**Encoding Molecular Graph Features.** The compound encoder takes a molecular graph $\mathcal{G}_C = (\mathcal{V}_C, \mathcal{E}_C)$ as input and learns a $F$-dimensional representation for each atom (node). In this work, we employ Graph Convolutional Networks (GCNs) (Kipf and Welling 2016) as the compound encoder, which are a powerful variant of GNNs widely used for feature extraction on graph-structured data. Given a molecular graph $\mathcal{G}_C = (\mathcal{V}_C, \mathcal{E}_C)$, GCNs take the adjacency matrix $\mathbf{A}_G$ and node feature matrix $\mathbf{X}_G$ as input and output node-level embeddings. We utilize a 3-layer GCN, which is formulated as:

$$\mathcal{M}^G = \hat{\mathbf{A}} \cdot \text{ReLU}\left(\hat{\mathbf{A}} \cdot \text{ReLU}\left(\hat{\mathbf{A}}\mathbf{X}_C\mathbf{W}_G^0\right)\mathbf{W}_G^1\right)\mathbf{W}_G^2, \tag{1}$$

where $\hat{\mathbf{A}} = \hat{\mathbf{D}}^{-1/2}(\mathbf{A}_G + \mathbf{I})\hat{\mathbf{D}}^{-1/2}$ is the symmetrically normalized adjacency matrix with self-loops. Here, $\hat{\mathbf{D}}$ is the degree matrix of $\mathbf{A}_G + \mathbf{I}$, and $\mathbf{I}$ denotes the identity matrix. The trainable weight matrices $\mathbf{W}^0 \in \mathbb{R}^{d \times F}$, $\mathbf{W}^1 \in \mathbb{R}^{F \times F}$, and $\mathbf{W}^2 \in \mathbb{R}^{F \times F}$ are used to transform node features across layers, where $F$ is the hidden dimension size.

**Encoding Drug SMILES Features.** RetNet (Sun et al. 2023) is a powerful architecture designed for large language models, offering advantages such as efficient training parallelism, low computational cost, and enhanced performance. In this work, we adopt RetNet to learn feature

representations of drugs from SMILES sequence data. The SMILES string of a drug is composed of 64 distinct characters. BioText-CPI begins by embedding each SMILES token $s_i$ into a dense vector using the embedding layers. As a result, we obtain the compound embedding matrices: $\mathbf{F}_{\mathcal{S}} \in \mathbb{R}^{n \times d}$, where $d$ is the embedding dimension.

Next, these embedded features are fed into RetNet for encoding. An $L$-layer RetNet block is constructed by stacking multiscale retention (MSR) modules and feed-forward networks (FFNs). Taking the initial embedding matrix $\mathcal{M}_{\mathcal{S}}^{(0)} = \mathbf{F}_{\mathcal{S}}$ as input, RetNet outputs the final drug feature matrix $\mathcal{M}^S = \mathcal{M}_{\mathcal{S}}^{(L)}$, computed iteratively as follows:

$$\begin{cases} \mathbf{Y}_{\mathcal{S}}^{(l)} = \text{MSR}(\text{LN}(\mathcal{M}_{\mathcal{S}}^{(l)})) + \mathcal{M}_{\mathcal{S}}^{(l)} \\ \mathcal{M}_{\mathcal{S}}^{(l+1)} = \text{FNN}(\text{LN}(\mathbf{Y}_{\mathcal{S}}^{(l)})) + \mathbf{Y}_{\mathcal{S}}^{(l)}, \end{cases} \tag{2}$$

where LN$(\cdot)$ denotes Layer Normalization, and MSR$(\cdot)$ refers to Multiscale Retention. The feature matrices $\mathcal{M}_{\mathcal{S}}^{(l)}, \mathbf{Y}_{\mathcal{S}}^{(l)} \in \mathbb{R}^{n \times d}$, and the FFN is implemented as:

$$\text{FNN}(\mathbf{X}_S) = \text{gelu}(\mathbf{X}_S \mathbf{W}_1)\mathbf{W}_2, \tag{3}$$

where $\mathbf{X}_S \in \mathbb{R}^{n \times d}$, and $\mathbf{W}_1 \in \mathbb{R}^{d \times d/2}$, $\mathbf{W}_2 \in \mathbb{R}^{d/2 \times d}$ are learnable parameter matrices.

