



## Full length article

# Improving protein–protein interaction modulator predictions via knowledge-fused language models

Zitong Zhang<sup>a,b</sup>, Quan Zou<sup>b,c</sup>, Chunyu Wang<sup>a</sup>, Junjie Wang<sup>d,\*</sup>, Lingling Zhao<sup>a,\*</sup>

<sup>a</sup> Faculty of Computing, Harbin Institute of Technology, Harbin, 150001, China

<sup>b</sup> Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, 324003, China

<sup>c</sup> Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, 610054, China

<sup>d</sup> Department of Medical Informatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, 211166, China

## ARTICLE INFO

## Keywords:

Protein-protein interaction modulator

Knowledge fusion

Language model

Gene ontology

Pretrain

## ABSTRACT

Protein-protein interactions (PPIs) play key roles in numerous biological processes and their dysregulation can lead to various human diseases. Modulating these interactions with small molecule PPI modulators has emerged as a promising strategy for treating such diseases. However, current computational approaches for screening PPI modulators often fail to integrate biomolecular expertise and lack the elucidation of interaction mechanisms. Here, we propose a knowledge-fused modulator-PPI interaction prediction method (KFPPIMI) to alleviate these issues. KFPPIMI constructs separate representation models for modulators and proteins, each of which integrates external knowledge from textual and graph-based data sources via a language modeling framework. The fusion of the nuanced expression of natural language with the structural attributes of biomolecules provides KFPPIMI with a holistic view of modulator-PPI interactions. Extensive experiments are conducted to evaluate the effectiveness of KFPPIMI and its individual components. The results show that KFPPIMI outperforms existing methods in different scenarios. Moreover, the modulator and protein representation model can be successfully applied to their respective downstream tasks with comparable performance.

## 1. Introduction

Protein-protein interactions (PPIs) are fundamental to numerous biological processes, including signal transduction, immune regulation, and cell division [1,2]. Dysregulation of PPIs, whether through imbalance, under-expression, or overexpression, is responsible for the pathogenesis of various diseases such as cancer and neurodegenerative disorders [3–7]. As a result, modulating these interactions with small molecules has become a promising therapeutic strategy for treating related diseases. The growing landscape of approved and developing PPI modulators has prompted efforts to improve computational methods for identifying these modulators [8–12].

Molecular docking and molecular dynamics simulations are two widely used computational approaches for screening drug candidates for specific targets [13–18]. However, the high dependence on computational resources and protein structure limits their feasibility for large-scale testing [19]. Moreover, even if the docking is successful, it is only for single target proteins and cannot be applied directly to identify modulators of protein complexes. In recent years, the increasing number of high-quality experimentally validated PPI targets and their corresponding small molecule modulators has opened up avenues

for machine learning-based modulator prediction methods. Depending on whether these methods incorporate target information, they can be divided into target-free and target-based methods.

Target-free methods [20–25] analyze the intrinsic properties and structural features of modulators without relying on prior knowledge of specific PPI targets. For example, SMMPPPI [26] adopts a two-stage prediction strategy based on Morgan fingerprints and Random Forest algorithm to first screen potential PPI inhibitors from a large compound library and then predict their inhibitory activities for 11 PPI families in the second stage. pdCSM-PPI [27] integrates graph-based signatures and feature engineering to improve inhibitor screening for different PPI targets. The recent HiGPPIM [28] provides new insights into modulator discovery by introducing a hierarchical graph neural network (GNN) based on functional group information. These target-free methods are useful for the preliminary screening of large-scale chemical libraries, but their inability to elucidate interaction mechanisms limits the precise design and optimization of modulators.

Target-based methods formulate the PPI modulator prediction problem as a binary classification task with the goal of determining whether there is an interaction between a modulator and a specific target

\* Corresponding authors.

E-mail addresses: [junjie2021@njmu.edu.cn](mailto:junjie2021@njmu.edu.cn) (J. Wang), [zhaoll@hit.edu.cn](mailto:zhaoll@hit.edu.cn) (L. Zhao).

<https://doi.org/10.1016/j.infus.2025.103227>

Received 22 February 2025; Received in revised form 15 March 2025; Accepted 16 April 2025

Available online 26 April 2025

1566-2535/© 2025 Elsevier B.V. All rights reserved, including those for text and data mining, AI training, and similar technologies.

(PPIMI). For this kind of method, only MultiPPIMI [29] and Yaseen et al.'s approach (hereafter referred to as YASEEN) [30] are reported. MultiPPIMI integrates structural and physicochemical property embeddings of modulators and PPI targets based on GraphMVP and ESM2 models and uses a bilinear attention network to capture intermolecular interactions. YASEEN utilizes fingerprint features of compounds and GNN-extracted structural features of protein complexes to predict the interaction potential of a given compound-PPI complex pair. MultiPPIMI surpasses YASEEN in terms of generalization performance and computational efficiency, but further enhancements are needed for biomolecular representation.

Although target-based PPI modulator prediction research is still in its infancy, the closely related PPI prediction research has flourished over the years [31,32] and two main methodological frameworks have been established, i.e., the network-based method and the integrated method [33–40]. Network-based methods usually formalize the PPI prediction problem as a link prediction problem and calculate the binding probability based on the topological similarity between protein nodes in the PPI network. Integrated methods focus on extracting homogeneous features from proteomics and genomics data, which will be integrated into popular classifiers to accomplish prediction tasks. With the development of artificial intelligence, PPI prediction methods have undergone a paradigm shift from machine learning to deep learning and have been continuously optimized in data processing, feature extraction, and network architecture. These advancements provide a solid theoretical foundation and technical support for the screening and rational design of PPI modulators.

Small molecules and proteins are two essential bioentities involved in the PPI modulator discovery process [41–43]. Computational analysis of properties and interaction mechanisms relies heavily on robust and expressive representations of entities. Common forms include biological sequences, molecular graphs, and experimentally verified or predicted three-dimensional structures [44,45]. Language model-based approaches such as ChemBERTa [46] and ESM [47] have been successfully used to capture sequence properties of biomolecules. Similarly, graph model-based approaches such as GraphMVP [48], and GearNet [49] have demonstrated strong performance, particularly in modeling functional properties of molecular graphs or geometric structures.

While powerful for extracting intrinsic features of biomolecules from different perspectives, these methods often overlook rich external knowledge that can enhance the depth and breadth of molecular understanding. In the biomedical field, scientific literature, biological databases, and knowledge graphs serve as crucial knowledge sources, providing detailed textual descriptions of the properties, functions, and interactions of various biological molecules. These cannot be directly inferred from molecular or protein sequences alone. Accordingly, a popular trend involves jointly modeling biomolecules and natural language to enhance molecule or protein representation with external knowledge. For instance, Molt5 [50] pretrains the T5 architecture on a large number of SMILES strings and textual descriptions to establish translations between molecules and natural language. ProtST [51] learns a joint representation of protein sequences and biomedical text by designing three types of tasks, i.e. unimodal mask prediction, multimodal representation alignment, and multimodal mask prediction tasks.

However, directly applying state-of-the-art molecular and protein representation models to PPIMI prediction tasks has limitations. On the one hand, high-quality textual descriptions of PPI modulators are difficult to retrieve from public knowledge repositories in a similar strategy to existing models. On the other hand, most models represent proteins solely based on brief knowledge at the sequence level, which cannot adequately capture the biological facts related to their functions. But this is exactly what is needed for PPIMI prediction. In general, determining whether a protein participates in a biological process, such as interactions with other biomolecules, relies on a deep understanding

of its functional properties and signaling pathways [52]. Even if two proteins exhibit a high degree of sequence identity, their functions may still diverge significantly [53]. In contrast, Gene Ontology (GO) provides comprehensive and standardized protein functional annotations, covering three categories: molecular function (MF), cellular component (CC), and biological process (BP). GO terms are constructed as a directed acyclic graph (DAG) to describe their hierarchical relationships. OntoProtein [54] and KeAP [55] are two prominent models that leverage GO to capture protein properties. However, they both model knowledge graphs through mere triples (Protein, Relation, GO), neglecting contextual dependencies with terms as the smallest unit. Besides, several specialized GO term representation methods [56,57] also fail to effectively fuse structural and semantic information.

To address these issues, we propose KFPPIMI, a PPIMI prediction framework that integrates independent modulator and protein representation models (Fig. 1). By combining external knowledge with language modeling architectures, KFPPIMI can improve the expressiveness of biomolecular properties. Specifically, for the modulator model, we first collect SMILES sequences and textual descriptions from the PubChem database and enrich them via GPT-3.5 to build a large-scale corpus. We then pretrain a RoBERTa model on this corpus as the backbone encoder. Due to the lack of detailed descriptions for many modulator molecules in PubChem and the potential inaccuracies in the generation of large language models, we additionally introduce a fingerprint-based structure encoder to calibrate the knowledge only provided by GPT-3.5. For the protein model, we fuse factual knowledge derived from both GO graphs and GO term annotations. The structural information of each GO term is injected into the BERT-based language model through a special token and processed together with the textual annotations. Furthermore, the direct neighbor prediction and sub-ontology member prediction tasks are designed to enhance the language model's understanding of the basic semantics and contextual relationships of terms. Finally, we apply these two representation models to the PPIMI prediction task, in which modulator and protein representations are integrated via an attention-based fusion block. We systematically evaluate the performance of KFPPIMI and its individual components. The experimental results show that the proposed KFPPIMI can improve the prediction of modulator-PPI interactions and provide new possibilities for future research and practical applications.

## 2. Method

### 2.1. Modulator model

The modulator model in KFPPIMI consists of two distinct submodules: a backbone text encoder and a structure encoder for knowledge calibration. They operate synergistically to generate comprehensive and robust modulator representations.

#### 2.1.1. Backbone text encoder

The text encoder is implemented based on a RoBERTa architecture [58] for processing molecular SMILES sequences and textual descriptions. These text descriptions are retrieved from PubChem and enriched by GPT-3.5, which provides detailed information about the chemical properties and functional groups of the molecules.

The original RoBERTa tokenizer is derived from natural language, and applying it directly to SMILES strings may overlook meaningful chemical structures. To efficiently handle SMILES strings and domain-specific vocabulary, we retrain the tokenizer using Byte Level Byte-Pair Encoding. The definition of special tokens is consistent with RoBERTa, and [SEP] is added as a precise boundary between SMILES sequences and textual descriptions to help the model understand input texts. Then, we pretrain the text encoder using the masked language modeling (MLM) task, where each token is randomly masked with a probability of 15% and reconstructed based on the context. The loss function is:

$$\mathcal{L}_{MLM} = - \sum_{i \in T} \log P(x_i | \bar{X}) \quad (1)$$

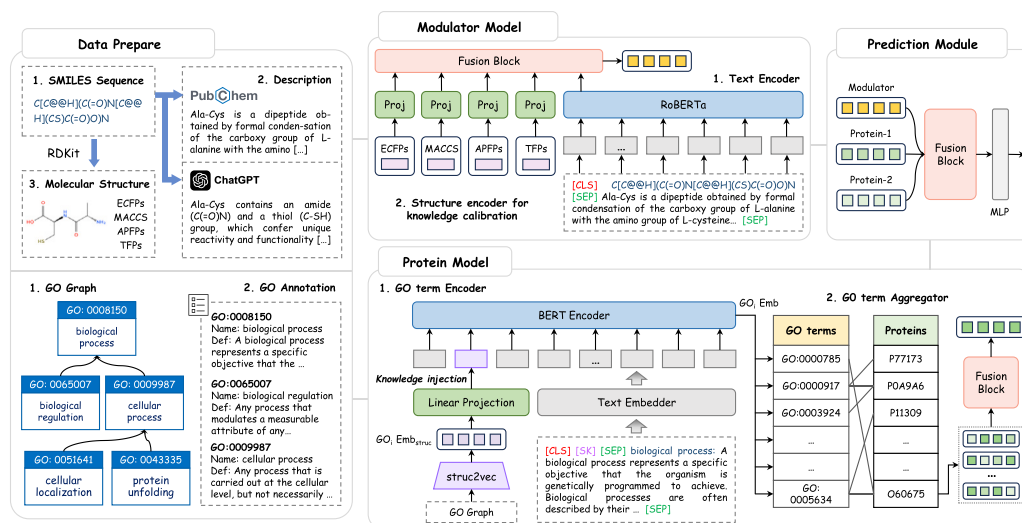


Fig. 1. The overall workflow of KFPPIMI.

where  $T$  is a random set of masked positions, and  $x_i$  and  $\tilde{x}$  denote the  $i$ th token of the input text and the masked input, respectively. After pretraining, the text encoder can capture the correlation between molecular structure and chemical properties to provide knowledge support for various downstream tasks.

### 2.1.2. Structure encoder for knowledge calibration

As mentioned above, the backbone text encoder extracts molecular information mainly from textual descriptions generated by GPT-3.5. However, as a general-purpose large language model, GPT is not immune to occasional inaccuracies, which may compromise the reliability of structure–activity analyses. To alleviate this problem, we introduce a fingerprint-based structure encoder for knowledge calibration during the execution of downstream tasks. Specifically, we employ different types of molecular fingerprints to represent various aspects of the molecular structure. The fingerprints utilized include Extended-Connectivity Fingerprints (ECFPs), which capture the local atomic environment; Molecular ACCess System (MACCS) fingerprints, which encode predefined substructures; Atom-Pair Fingerprints (APFPs), which represent atom pairs and their topological distances; and Topological Fingerprints (TFPs), which encode molecular graph-based features. We then align each type of molecular fingerprint with the embedding space of the language model using separate linear projection heads to ensure efficient integration of structural and textual information.

The final molecular representations are constructed by concatenating five molecular embeddings and feeding them into an elaborated fusion block (Section 2.3). By incorporating explicit fingerprint knowledge, the modulator model can bridge potential discrepancies between molecular real structure and textual description, avoiding being misled by ambiguous or incomplete information.

### 2.2. Protein model

The proposed protein model consists of two stacked modules: (1) GO term encoder injects knowledge graph information into the language model to generate rich GO term embeddings; and (2) GO term aggregator jointly models relevant GO term embeddings to generate protein representations. We will introduce the implementation of these two modules in the following subsections.

#### 2.2.1. GO term encoder

**Architecture.** Our aim is to integrate the GO graph and text annotations into a unified architecture to build deep representation interactions. To this end, we utilize the BERT model [59] as a backbone

network and inject the extracted structural information of GO terms into its textual representation for joint modeling. The model weights are initialized by ouBioBERT [60], which is pretrained on a large scientific corpus spanning disciplines like biology and medicine.

We train the graph embedding model struc2vec [61] on the entire GO graph to extract contextual information implied in GO terms and their interconnections. Compared with DeepWalk [62] and node2vec [63], struc2vec shows superior performance in measuring the structural similarity and hierarchical relationships of different nodes, which is more in line with the modeling requirements of GO graphs. We use the trained struc2vec to generate embedding vectors for each term node and inject them into the backbone model as structural knowledge to enhance GO term representation. To bridge the gap between GO structure and annotation text, struc2vec embeddings are projected into text space using affine transformation before injection.

Given a GO term, we first concatenate its name and definition to form a text sequence that provides a detailed description of protein function. Following BERT, we tokenize the GO text sequences by WordPiece and add special tokens [CLS], [PAD], and [SEP]. To inject structural knowledge into the GO term representation framework, we additionally define a special token [SK] to act as a placeholder, whose embedding will be replaced by the struc2vec-encoded structural features for improving GO representations in subsequent information dissemination. The interleaved GO term input can be formalized as  $\{[CLS] [SK] [SEP] \text{ Name} + \text{Def} [PAD] \dots [PAD] [SEP]\}$ .