## Multimodal Protein Representation Learning

BioText-CPI incorporates two distinct modalities for representing proteins: amino acid sequences and biomedical textual descriptions. By allowing different protein modalities as input, the framework can be adapted to various application scenarios. To enable the model to infer potential sequence-level information from textual descriptions, we first pretrain a multimodal protein encoder that aligns amino acid sequences with their corresponding biomedical texts. During the downstream CPI prediction stage, the model can then generate protein representations based on either the sequence or the textual modality, depending on the available input. In this subsection, we provide a detailed description of these two components.

**Multimodal Representation Alignment.** Pretrained protein language models (PLMs) (Ahmed et al. 2020; Rives et al. 2021; Meier et al. 2021), trained on large-scale corpora of protein sequences, have demonstrated remarkable performance in a range of downstream tasks. Pretrained biomedical language models (BLMs) (Lee et al. 2020; Gu et al. 2021; Yue et al. 2024) are capable of capturing the semantic meaning of biomedical texts. When provided with protein property descriptions, BLMs can generate rich and informative textual representations. In this paper, we leverage the SOTA protein language model ESM-2 (Lin et al. 2022) to encode amino acid sequences, and adopt PubMedBERT (Gu et al. 2021), a domain-specific transformer pretrained on biomedical literature, to embed functional textual descriptions. We enable the text encoder to implicitly capture structural and functional cues from the underlying protein sequence through modal alignment.

To achieve this alignment, we adopt a contrastive learning approach between protein sequences and their associated textual descriptions. Given a batch of $M$ proteins $\{P_i = (S_i, T_i)\}_{i=1}^{M}$, we encode the sequences using a PLM, yielding $\{z_i^P\}_{i=1}^{M}$, and obtain textual embeddings $\{z_i^T\}_{i=1}^{M}$ via a BLM. The alignment is optimized using the InfoNCE loss (Oord, Li, and Vinyals 2018), defined as:

$$\begin{aligned} \mathcal{L}_{\text{C}} = -\frac{1}{2M} \sum_{i=1}^{M} \Bigg( & \log \frac{\exp(z_i^P \cdot z_i^T / \tau)}{\sum_{j=1}^{M} \exp(z_i^P \cdot z_j^T / \tau)} \\ & + \log \frac{\exp(z_i^P \cdot z_i^T / \tau)}{\sum_{j=1}^{M} \exp(z_j^P \cdot z_i^T / \tau)} \Bigg), \end{aligned} \tag{4}$$

where $\tau$ is a learnable temperature parameter. Following standard practice under multi-GPU data parallelism, we aggregate samples across all devices to construct a larger set of negative examples, and thus refer to this objective as the global contrastive loss $\mathcal{L}_{\text{C}}$.

During the pre-training phase, the model is optimized by minimizing the objective function associated with the pre-training task: $\min_\theta \mathcal{L}_{\text{C}}$, where $\theta$ represents all learnable parameters.

**Encoding Protein Features.** After the pre-training stage, we obtain a PLM and a BLM. During CPI prediction, we use these models by freezing parameters. The protein is represented as $\mathcal{P} = (S, T)$, where $S = \{r_1, \ldots, r_{N_p}\}$ is its amino-acid sequence and $T = \{t_1, \ldots, t_{N_t}\}$ is the associated biomedical text. The PLM and BLM generate $d$-dimensional representations for every amino acid and every text token, respectively:

$$\begin{cases} \mathcal{Z}^P = \text{PLM}(S) \in \mathbb{R}^d \\ \mathcal{Z}^T = \text{BLM}(T) \in \mathbb{R}^d. \end{cases} \tag{5}$$

These two sets of representations are subsequently fed into the downstream CPI prediction module.

## Gating Network for Interaction Prediction

To effectively model drug information, we adopt a multimodal encoding strategy that jointly captures both sequential (SMILES) and structural (molecular graph) representations, as illustrated in Figure 2.

While multimodal fusion provides complementary views of each compound, it may also introduce redundant or noisy signals. To address this, we incorporate a gating mechanism (Gu et al. 2020) that dynamically modulates the flow of modality-specific features, selectively emphasizing task-relevant information. This gated fusion framework facilitates more discriminative CPI modeling by integrating both expressive multimodal representations and adaptive feature control.

As shown in Figure 2, we adopt a Gated Linear Unit (GLU) (Dauphin et al. 2017) to filter out uninformative features. A max pooling (MP) operation is subsequently applied to downsample the features. The overall transformation is defined as follows:

$$\begin{aligned} \mathbf{Z}_g &= \text{MP}((\mathcal{M}^G \mathbf{W}_g + \mathbf{b}_g) \odot \sigma(\mathcal{M}^G \mathbf{V}_g + \mathbf{c}_g)) \\ \mathbf{Z}_s &= \text{MP}((\mathcal{M}^S \mathbf{W}_s + \mathbf{b}_s) \odot \sigma(\mathcal{M}^S \mathbf{V}_s + \mathbf{c}_s)) \\ \mathbf{Z}_p &= (\mathcal{Z}^p \mathbf{W}_p + \mathbf{b}_p) \odot \sigma(\mathcal{Z}_p \mathbf{V}_p + \mathbf{c}_p) \\ \mathbf{Z}_t &= (\mathcal{Z}^T \mathbf{W}_t + \mathbf{b}_t) \odot \sigma(\mathcal{Z}^T \mathbf{V}_t + \mathbf{c}_t), \end{aligned} \tag{6}$$

where $\mathbf{W} \in \mathbb{R}^{d\times d}, \mathbf{b} \in \mathbb{R}^d, \mathbf{V} \in \mathbb{R}^{d\times d}, \mathbf{c} \in \mathbb{R}^d$ are learnable parameters. $\sigma(\cdot)$ denotes the sigmoid activation function, $\odot$ denotes element-wise multiplication, and MP$(\cdot)$ denotes max pooling.

We acquire sequence features of a drug by combining RetNet and GLU with max pooling, and extract its molecular graph features via a GCN and GLU, also followed by max pooling. During training, we optimize drug multimodal representation learning using a Mean Square Error (MSE) loss. This enables the model to better preserve the unique characteristics of a drug by aligning its representations derived from both the SMILES sequence and the molecular graph. The MSE loss is formulated as follows:

$$\mathcal{L}_{\text{mse}} = \frac{1}{d}\sum_{i=1}^{d}\left(\mathbf{Z}_s^i - \mathbf{Z}_g^i\right)^2, \tag{7}$$

where $\mathbf{Z}_s^i$ and $\mathbf{Z}_g^i$ denote the $i$-th feature of the SMILES-based and graph-based drug representations, respectively. Since the multimodal protein information has been aligned through pre-training, it is unnecessary to impose constraints on protein characterization using MSE loss.

Finally, a compound–protein pair (CPP) representation can be defined. Let $\mathbf{Z}_{sp} \in \mathbb{R}^d$ represent the element-wise product between the drug SMILES representation $\mathbf{Z}_s$ and the protein representation $\mathbf{Z}_p$. $\mathbf{Z}_{gp}$, $\mathbf{Z}_{st}$, $\mathbf{Z}_{gt}$ are similar as well. The final CPP representation $\mathbf{f} \in \mathbb{R}^{4d}$ is then obtained by concatenating these two components and applying a nonlinear transformation, as defined in following equation:

$$\begin{cases} \mathbf{Z}_{sp} = \mathbf{Z}_s \odot \mathbf{Z}_p \\ \mathbf{Z}_{gp} = \mathbf{Z}_g \odot \mathbf{Z}_p \\ \mathbf{Z}_{st} = \mathbf{Z}_s \odot \mathbf{Z}_t \\ \mathbf{Z}_{gt} = \mathbf{Z}_g \odot \mathbf{Z}_t \\ \mathbf{Z_{joint}} = \tanh(\text{Concat}(\mathbf{Z}_{sp}, \mathbf{Z}_{gp}, \mathbf{Z}_{st}, \mathbf{Z}_{gt})) \end{cases} \tag{8}$$

where $Z_{joint}$ represents joint representation.

To calculate the interaction probability, we input the joint representation $Z_{joint}$ into the decoder, which consists of a fully connected classification layer followed by a sigmoid activation function:

$$p = \sigma(W_f Z_{joint} + b_{joint}), \tag{9}$$

where $W_f$ and $b_{joint}$ represent the learnable weight matrix and bias vector, respectively.

As a downstream task, the primary CPI prediction objective is to estimate the probability of molecular interactions with proteins. Meanwhile, training aims to minimize the cross entropy loss function:

$$\mathcal{L}_{\text{ce}} = -\sum_i (y_i \log(p_i) + (1-y_i)\log(1-p_i)) + \frac{\lambda}{2}\|\Theta\|_2^2. \tag{10}$$

In this context, $\Theta$ represents the collection of all learnable weight matrices and bias vectors, $y_i$ denotes the ground-truth label for the $i$-th CPI pair, $p_i$ is the predicted output probability generated by the model, and $\lambda$ is a hyperparameter controlling the L2 regularization function's strength.

During the training process, the final loss functions is:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{mse}}. \tag{11}$$

Finally, we minimize $\mathcal{L}$ to train BioText-CPI.