With clever knowledge injection, our model not only captures explicit hierarchical relationships between GO terms but also deeply models implicit semantic correlations. The contextual information of all terms is captured through a one-time embedding and delegated to the BERT module for subsequent processing, replacing the setup of a separate graph embedder in dual-stream methods. This design can significantly reduce model parameters and running time because the GO graph is massive in scale. After pretraining, the architecture can be applied to various protein tasks with minimal modifications.

**Multi-task training.** To enhance the joint representation of semantics and structure, we add different prediction heads to the top of the GO encoder and train the whole architecture through multi-task learning. The designed training tasks include: (1) direct neighbor prediction of GO terms, where direct neighbors are defined as the union of children and parent terms; and (2) sub-ontology membership prediction of GO terms, where sub-ontology categories include CC, MF, and BP. We argue that the two prediction tasks are related with the goal of encouraging the model to effectively integrate structural features with language representations while preserving their properties. Both

use cross-entropy as the loss function and are jointly optimized by adding different classification heads:

$$\mathcal{L}_{GO} = - \sum_i^N y_i^{nbr} \log(\hat{y}_i^{nbr}) - \sum_j^3 y_j^s \log(\hat{y}_j^s) \quad (2)$$

where  $y^{nbr} \in \{0, 1\}^N$  denotes the actual direct neighbors of the input term,  $y^s \in \{0, 1\}^3$  denotes the actual category of the sub-ontology,  $\hat{y}^{nbr}$  and  $\hat{y}^s$  are prediction vectors, and  $N$  is the total number of GO terms.

### 2.2.2. GO term aggregator

In our approach, protein representations are created through GO term embeddings. We retrieve the set of GO terms  $S = \{GO_1, GO_2, \dots, GO_n\}$  annotated to a protein, where  $GO_i$  corresponds to a single GO term associated with the protein. Then, we use the trained GO term encoder to infer the embedding of each term in the set  $S$  and assemble them to construct the protein embedding  $E = \{Emb(GO_1), Emb(GO_2), \dots, Emb(GO_n)\}$ .

The specificity (importance) of GO terms is influenced by several factors, especially the cellular processes associated with protein interactions. Consequently, we further input  $E$  into the feature fusion block to generate a comprehensive representation of the protein in the context of its functional annotation terms. The fusion block is constructed based on the attention mechanism that dynamically evaluates the contributions of GO terms and weights their embeddings.

### 2.3. Fusion block

In this section, we introduce a general fusion block to effectively address the need for repeated feature fusion operations within the KFPPIMI framework. As illustrated in Fig. 2, the fusion block is built upon the standard Transformer encoder architecture and enhanced with the SwiGLU feedforward network (FFN) [64]. Specifically, it mainly contains a multi-head self-attention (MSA) layer, a SwiGLU FFN layer, and a fully connected (FC) layer. Layer norm (LN) and residual connections are applied between each layer. Given  $M$  vectors to be fused, we combine them into an input feature  $z \in \mathbb{R}^{M \times D}$ . Then, the fusion process can be formalized as:

$$\begin{aligned} z' &= MSA(LN(z)) + z, \\ z'' &= FFN_{SwiGLU}(LN(z')) + z', \\ z^{out} &= FC(z''), \end{aligned} \quad (3)$$

where

$$FFN_{SwiGLU}(x, W, U, V) = (Swish(xW) \odot xU)V, \quad (4)$$

$W$ ,  $U$ , and  $V$  are weight matrices,  $\odot$  is the element-wise product,  $Swish(\cdot)$  is the Swish activation function, and  $Swish(x) = x * Sigmoid(x)$ .

We reuse this fusion block in the modulator model, protein model, and later prediction module to capture the complex interactions and dependencies between input features.

### 2.4. Prediction module

In the last section, we perform PPIMI prediction tasks using a two-layer MLP integrated with a SwiGLU FFN layer. We concatenate modulator features and partner protein features, and feed them into the fusion block to generate interaction features. The interaction features are subsequently passed through the MLP variant to produce final predictions. The training pipeline of KFPPIMI is summarized in Fig. 3. We use binary cross-entropy as the loss function for PPIMI prediction, calculated as follows:

$$\mathcal{L} = -\frac{1}{\Gamma} \sum_{i=1}^{\Gamma} [y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i))], \quad (5)$$

where  $\Gamma$  is the number of training samples,  $\sigma$  is the sigmoid function,  $\hat{y}$  denotes the predicted score, and  $y$  denotes the actual label.

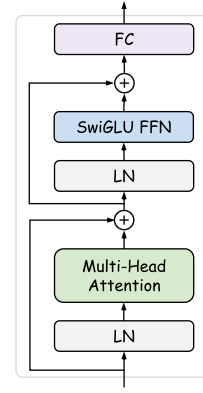


Fig. 2. An illustration of the fusion block.

## 3. Experiment setup

### 3.1. Data preparation

**Molecular data.** We begin by collecting 301,566 molecular SMILES sequences and corresponding textual descriptions from PubChem to build a pretraining corpus for the modulator model, carefully excluding all known PPI modulators to avoid potential data leakage. We note that the textual descriptions provided in PubChem are generally limited in scope and detail. Furthermore, for the majority (94.82%) of modulators in the PPIMI dataset, the PubChem descriptions are either inadequate or not retrievable. To overcome this limitation, we use GPT-3.5 to enrich these molecular descriptions to obtain detailed information about the chemical properties and functional groups. The modulator data in PPIMI datasets are curated with the same strategy.

**GO data.** We use the latest Gene Ontology released in September 2024 (<https://geneontology.org/docs/download-ontology/>). Following the practice in [56], we remove GO terms marked as “obsolete”, resulting in a total of 44,261 terms with 8888 BP terms, 11,177 CC terms, and 4196 MF terms. All GO terms and the set of inclusion relationships are organized as a DAG, where each term is labeled with its direct neighbors and sub-ontology memberships.

**PPIMI data.** We extract the interactions between small molecule modulators and PPI targets from the public DLiP database [65] to construct the PPIMI benchmark dataset. We use a strict filtering procedure [29] and further remove data entries for illegal PPI targets and non-human species. After this processing, a total of 11,145 PPIMI tuples are generated, including 9343 small molecule modulators and 117 PPI targets. These verified PPIMI tuples are regarded as positive samples. Consequently, negative samples are generated by replacing modulators or PPI targets in positive samples. We control the ratio of positive and negative samples to 1:1 and ensure that the sampled negative samples are not present in the positive samples for effective training.

To systematically evaluate the generalization ability of models, we use four data split schemes, including one transductive setting (S1) and three inductive settings (S2–S4):

- S1 (random): samples are randomly divided into training and test sets.
- S2 (new modulator-old PPI target): modulators in the test set do not appear in the training set.
- S3 (old modulator-new PPI target): PPI targets in the test set do not appear in the training set.
- S4 (new modulator-new PPI target): modulators and PPIs in the test set are completely invisible during training.

For each split scheme, the ratio of training, validation, and test sets is approximately 8:1:1.



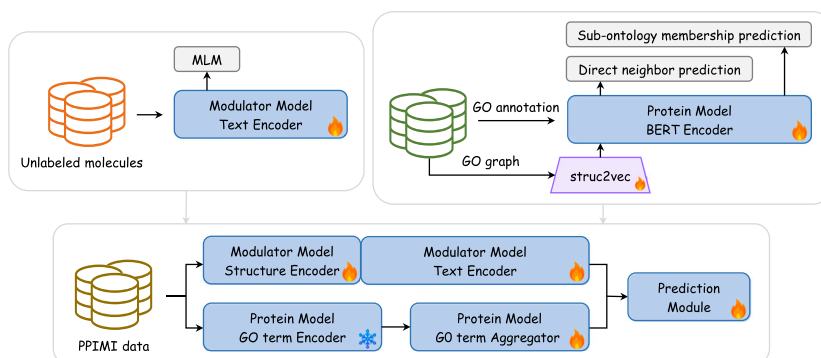


Fig. 3. The training pipeline of KFPPIMI.

Table 1

Performance evaluation of KFPPIMI and other methods in different experimental settings.

| Method        | Transductive setting |              | Inductive setting |              |              |              |              |              |
|---------------|----------------------|--------------|-------------------|--------------|--------------|--------------|--------------|--------------|
|               | S1                   |              | S2                |              | S3           |              | S4           |              |
|               | AUROC                | AUPR         | AUROC             | AUPR         | AUROC        | AUPR         | AUROC        | AUPR         |
| Kernel-based  | 0.950                | 0.933        | 0.888             | 0.865        | 0.529        | 0.601        | 0.559        | <b>0.642</b> |
| GearNet-based | 0.975                | 0.973        | 0.924             | 0.914        | 0.788        | 0.800        | 0.456        | 0.449        |
| YASEEN        | 0.970                | 0.965        | 0.925             | 0.911        | 0.638        | 0.681        | 0.432        | 0.464        |
| MultiPPIMI    | 0.988                | 0.988        | 0.968             | 0.967        | 0.789        | 0.808        | 0.579        | 0.566        |
| KFPPIMI       | <b>0.990</b>         | <b>0.989</b> | <b>0.972</b>      | <b>0.969</b> | <b>0.832</b> | <b>0.844</b> | <b>0.647</b> | 0.613        |

### 3.2. Implementation detail

For the modulator model, we pretrain the text encoder using the RoBERTa procedure for 100 epochs. The maximum vocabulary size, maximum sequence length, hidden size, number of attention heads, and batch size are set to 52K, 512, 768, 6, and 32, respectively.

For the protein model, we first train the struc2vec model on the entire GO graph with an embedding dimension of 128, a walk length of 10, and a number of walks of 80. Using these struc2vec embeddings, we then train the GO term encoder with the AdamW optimizer, setting the learning rate to  $1e-5$ , weight decay to 0.01, and batch size to 64.

Finally, we train the whole KFPPIMI model using AdamW with a learning rate of  $1e-4$  and a batch size of 64. During training, the text encoder of the moderator model is fine-tuned and the GO term encoder of the protein model is frozen.

All experiments are run on 1 NVIDIA GeForce RTX 4090 GPU. The area under the receiver operating characteristic curve (AUROC) and the area under the precise-recall curve (AUPR) are utilized to report model performance, with higher values indicating better performance.

## 4. Results and discussion

### 4.1. Performance evaluation

We compare KFPPIMI with four state-of-the-art methods on the benchmark dataset. The details of these methods are listed below:

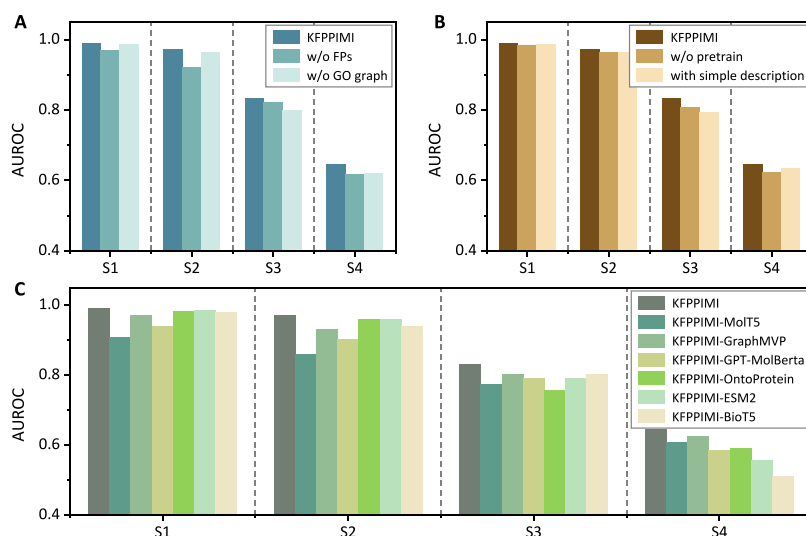
- **MultiPPIMI** integrates GraphMVP features of modulators and ESM2 features of PPI targets and employs a bilinear attention network to capture biomolecular interactions.
- **YASEEN** uses ECFP features of compounds and GNN-extracted structural features of protein complexes to predict PPIMI.
- **Kernel-based** method combines nonlinear radial basis function (RBF) similarity kernel representations of compounds and protein complexes and passes them to a support vector machine for PPIMI classification.

- **GearNet-based** method introduces GearNet (Geometry-Aware Relational Graph Neural Network) to extract the structural features of protein complexes and then combines them with ECFPs features of compounds to jointly predict PPIMI.

To ensure the reliability and comprehensiveness of our evaluation, protein complex structure inputs for both the YASEEN and GearNet-based methods are generated using AlphaFold3 [66], an advanced protein structure prediction tool. Table 1 presents the comparison results, with the best values highlighted in bold.

For the inductive setting, the modulators and PPI targets used during training can also appear in the test set. In this scenario, all methods show excellent performance, with AUROC and AUPR values exceeding 0.9. Nevertheless, KFPPIMI still achieves further improvements, with AUROC and AUPR as high as 0.99 and 0.989, respectively. This result demonstrates the effectiveness and robustness of KFPPIMI in PPIMI prediction. Unlike the transductive setting, the inductive setting is closer to real-world scenarios where models are required to generalize to unseen biological entities. As shown in Table 1, the performance of all methods degrades in the inductive setting, which is expected due to the increased complexity of the task. This degradation is particularly evident in the S3 (new PPI) and S4 (new modulator-new PPI) scenarios, implying the greater challenge in accurately characterizing protein interactions. Despite the difficulty, our method still achieves substantial improvements, with AUROC increases of 4.3% and 6.8% on S3 and S4 than the second-best MultiPPIMI, respectively.

Kernel-based method shows relative inadequacy in performance compared to the other four deep learning methods. This discrepancy highlights the capacity of deep learning models to address intricate computational challenges. The effectiveness of hand-selected features or kernel functions in conventional machine learning is often limited in their ability to encapsulate the complexities of data distributions, which are characteristic of complex tasks [67–70]. Among all the evaluated methods, three methods utilizing pretrained models, KFPPIMI, MultiPPIMI, and GearNet-based, generally outperform other baselines. This means that pretraining can improve model performance on corresponding downstream tasks by incorporating valuable biological insights. Among them, our KFPPIMI achieves the best performance, which can be



**Fig. 4.** Ablation studies on (A) the importance of structural knowledge, (B) the impact of pretraining on the text encoder in the modulator model, and (C) the effectiveness of the modulator and protein representation models in KFPPIMI.

attributed to its unique knowledge-fused architecture that establishes the combined properties of mappings between SMILES strings, substructures, and text descriptions to effectively capture the complex functions of modulators and proteins. Additionally, YASEEN and GearNet-based methods exhibit comparable performance even when using unverified structural data. This finding demonstrates the validity of incorporating spatial information, such as protein folding patterns and interface features, into PPIMI predictions and warrants further exploration.

#### 4.2. Ablation studies

To validate the effectiveness of the specific design of KFPPIMI, we conduct a series of ablation studies.

##### (1) Importance of structural knowledge

We explore the importance of structural knowledge by removing the fingerprint-based structural knowledge calibration in the modulator model (w/o FPs) and the GO graph-based structural knowledge injection in the protein model (w/o GO graph), respectively. The results are shown in Fig. 4A. It can be seen that the lack of explicit structural knowledge adversely affects the predictive performance of KFPPIMI. In most scenarios, this effect is not significant because textual descriptions usually contain implicit structural relationships that can partially compensate for the absence of explicit structural features. However, in the S2 scenario, removing the fingerprint information led to a 4% performance decrease. This may be attributed to the structural differences between the chemical compositions of new modulators in the test set and those present in the training set. Accordingly, the introduction of fingerprint information facilitates the comprehensive modeling of associations between molecular structures, SMILES strings, and intrinsic properties, thereby improving the model's ability to generalize the properties of new modulators.