# Experiments

**Datasets.** To evaluate the model's effectiveness in the CPI prediction task, experiments are conducted primarily using a publicly available molecular protein dataset, BioSNAP, derived from the DrugBank database. we reconstruct the BioSNAP dataset by accessing the public protein database Uniprot to obtain its relevant biomedical description. To better evaluate the model's generalization ability, we divide the data into four subsets according to the presence of compounds and proteins in the training data: (1) Seen-Both (4,166 pairs): both compounds and proteins are seen; (2) Unseen-Comp (5,410 pairs): only proteins are seen; (3) Unseen-Prot (5,082 pairs): only compounds are observed; and (4) Unseen-Both (1,540 pairs): both molecules and proteins are not observed.

**Experimental Settings.** Our model is trained for 50 epochs with an initial learning rate of 5e-5 and a batch size of 8, using the Adam optimizer. The amino acids of proteins and their associated functional textual descriptions are encoded using 512-dimensional embeddings. Similarly, the atomic representations of drugs are embedded in a 512-dimensional space. At the fine-tuning stage, the model that achieves the highest AUROC score on the validation set is selected for evaluation on the test set. The maximum protein sequence length is the same as the maximum protein length of the ESM-2 word order, which is 1024. Proteins exceeding the maximum length are truncated. The maximum number of atoms in compounds is limited to 290. In the protein encoding module, the amino acid type $N_a$ is set to 20. Furthermore, the experiments are conducted on a system equipped with an Intel Xeon E5-2690 v3 processor and a RTX 4090 GPU for accelerated computations.

# Result

## Overall Evaluation

To evaluate the BioText-CPI model's effectiveness, we employ the AUROC and AUPRC metrics to evaluate the model's performance in CPI tasks. We conducte five independent rounds of experiments for diverse testing scenarios using distinct random seeds. Furthermore, we report each metric's mean. Based on this setup, we compare BioText-CPI with other benchmark methods, including DrugBAN (Bai et al. 2023), PerceiverCPI (Nguyen et al. 2023), MGNDTI (Peng et al. 2024a), PSC-CPI (Wu et al. 2024), SiamDTI (Zhang et al. 2024), PSRP-CPI (Zhang et al. 2025). The key observations from Table 2 are summarized as follows:

(1) **Sequence-free setting:** When only biomedical descriptions of proteins are provided as input, BioText-CPI achieves the second-best performance in terms of average AUROC across existing baselines (except PSRP-CPI). Furthermore, it is only less than 3% lower than the highest average AUROC. Notably, its performance is the best-performing method in the *Unseen-Both* scenarios across

| Method | Seen-Both | | Unseen-Compound | | Unseen-Protein | | Unseen-Both | | Avarage | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC |
| DrugBAN | 84.74 | 90.66 | 77.58 | 76.94 | 69.16 | 53.46 | 57.96 | 26.10 | 72.36 | 61.79 |
| PerceiverCPI | 82.52 | 89.74 | 80.94 | 77.96 | 53.30 | 45.68 | 52.16 | 28.50 | 67.23 | 60.47 |
| SiamDTI | <u>90.62</u> | **94.38** | 85.14 | <u>85.24</u> | 73.93 | 61.72 | 63.61 | 32.05 | 78.32 | 68.35 |
| MGNDTI | 89.87 | 93.48 | 85.18 | 83.54 | 73.08 | 60.04 | 60.08 | 27.41 | 77.05 | 66.12 |
| PSC-CPI | 86.28 | 91.23 | 80.18 | 78.96 | 61.83 | 41.55 | 57.54 | 25.05 | 71.46 | 59.20 |
| PSRP-CPI | 90.28 | <u>93.98</u> | 85.19 | 84.43 | 75.70 | 61.69 | 66.01 | 33.70 | 79.29 | 68.45 |
| ours (Seq only) | 90.33 | 92.82 | <u>85.78</u> | 84.01 | <u>79.42</u> | <u>66.05</u> | <u>68.03</u> | **38.73** | <u>80.89</u> | <u>70.40</u> |
| ours (Text only) | 89.27 | 92.14 | 84.28 | 82.76 | 75.60 | 62.58 | 67.65 | 36.60 | 79.20 | 68.52 |
| ours (Seq+Text) | **90.84** | 93.50 | **86.76** | **85.57** | **81.61** | **70.03** | **69.09** | <u>38.45</u> | **82.08** | **71.89** |

Table 2: In the four BioSNAP dataset scenarios, the predictive performance of existing SOTA methods are evaluated in comparison with BioText-CPI. The best result for each metric is marked in bold and the second-best result is underlined.