##### (2) Impact of pretraining

To investigate the impact of pretraining on the text encoder in the modulator model, we compare KFPPIMI with two variants: without pretrained and pretrained on simple PubChem text without GPT-3.5 enrichment. The results are presented in Fig. 4B.

First, we observe that pretraining on a large corpus improves the performance of KFPPIMI. This suggests that KFPPIMI successfully learns molecular knowledge from the pretraining process and this can help it understand modulator properties. Second, pretraining with simple

texts leads to a decrease in KFPPIMI performance, especially in the S3 scenario, where AUROC is even lower than without pretraining. This is because removing GPT-enriched text exacerbates the distribution gap between pretraining and downstream data. Moreover, the enriched pretraining corpus can enhance the model's language understanding ability.

##### (3) Effectiveness of representation models

To assess the effectiveness of the modulator and protein representation models in KFPPIMI, we construct a series of variants by replacing the representation components with state-of-the-art pretrained models. These variants include KFPPIMI-MolT5 (replacing the modulator model with MolT5 [50]), KFPPIMI-GraphMVP (replacing the modulator model with GraphMVP [48]), KFPPIMI-GPT-MolBerta (replacing the modulator model with GPT-MolBerta [71]), KFPPIMI-OntoProtein (replace the protein model with OntoProtein [54]), KFPPIMI-ESM2 (replace the protein model with ESM2 [72]), and KFPPIMI-BioT5 (replace the modulator and protein model with BioT5 [73]). All replacement models use their default configurations for fair comparisons.

As shown in Fig. 4C, KFPPIMI outperforms its variants. In the S2 scenario, replacing the modulator model led to a significant decline in the AUROC metric, indicating that our modulator model is particularly well-suited for handling scenarios where the modulator is new or unseen. Similarly, in the S3 scenario, replacing the protein model within KFPPIMI resulted in a notable decline, further demonstrating the effectiveness and generalization capability of our protein model. Among all variants, KFPPIMI-GraphMVP shows the most promising results. This emphasizes the importance of explicit graph structure information in the modulator representation, consistent with the experimental results in subsection (1). When comparing KFPPIMI with the variant models that also pretrain molecular representations using external knowledge, we identify key factors contributing to KFPPIMI's superiority. Compared with KFPPIMI-MolT5, KFPPIMI benefits from a large and rich pretraining dataset, which provides a more comprehensive learning foundation. Compared with GPT-MolBerta, KFPPIMI combines real biological knowledge, including explicit structural features and PubChem descriptions, which provides an advantage in capturing molecular features. A further factor is that the word patterns in the downstream PPIMI dataset are different from those in the pretraining dataset of variant models. This difference emphasizes the need to design specialized modulator representation models, as evidenced by the comparison with the generic model BioT5, a general model for representing biological

**Table 2**

Performance of KFPPIMI-MTE on the potency prediction task.

| PPI target           | Method      | $\rho$      | $\tau$      | $r_s$       | RMSE        | MAE         |
|----------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Bcl2-Like/Bak-Bax    | pdCSM-PPI   | 0.64        | 0.43        | 0.63        | 0.79        | 0.65        |
|                      | KFPPIMI-MTE | <b>0.69</b> | <b>0.46</b> | <b>0.66</b> | <b>0.74</b> | <b>0.61</b> |
| Bromodomain/Histone  | pdCSM-PPI   | <b>0.44</b> | <b>0.34</b> | <b>0.51</b> | 0.84        | 0.61        |
|                      | KFPPIMI-MTE | 0.42        | 0.20        | 0.25        | <b>0.83</b> | <b>0.60</b> |
| Cyclophilins         | pdCSM-PPI   | 0.87        | 0.48        | 0.67        | 0.57        | 0.39        |
|                      | KFPPIMI-MTE | <b>0.87</b> | <b>0.67</b> | <b>0.81</b> | <b>0.54</b> | <b>0.39</b> |
| HIF-1 $\alpha$ /p300 | pdCSM-PPI   | 0.28        | 0.09        | 0.13        | 0.83        | 0.62        |
|                      | KFPPIMI-MTE | <b>0.33</b> | <b>0.29</b> | <b>0.39</b> | <b>0.73</b> | <b>0.58</b> |
| Integrins            | pdCSM-PPI   | 0.49        | 0.33        | 0.47        | 1.31        | <b>1.00</b> |
|                      | KFPPIMI-MTE | <b>0.53</b> | <b>0.37</b> | <b>0.52</b> | <b>1.28</b> | 1.01        |
| LEDGF/IN             | pdCSM-PPI   | -0.03       | -0.02       | -0.09       | 0.60        | 0.52        |
|                      | KFPPIMI-MTE | <b>0.28</b> | <b>0.17</b> | <b>0.30</b> | <b>0.54</b> | <b>0.45</b> |
| LFA/ICAM             | pdCSM-PPI   | 0.67        | 0.37        | 0.59        | 0.85        | 0.65        |
|                      | KFPPIMI-MTE | <b>0.84</b> | <b>0.65</b> | <b>0.85</b> | <b>0.72</b> | <b>0.57</b> |
| Mdm2-Like/P53        | pdCSM-PPI   | <b>0.63</b> | <b>0.45</b> | <b>0.61</b> | <b>0.76</b> | <b>0.61</b> |
|                      | KFPPIMI-MTE | 0.58        | 0.34        | 0.48        | 0.77        | 0.65        |
| XIAP/Smac            | pdCSM-PPI   | 0.44        | 0.37        | 0.50        | 0.89        | <b>0.57</b> |
|                      | KFPPIMI-MTE | <b>0.69</b> | <b>0.51</b> | <b>0.62</b> | <b>0.80</b> | 0.71        |

entities. Regarding the comparison with protein representation variant-based models, KFPPIMI adopts a fine-grained GO term representation and injects the GO graph structure. This approach can more accurately capture the function information of proteins, thereby improving prediction performance.

In summary, the results of ablation studies affirm the effectiveness and unique advantages of KFPPIMI, both in its overall architecture and the specific design of modulator and protein representation models, as well as its superiority over existing models in similar fields.

#### 4.3. Investigation of pretrained components

The KFPPIMI framework is supported by two independent pre-trained components: the text encoder in the modulator model (KFPPIMI-MTE) and the GO term encoder in the protein model (KFPPIMI-GOE). In addition to being integrated into the KFPPIMI architecture, we envision them having practical applications that facilitate the development of more effective PPI modulator discovery and treatment strategies. In this section, we demonstrate the generalization capabilities of these two components by applying them to relevant downstream tasks, respectively: modulator potency prediction and PPI prediction. To avoid the impact of other components and fine-tuning configurations on performance, we employ pretrained components to extract features and directly input them into Random Forest predictor implemented in the Scikit-learn library to generate prediction results, in which protein representations are concatenated by the corresponding GO term embeddings. For both experiments, the hyperparameter settings for Random Forest are the same, with “n\_estimators” (number of trees) set to 500 and other hyperparameters kept at their default values.

##### (1) KFPPIMI-MTE.

We evaluate the effectiveness of KFPPIMI-MTE on the modulator potency prediction task. The benchmark dataset is provided by pdCSM-PPI [27], which contains nine modulator datasets for nine PPI families, one for each family. The experimental pIC50 (negative logarithm of the half maximal inhibitory concentration) value of each modulator sample is regarded as a predictive label. We ensure that the preprocessing of this dataset is consistent with the pretraining dataset. We use pdCSM-PPI, the gold standard method for PPI modulator prediction, as a baseline and use Pearson correlation coefficient ( $\rho$ ), Kendall's tau coefficient ( $\tau$ ), Spearman correlation coefficient ( $r_s$ ), root mean squared error (RMSE), and mean absolute error (MAE) as evaluation metrics. The results are listed in Table 2. We can observe that KFPPIMI-MTE shows better performance than pdCSM-PPI on seven out of nine PPI families. This result validates the accuracy and generalization ability

**Table 3**

Performance of KFPPIMI-GOE on the PPI prediction task.

| Method      | Random       | DFS          | BFS          |
|-------------|--------------|--------------|--------------|
| DPPI        | 0.705        | 0.437        | 0.439        |
| DNN-PPI     | 0.752        | 0.489        | 0.516        |
| PIPR        | 0.796        | 0.522        | 0.471        |
| GNN-PPI     | 0.837        | 0.665        | 0.631        |
| SemiGNN-PPI | 0.856        | 0.693        | 0.679        |
| HIGH-PPI    | <b>0.862</b> | 0.702        | 0.684        |
| KFPPIMI-GOE | 0.849        | <b>0.703</b> | <b>0.695</b> |

of KFPPIMI-MTE and emphasizes its great potential to facilitate and accelerate the discovery of PPI modulators.