| BioText | Unseen-Protein | | Unseen-Both | |
|---|---|---|---|---|
| | AUROC | AUPRC | AUROC | AUPRC |
| Function | **75.60** | **62.58** | **67.65** | **36.60** |
| SubcellLoc. | 69.07 | 52.32 | 60.39 | 27.58 |
| Both | <u>75.17</u> | <u>61.14</u> | <u>65.88</u> | <u>33.92</u> |

Table 3: The CPI prediction results based on different biological textual inputs for proteins in the *Unseen-Protein* and *Unseen-Both* Scenarios. *Abbr.,* BioText: biomedical text; SubcellLoc.: subcellular location.

existing baselines. These results indicate that BioText-CPI is capable of approaching SOTA-level performance *without any protein sequence information*, relying solely on semantic-level descriptions.

(2) **Sequence-only setting:** When fed with only protein sequence inputs, BioText-CPI outperforms all competing methods in both AUROC and AUPRC, showcasing its strong ability to capture sequence-derived interactions even without textual context.

(3) **Multimodal setting:** When both protein sequences and biomedical texts are available, BioText-CPI consistently achieves the best performance across all benchmarks. This demonstrates the complementary nature of multimodal protein representations and highlights the effectiveness of our gated mechanism in harnessing heterogeneous information.

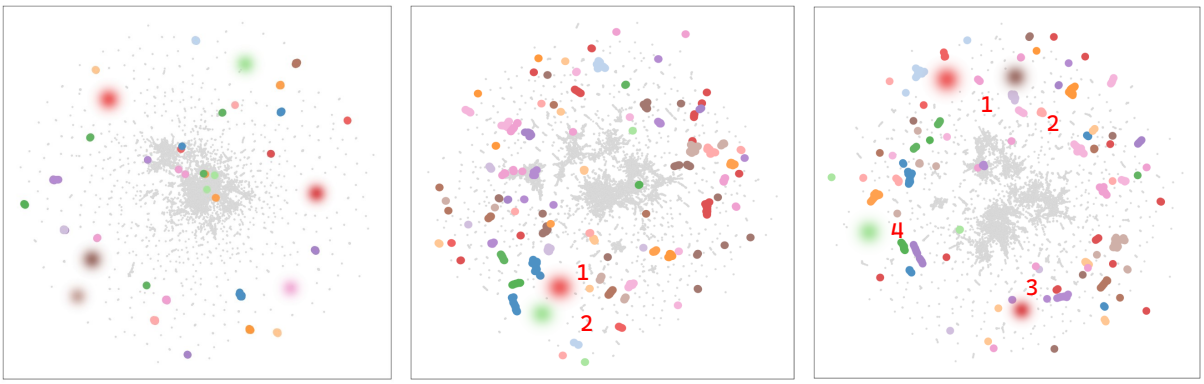
This experiment demonstrates that the sequence-free CPI prediction framework using BioText-CPI can achieve sequence-free CPI prediction and effectively reduce the sequencing cost in the early stage of drug development. Based on prevailing RNA / protein sequencing prices (Aebersold and Mann 2016; Van Dijk et al. 2018), we estimate that this approach could save approximately $100,000 to $200,000 in sequencing expenses when scaling to thousands of proteins. In addition, its feature of supporting multimodal input can meet the practical value of modal input for different proteins at different drug development stages.

| Methods | Input | UP | UB | Average |
|---|---|---|---|---|
| Full Model | Seq | **79.42** | **68.03** | **80.89** |
| | Text | **75.60** | **67.65** | **79.20** |
| | Both | **81.61** | **69.09** | **82.08** |
| w/o PMCL | Seq | 76.05 | 64.37 | 77.88 |
| | Text | 73.72 | 64.47 | <u>76.99</u> |
| | Both | 76.71 | 64.60 | <u>78.75</u> |
| w/o GM | Seq | <u>76.70</u> | <u>67.65</u> | <u>78.56</u> |
| | Text | <u>75.50</u> | 64.03 | 76.47 |
| | Both | <u>77.02</u> | 65.48 | 78.34 |
| w/o DM | Seq | 70.97 | 64.41 | 74.61 |
| | Text | 70.76 | <u>64.98</u> | 74.27 |
| | Both | 72.26 | <u>65.69</u> | 75.86 |

Table 4: Ablation study on protein multimodal contrastive learning, gating network & mse loss, and drug multimodal used for CPI prediction (measured using AUROC). The **bold** and underline highlight indicate the best and second performance under joint text–sequence input, sequence-only input, and text-only input, respectively.

## Biomedical Texts Evaluation

When protein sequences are unavailable, two alternative biomedical textual descriptors can be utilized: functional annotation and subcellular localization. To assess their individual contributions to CPI prediction, we isolate and evaluate each descriptor. As summarized in Table 3, only using functional text generates the highest AUROC and AUPRC, while the effect of jointly inputting functional text and subcellular localization information is second, and the performance relying solely on subcellular localization information is the worst. We attribute the degradation to label redundancy: many proteins share identical localization tags, injecting noise that hampers discriminative learning. Together, these findings suggest that, in sequence-free settings, functional descriptions alone suffice to achieve CPI performance comparable to SOTA sequence-based models.




(a) Text only (b) Sequence only (c) Text + Sequence

Figure 3: Latent space visualization of only protein biomedical text input (a), only protein sequence input (b), joint protein biomedical text and sequence input (c) on UniProt Database. Gradient circles are obtained by applying DBSCAN clustering to the aggregated points with a specified threshold.

## Ablation Study

We conduct ablation studies to evaluate the contribution of key components in BioText-CPI. Results are summarized in Table 4. These results validate the effectiveness of each module across all settings, including multimodal, text-only, and sequence-only inputs.

**Drug Multimodality.** Removing SMILES-based encoding (w/o Drug Multimodal) and using only molecular graphs leads to results in a notable decrease in performance in both scenarios, confirming that complementary information from multimodal drug representations enhances prediction accuracy. We omit SMILES-only variants, as prior work (Nguyen et al. 2021) has shown graph-based representations to be more effective for CPI prediction.

**Gating Network & MSE Loss.** Replacing the gating mechanism (Gu et al. 2020) with simple concatenation and remove the MSE loss at the same time (w/o GM) results in degraded performance, indicating the importance of dynamic feature selection and feature alignment of compounds when fusing heterogeneous modalities.

**Protein Contrastive Learning.** Removing multimodal contrastive pretraining (w/o PMCL) and using only masked modeling reduces performance, highlighting that contrastive alignment of sequence and text modalities promotes more robust and transferable protein representations.

## Visualization of Protein Representations

To qualitatively assess the quality of learned protein representations, we visualize the latent space of BioText-CPI using UMAP (McInnes, Healy, and Melville 2018). We extract all protein embeddings from the UniProt database and project them into 2D space. Following (Akdel et al. 2022), the 20 most frequent protein families are color-coded for comparison across input conditions: text-only, sequence-only, and multimodal. In addition, we apply DBSCAN (Schubert et al. 2017) clustering with a distance threshold of 3 and a minimum cluster size of 300. We can directly observe the clustering effects of different representations by the number of gradient circles. Results are shown in Figure 3, with the following key observations: **Text-only input.** Proteins are effectively clustered according to their families, suggesting that the BLM encoder, trained via contrastive learning, captures high-level functional semantics. However, such clustering does not translate into strong CPI prediction, likely because functional text alone lacks sufficient structural detail for modeling molecular interactions. **Sequence-only vs. multimodal.** Sequence-only representations yield less distinct clustering compared to multimodal inputs. The multimodal setting enhances family separation, highlighting the complementary role of biomedical text. These observations are consistent with quantitative results, confirming the advantage of integrating both modalities in CPI prediction.

## Conclusion

In this work, we propose a novel *sequence-free* framework for CPI prediction, which leverages pretraining via multimodal contrastive learning between biomedical text and protein functions. This design enables our method to be applicable in early-stage drug discovery scenarios where protein sequence information is unavailable. Extensive experiments demonstrate that BioText-CPI, when using only protein function descriptions as input, achieves performance competitive with leading sequence-based CPI prediction methods. This validates the effectiveness of our approach in scenarios where sequencing is infeasible or costly, offering a promising direction to reduce experimental overhead while maintaining virtual screening accuracy. Furthermore, when protein sequences are available, BioText-CPI can leverage its multimodal representation and gated fusion architecture to achieve superior performance, especially under challenging *Unseen-Protein* and *Unseen-Both* settings.

Despite the promising results, several limitations remain. We have not explored the computational efficiency yet. Additionally, we leave to future work a systematic investigation of the impact of functional description quality and quantity on predictive performance.

## Acknowledgments

This work was supported in part by the Natural Science Foundation of China (No. 62476203), Key Project of Traditional Chinese Medicine Joint Fund of Hubei Provincial Natural Science Foundation (No.2025AFD47), Hubei Province Science and Technology Innovation Plan Project (No.2025BCB035), the Guangdong Provincial Natural Science Foundation General Project (No. 2025A1515012155), the Shenzhen Natural Science Foundation Project (No. JCYJ20250604122534006).