##### (2) KFPPIMI-GOE.

We evaluate the effectiveness of KFPPIMI-GOE on the PPI prediction task. The experiments are conducted on the SHS27k dataset, which is a Homo sapiens subset extracted from the widely used multi-type PPI database STRING [74] with less than 40% sequence identity. Each PPI is annotated with at least one of seven types, i.e., reaction, binding, inhibition, activation, post-translational modification, catalysis, and expression. The PPI prediction problem is formalized as a multi-label classification problem, and Micro-F1 score is used as the performance evaluation metric. Following previous studies, three partition strategies: Random, Depth-First Search (DFS), and Breath-First Search (BFS) are used to split the training and test sets. We introduce six representative PPI prediction methods as baselines: DPPI [35], DNN-PPI [36], PIPR [37], GNN-PPI [38], SemiGNN-PPI [39], and HIGH-PPI [40], and list the comparison results in Table 3.

It can be seen that KFPPIMI-GOE outperforms all baselines in DFS and BFS partitions. Although KFPPIMI-GOE performs slightly lower than SemiGNN-PPI and HIGH-PPI in Random partition, it still improves by 1.2%–14.4% over the other four baselines. Notably, DFS and BFS partitions impose stricter requirements on the generalization of models, as they involve more unseen proteins. KFPPIMI-GOE's strong performance in these challenging scenarios highlights its ability to understand protein properties by integrating biological knowledge, which is critical for robust applications in the real world.

#### 4.4. Model interpretation

KFPPIMI not only demonstrates competitive performance but also offers a significant advantage in enhanced interpretability through attention mechanisms. In this section, we present the interpretability of KFPPIMI by using the case of the Caspase-9/XIAP interaction and its modulator  $C_{17}H_{20}N_4O_2S$ . Caspases are central executors of the apoptotic program that play an indispensable role in maintaining cellular homeostasis and ensuring proper organism development. Dysregulation of Caspases activity can lead to pathological conditions, most notably various types of cancer. Among the family of inhibitors of apoptosis proteins (IAPs), XIAP is considered to be the most potent Caspase inhibitor. XIAP interacts with specific Caspases (including Caspase-3, -7, and -9) through its BIR3 and BIR2 domains, thereby inhibiting enzymatic activity and regulating the apoptotic process. Given their key role in cancer biology, the development of novel compounds that can specifically modulate Caspase/XIAP interactions has emerged as a promising therapeutic strategy.

We initiate our investigation by observing whether and how KFPPIMI focuses on crucial interaction features during the prediction process. We visualize the attention weights of different GO terms within the Caspase-9 and XIAP proteins. As shown in Fig. 5A, certain GO terms received higher attention scores, presenting in a vertical line pattern. For the Caspase-9 protein, the two most important terms are GO:0008047 (enzyme activator activity) and GO:0097153 (cysteine-type endopeptidase activity involved in apoptotic process). For the XIAP protein, the two top-ranked terms are GO:0004869 (cysteine-type endopeptidase inhibitor activity) and GO:0043027 (cysteine-type

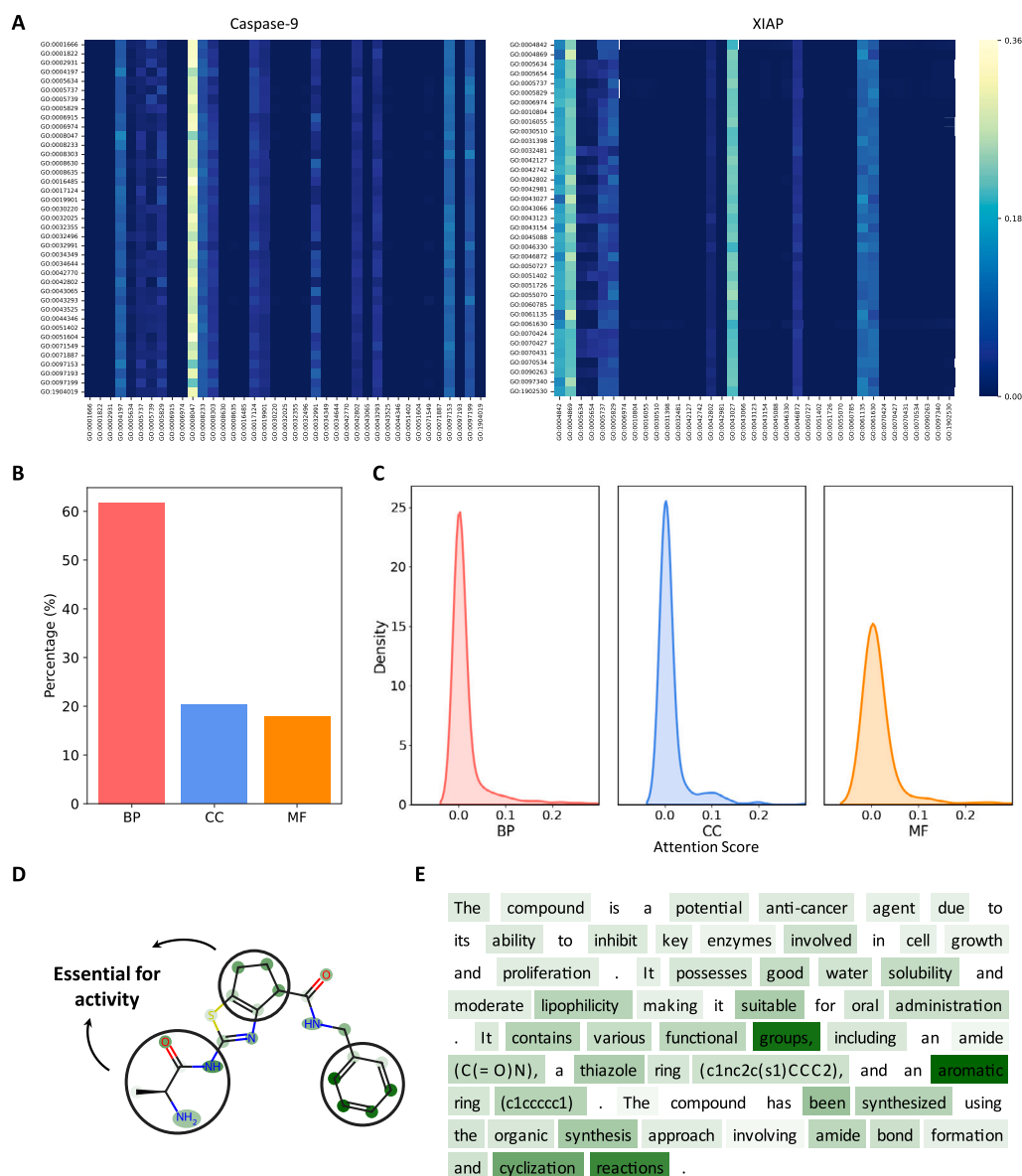


Fig. 5. Model interpretability studies on the Caspase-9/XIAP target and its modulator  $C_{17}H_{20}N_4O_2S$ .

endopeptidase inhibitor activity involved in apoptotic process). It is evident that the key terms identified by KFPPIMI are closely related to the functional characteristics of the Caspase-9/XIAP interaction. Further, we analyze the distribution and significance of different types of GO terms across all 117 PPI targets in the S1 scenario. As shown in Fig. 5B and C, BP and CC terms contribute more to the interaction prediction overall than MF terms, especially CC terms, which have higher attention scores despite their lower background frequency. This finding aligns with our expectations. Given that PPI is inherently highly dependent on physical proximity, the cellular component of the protein is more important when modeling the interaction between two proteins. If two proteins are located in different cellular regions, spatial separation poses a significant barrier to their interaction.