## References

Aebersold, R.; and Mann, M. 2016. Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620): 347–355.

Ahmed, E.; Heinzinger, M.; Dallago, C.; Rihawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Martin, S.; et al. 2020. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *bioRxiv*.

Akdel, M.; Pires, D. E.; Pardo, E. P.; Jänes, J.; Zalevsky, A. O.; Mészáros, B.; Bryant, P.; Good, L. L.; Laskowski, R. A.; Pozzati, G.; et al. 2022. A structural biology community assessment of AlphaFold2 applications. *Nature Structural & Molecular Biology*, 29(11): 1056–1067.

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.

Bian, J.; Zhang, X.; Zhang, X.; Xu, D.; and Wang, G. 2023. MCANet: shared-weight-based MultiheadCrossAttention network for drug–target interaction prediction. *Briefings in Bioinformatics*, 24(2): bbad082.

Chu, X.; Jaeger, M.; Beumer, J.; Bakker, O. B.; Aguirre-Gamboa, R.; Oosting, M.; Smeekens, S. P.; Moorlag, S.; Mourits, V. P.; Koeken, V. A.; et al. 2021. Integration of metabolomics, genomics, and immune phenotypes reveals the causal roles of metabolites in disease. *Genome biology*, 22(1): 198.

Consortium, U. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1): D506–D515.

Dalkıran, A.; Atakan, A.; Rifaioğlu, A. S.; Martin, M. J.; Atalay, R. Ç.; Acar, A. C.; Doğan, T.; and Atalay, V. 2023. Transfer learning for drug–target interaction prediction. *Bioinformatics*, 39(Supplement_1): i103–i110.

Dauphin, Y. N.; Fan, A.; Auli, M.; and Grangier, D. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*, 933–941. PMLR.

Gu, A.; Gulcehre, C.; Paine, T.; Hoffman, M.; and Pascanu, R. 2020. Improving the gating mechanism of recurrent neural networks. In *International conference on machine learning*, 3800–3809. PMLR.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23.

Hughes, R. E.; Elliott, R. J.; Dawson, J. C.; and Carragher, N. O. 2021. High-content phenotypic and pathway profiling to advance drug discovery in diseases of unmet need. *Cell Chemical Biology*, 28(3): 338–355.

Kang, H.; Goo, S.; Lee, H.; Chae, J.-w.; Yun, H.-y.; and Jung, S. 2022. Fine-tuning of BERT model to accurately predict drug–target interactions. *Pharmaceutics*, 14(8): 1710.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

LeCun, Y.; Bengio, Y.; et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10): 1995.

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.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Lehmann, S.; Atika, B.; Grossmann, D.; Schmitt-Engel, C.; Strohlein, N.; Majumdar, U.; Richter, T.; Weißkopf, M.; Ansari, S.; Teuscher, M.; et al. 2022. Phenotypic screen and transcriptomics approach complement each other in functional genomics of defensive stink gland physiology. *BMC genomics*, 23(1): 608.

Li, K.; Liu, W.; Luo, Y.; Cai, X.; Wu, J.; and Hu, W. 2024. Zero-shot learning for preclinical drug screening. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2117–2125.

Li, K.; Xiong, Y.; Zhang, H.; Cai, X.; Wu, J.; Du, B.; and Hu, W. 2025a. Graph-Structured Small Molecule Drug Discovery Through Deep Learning: Progress, Challenges, and Opportunities. *2025 IEEE International Conference on Web Services (ICWS)*, 1033–1042.

Li, K.; Zeng, Y.; Xiong, Y.-d.; Wu, H.-c.; Fang, S.; Qu, Z.-y.; Zhu, Y.; Du, B.; Gao, Z.-b.; and Hu, W.-b. 2025b. Contrastive learning-based drug screening model for GluN1/GluN3A inhibitors. *Acta Pharmacologica Sinica*, 1–13.

Li, T.; Fang, Z.; Zhang, X.; Tang, K.; Chen, H.; Jiang, Z.; Zhao, T.; Xu, R.; Cheng, F.; Li, X.; et al. 2025c. DrugLM: A Unified Framework to Enhance Drug-Target Interaction Predictions by Incorporating Textual Embeddings via Language Models. *bioRxiv*, 2025–07.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; 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*, 2022: 500902.

Liu, M.; Meng, X.; Mao, Y.; Li, H.; and Liu, J. 2024. ReduMixDTI: prediction of drug–target interaction with feature redundancy reduction and interpretable attention mech-

anism. *Journal of Chemical Information and Modeling*, 64(23): 8952–8962.

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.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; and Rives, A. 2021. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34: 29287–29303.

Nguyen, N.-Q.; Jang, G.; Kim, H.; and Kang, J. 2023. Perceiver CPI: a nested cross-attention network for compound–protein interaction prediction. *Bioinformatics*, 39(1): btac731.

Nguyen, T. M.; Nguyen, T.; Le, T. M.; and Tran, T. 2021. Gefa: early fusion approach in drug-target affinity prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(2): 718–728.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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

Öztürk, H.; Ozkirimli, E.; and Özgür, A. 2019. WideDTA: prediction of drug-target binding affinity. *arXiv preprint arXiv:1902.04166*.

Peng, J.; Wang, Y.; Guan, J.; Li, J.; Han, R.; Hao, J.; Wei, Z.; and Shang, X. 2021. An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction. *Briefings in bioinformatics*, 22(5): bbaa430.

Peng, L.; Liu, X.; Chen, M.; Liao, W.; Mao, J.; and Zhou, L. 2024a. MGNDTI: A Drug-Target Interaction Prediction Framework Based on Multimodal Representation Learning and the Gating Mechanism. *Journal of Chemical Information and Modeling*, 64(16): 6684–6698.

Peng, L.; Liu, X.; Yang, L.; Liu, L.; Bai, Z.; Chen, M.; Lu, X.; and Nie, L. 2024b. BINDTI: a bi-directional intention network for drug-target interaction identification based on attention mechanisms. *IEEE Journal of Biomedical and Health Informatics*, 29(3): 1602–1612.

Qiu, S.; Cai, Y.; Yao, H.; Lin, C.; Xie, Y.; Tang, S.; and Zhang, A. 2023. Small molecule metabolites: discovery of biomarkers and therapeutic targets. *Signal Transduction and Targeted Therapy*, 8(1): 132.

Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118.

Schubert, E.; Sander, J.; Ester, M.; Kriegel, H. P.; and Xu, X. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, 42(3): 1–21.

Sun, Y.; Dong, L.; Huang, S.; Ma, S.; Xia, Y.; Xue, J.; Wang, J.; and Wei, F. 2023. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*.

Torng, W.; and Altman, R. B. 2019. Graph convolutional neural networks for predicting drug-target interactions. *Journal of chemical information and modeling*, 59(10): 4131–4149.

Van Dijk, E. L.; Jaszczyszyn, Y.; Naquin, D.; and Thermes, C. 2018. The third revolution in sequencing technology. *Trends in Genetics*, 34(9): 666–681.

Wu, L.; Huang, Y.; Tan, C.; Gao, Z.; Hu, B.; Lin, H.; Liu, Z.; and Li, S. Z. 2024. Psc-cpi: Multi-scale protein sequence-structure contrasting for efficient and generalizable compound-protein interaction prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 310–319.

Wu, L.; Lin, H.; Huang, Y.; and Li, S. Z. 2022. Knowledge distillation improves graph structure augmentation for graph neural networks. *Advances in Neural Information Processing Systems*, 35: 11815–11827.

Xu, M.; Yuan, X.; Miret, S.; and Tang, J. 2023. Protst: Multimodality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, 38749–38767. PMLR.

Yang, X.; Yang, G.; and Chu, J. 2024. GraphCL-DTA: a graph contrastive learning with molecular semantics for drug-target binding affinity prediction. *IEEE Journal of Biomedical and Health Informatics*, 28(8): 4544–4552.

Yue, L.; Xing, S.; Lu, Y.; and Fu, T. 2024. Biomamba: A pre-trained biomedical language representation model leveraging mamba. *arXiv preprint arXiv:2408.02600*.

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

Zhang, H.; Liu, Z.; Meng, K.; Chen, J.; Wu, J.; Du, B.; Lin, D.; Che, Y.; and Hu, W. 2025. Zero-Shot Learning with Subsequence Reordering Pretraining for Compound-Protein Interaction. arXiv:2507.20925.

Zhang, P.; Wei, Z.; Che, C.; and Jin, B. 2022. DeepMGT-DTI: Transformer network incorporating multilayer graph information for Drug–Target interaction prediction. *Computers in biology and medicine*, 142: 105214.

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

Zitnik, M.; Sosič, R.; Maheshwari, S.; and Leskovec, J. 2018. BioSNAP Datasets: Stanford Biomedical Network Dataset Collection. http://snap.stanford.edu/biodata.