$C_{17}H_{20}N_4O_2S$  is a novel Caspase-9/XIAP inhibitor designed and synthesized based on the AVPI tetrapeptide [75]. We compute the attention scores for  $C_{17}H_{20}N_4O_2S$  in the KFPPIMI model and color different parts of the input text based on these scores. For clarity, the SMILES sequence and textual description of  $C_{17}H_{20}N_4O_2S$  are visualized separately. Darker colors indicate higher attention scores and lighter colors indicate lower attention scores. As presented in Fig. 5D, most atoms

in the alanine, proline, and aromatic groups are assigned high attention values. These groups are closely associated with the actual binding coordination. Specifically, the alanine and proline have been determined as essential amino acid residues for preserving the activity of the AVPI tetrapeptide [76], while the aromatic group connected by an amide bond distinguishes  $C_{17}H_{20}N_4O_2S$  from other analogs. Additionally, KFPPIMI emphasizes the terms such as “lipophilicity”, “functional groups”, “aromatic”, and “cyclization” in the textual description, which are consistent with key elements of  $C_{17}H_{20}N_4O_2S$  (Fig. 5E). This result indicates that KFPPIMI can effectively understand the chemical properties and structure of molecules and selectively focus on functional groups. Collectively, the case study validates the effectiveness of KFPPIMI in capturing implicit relationships in PPI-modulator interactions and providing interpretable results.

## 5. Conclusion

In this study, we introduce a knowledge-fused PPIMI prediction approach that incorporates independent modulator and protein representation models. With language modeling, KFPPIMI can effectively fuse



factual knowledge derived from textual and graphical data to enhance a comprehensive understanding of biomolecules. Compared with several baseline methods, KFPPIMI achieves improvements in both transductive and inductive settings. The modulator and protein representation components also yield comparable results in their corresponding modulator potency prediction and PPI prediction tasks, respectively, showing outstanding practical application capabilities to various downstream scenarios. Another key advantage of KFPPIMI is its ability to robustly address the limitations of previous approaches in elucidating potential interaction mechanisms. As demonstrated by our interpretability experiments on the interaction between the Caspase-9/XIAP target and its inhibitor, KFPPIMI provides more detailed insights into the key factors influencing interaction relationships leveraging the attention mechanism. We anticipate that our work will inspire further innovations in language modeling for PPI modulator screening, ultimately leading to advances in practical applications for drug discovery.

Looking ahead, we aim to integrate more sources and types of biomedical knowledge. For the modulator model, KFPPIMI focuses primarily on the acquisition and fusion of molecular sequence and functional information, but 3D formats are also of significance. One promising direction is to construct larger multimodal datasets that explicitly include 3D structural information in textual descriptions. For the protein model, we plan to integrate a broader range of biological data, covering essential components such as PPI networks and disease associations. To this end, the knowledge injection module of KFPPIMI will be extended and enhanced, such as introducing more efficient tokenization schemes and loss-aware injection approaches.

Furthermore, due to the limitation of computing resources, KFPPIMI employs simple graph embedding models to encode structured features. In the future, we will explore advanced GNNs [77–79] and pretraining techniques to enhance the modeling of biomolecules and their interactions. GNNs provide a powerful framework for efficiently capturing topological and spatial relationships within molecular structures. Moreover, they can facilitate the analysis of complex biological networks and structural changes in subgraphs. An intriguing direction could be to model the evolutionary mechanisms underlying the dynamic transformations of PPI networks in response to biological stimuli, including modulator interventions.

### CRedit authorship contribution statement

**Zitong Zhang:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation, Conceptualization. **Quan Zou:** Writing – review & editing, Validation, Supervision, Project administration, Investigation, Data curation. **Chunyu Wang:** Writing – review & editing, Validation, Supervision, Investigation, Data curation. **Junjie Wang:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Conceptualization. **Lingling Zhao:** Writing – review & editing, Validation, Supervision, Project administration, Investigation.

### Code availability

The source code is available at <https://github.com/1zzt/KFPPIMI>.

### Declaration of competing interest

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

### Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC, Grant no. 62272136, 62231013, 62171164).

### Data availability

Data will be made available on request.

### References

- [1] F. Cheng, J. Zhao, Y. Wang, W. Lu, Z. Liu, Y. Zhou, W.R. Martin, R. Wang, J. Huang, T. Hao, et al., Comprehensive characterization of protein-protein interactions perturbed by disease mutations, *Nature Genet.* 53 (3) (2021) 342–353.
- [2] J. Zhang, T. Huang, Q. Sun, J. Zhang, Identifying pathological myopia associated genes with a random walk-based method in protein-protein interaction network, *Curr. Bioinform.* 19 (4) (2024) 375–384.
- [3] H. Ruffner, A. Bauer, T. Bouwmeester, Human protein-protein interaction networks and the value for drug discovery, *Drug Discov. Today* 12 (17–18) (2007) 709–716.
- [4] H. Nada, Y. Choi, S. Kim, K.S. Jeong, N.A. Meanwell, K. Lee, New insights into protein-protein interaction modulators in drug discovery and therapeutic advance, *Signal Transduct. Target. Ther.* 9 (1) (2024) 1–32.
- [5] H. Zhang, Z. Feng, C. Wu, Refining protein interaction network for identifying essential proteins, *Curr. Bioinform.* 18 (3) (2023) 255–265.
- [6] H.-L. Li, Y.-H. Pang, B. Liu, BioSeq-BLM: A platform for analyzing DNA, RNA and protein sequences based on biological language models, *Nucleic Acids Res.* 49 (22) (2021) e129–e129.
- [7] X. Liu, C. Ai, H. Yang, R. Dong, J. Tang, S. Zheng, F. Guo, RetroCaptioner: Beyond attention in end-to-end retrosynthesis transformer via contrastively captioned learnable graph representation, *Bioinformatics* 40 (9) (2024) btac561.
- [8] S. Kang, T. Tanaka, T. Kishimoto, Therapeutic uses of anti-interleukin-6 receptor antibody, *Int. Immunol.* 27 (1) (2015) 21–29.
- [9] S. Dhillon, Adagrasib: First approval, *Drugs* 83 (3) (2023) 275–285.
- [10] K. Yan, H. Lv, Y. Guo, W. Peng, B. Liu, Samppred-GAT: Prediction of antimicrobial peptide by graph attention network and predicted peptide structure, *Bioinformatics* 39 (1) (2023) btac715.
- [11] H. Zhu, H. Hao, L. Yu, Identification of microbe–disease signed associations via multi-scale variational graph autoencoder based on signed message propagation, *BMC Biol.* 22 (1) (2024) 172.
- [12] Z. Huang, Z. Xiao, C. Ao, L. Guan, L. Yu, Computational approaches for predicting drug-disease associations: A comprehensive review, *Front. Comput. Sci.* 19 (5) (2025) 1–15.
- [13] M. De Vivo, M. Masetti, G. Bottegoni, A. Cavalli, Role of molecular dynamics and related methods in drug discovery, *J. Med. Chem.* 59 (9) (2016) 4035–4061.
- [14] R. Sable, S. Jois, Surfing the protein-protein interaction surface using docking methods: application to the design of PPI inhibitors, *Molecules* 20 (6) (2015) 11569–11603.
- [15] T. Wang, J. Yang, Y. Xiao, J. Wang, Y. Wang, X. Zeng, Y. Wang, J. Peng, DFinder: A novel end-to-end graph embedding-based method to identify drug–food interactions, *Bioinformatics* 39 (1) (2023) btac837.
- [16] Z. Yang, J. Liu, X. Zhu, F. Yang, Q. Zhang, H.A. Shah, FragDPI: A novel drug-protein interaction prediction model based on fragment understanding and unified coding, *Front. Comput. Sci.* 17 (5) (2023) 175903.
- [17] H. Cheng, B. Rao, L. Liu, L. Cui, G. Xiao, R. Su, L. Wei, PepFormer: End-to-end transformer-based siamese network to predict and enhance peptide detectability based on sequence only, *Anal. Chem.* 93 (16) (2021) 6481–6490.
- [18] S. Ren, L. Chen, H. Hao, L. Yu, Prediction of cancer drug combinations based on multidrug learning and cancer expression information injection, *Future Gener. Comput. Syst.* 160 (2024) 798–807.
- [19] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, The rise of deep learning in drug discovery, *Drug Discov. Today* 23 (6) (2018) 1241–1250.
- [20] A. Neugebauer, R.W. Hartmann, C.D. Klein, Prediction of protein–protein interaction inhibitors by chemoinformatics and machine learning methods, *J. Med. Chem.* 50 (19) (2007) 4665–4668.
- [21] C. Reynès, H. Host, A.-C. Camproux, G. Laconde, F. Leroux, A. Mazars, B. Deprez, R. Fahraeus, B.O. Villoutreix, O. Sperandio, Designing focused chemical libraries enriched in protein-protein interaction inhibitors using machine-learning methods, *PLoS Comput. Biol.* 6 (3) (2010) e1000695.
- [22] T. Jana, A. Ghosh, S. Das Mandal, R. Banerjee, S. Saha, PPIMPred: A web server for high-throughput screening of small molecules targeting protein-protein interaction, *R. Soc. Open Sci.* 4 (4) (2017) 160501.
- [23] B.I. Díaz-Eufracio, J.L. Medina-Franco, Machine learning models to predict protein-protein interaction inhibitors, *Molecules* 27 (22) (2022) 7986.
- [24] Z. Zhang, L. Zhao, M. Gao, Y. Chen, J. Wang, C. Wang, PPII-AEAT: Prediction of protein-protein interaction inhibitors based on autoencoders with adversarial training, *Comput. Biol. Med.* 172 (2024) 108287.
- [25] Z. Zhang, Z. Wang, L. Zhao, J. Wang, C. Wang, Multimodal contrastive learning for protein-protein interaction inhibitor prediction, in: 2024 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2024, pp. 1327–1333.

- [26] P. Gupta, D. Mohanty, SMMPPi: A machine learning-based approach for prediction of modulators of protein-protein interactions and its application for identification of novel inhibitors for RBD: hACE2 interactions in SARS-CoV-2, *Brief. Bioinform.* 22 (5) (2021) bbab111.
- [27] C.H. Rodrigues, D.E. Pires, D.B. Ascher, PdCSM-PPI: Using graph-based signatures to identify protein-protein interaction inhibitors, *J. Chem. Inf. Model.* 61 (11) (2021) 5438–5445.
- [28] Z. Zhang, L. Zhao, J. Wang, C. Wang, A hierarchical graph neural network framework for predicting protein-protein interaction modulators with functional group information and hypergraph structure, *IEEE J. Biomed. Heal. Inform.* (2024).
- [29] H. Sun, J. Wang, H. Wu, S. Lin, J. Chen, J. Wei, S. Lv, Y. Xiong, D.-Q. Wei, A multimodal deep learning framework for predicting PPI-modulator interactions, *J. Chem. Inf. Model.* 63 (23) (2023) 7363–7372.
- [30] A. Yaseen, S. Roy, N. Akhter, A. Ben-Hur, F. Minhas, Predicting small-molecule inhibition of protein complexes, 2024, *BioRxiv*, 2024-2008.
- [31] L. Hu, X. Wang, Y.-A. Huang, P. Hu, Z.-H. You, A survey on computational models for predicting protein-protein interactions, *Brief. Bioinform.* 22 (5) (2021) bbab036.
- [32] J. Durham, J. Zhang, I.R. Humphreys, J. Pei, Q. Cong, Recent advances in predicting and modeling protein-protein interactions, *Trends Biochem. Sci.* 48 (6) (2023) 527–538.
- [33] X. Luo, L. Wang, P. Hu, L. Hu, Predicting protein-protein interactions using sequence and network information via variational graph autoencoder, *IEEE/ACM Trans. Comput. Biol. Bioinform.* 20 (5) (2023) 3182–3194.
- [34] L. Hu, S. Yang, X. Luo, H. Yuan, K. Sedraoui, M. Zhou, A distributed framework for large-scale protein-protein interaction data analysis and prediction using mapreduce, *IEEE/CAA J. Autom. Sin.* 9 (1) (2021) 160–172.
- [35] S. Hashemifar, B. Neyshabur, A.A. Khan, J. Xu, Predicting protein-protein interactions through sequence-based deep learning, *Bioinformatics* 34 (17) (2018) i802–i810.
- [36] H. Li, X.-J. Gong, H. Yu, C. Zhou, Deep neural network based predictions of protein interactions using primary sequences, *Molecules* 23 (8) (2018) 1923.
- [37] M. Chen, C.J.-T. Ju, G. Zhou, X. Chen, T. Zhang, K.-W. Chang, C. Zaniolo, W. Wang, Multifaceted protein-protein interaction prediction based on siamese residual RCNN, *Bioinformatics* 35 (14) (2019) i305–i314.
- [38] G. Lv, Z. Hu, Y. Bi, S. Zhang, Learning unknown from correlations: Graph neural network for inter-novel-protein interaction prediction, 2021, *arXiv preprint arXiv:2105.06709*.
- [39] Z. Zhao, P. Qian, X. Yang, Z. Zeng, C. Guan, W.L. Tam, X. Li, Semgnpp: Self-ensembling multi-graph neural network for efficient and generalizable protein-protein interaction prediction, 2023, *arXiv preprint arXiv:2305.08316*.
- [40] Z. Gao, C. Jiang, J. Zhang, X. Jiang, L. Li, P. Zhao, H. Yang, Y. Huang, J. Li, Hierarchical graph learning for protein-protein interaction, *Nat. Commun.* 14 (1) (2023) 1093.
- [41] C. Ai, H. Yang, X. Liu, R. Dong, Y. Ding, F. Guo, MTMol-GPT: De novo multi-target molecular generation with transformer-based generative adversarial imitation learning, *PLoS Comput. Biol.* 20 (6) (2024) e1012229.
- [42] C. Pang, J. Qiao, X. Zeng, Q. Zou, L. Wei, Deep generative models in de novo drug molecule generation, *J. Chem. Inf. Model.* 64 (7) (2023) 2174–2194.
- [43] J. Wang, J. Mao, C. Li, H. Xiang, X. Wang, S. Wang, Z. Wang, Y. Chen, Y. Li, K.T. No, et al., Interface-aware molecular generative framework for protein-protein interaction modulators, *J. Cheminform.* 16 (1) (2024) 142.
- [44] L. Lai, Y. Liu, B. Song, K. Li, X. Zeng, Deep generative models for therapeutic peptide discovery: A comprehensive review, *ACM Comput. Surv.* (2025).
- [45] M. Liu, C. Li, R. Chen, D. Cao, X. Zeng, Geometric deep learning for drug discovery, *Expert Syst. Appl.* 240 (2024) 122498.
- [46] S. Chithrananda, G. Grand, B. Ramsundar, ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction, 2020, *arXiv preprint arXiv:2010.09885*.
- [47] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C.L. Zitnick, J. Ma, et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, *Proc. Natl. Acad. Sci.* 118 (15) (2021) e2016239118.
- [48] S. Liu, H. Wang, W. Liu, J. Lasenby, H. Guo, J. Tang, Pre-training molecular graph representation with 3d geometry, 2021, *arXiv preprint arXiv:2110.07728*.
- [49] Z. Zhang, M. Xu, A. Jamash, V. Chenthamarakshan, A. Lozano, P. Das, J. Tang, Protein representation learning by geometric structure pretraining, 2022, *arXiv preprint arXiv:2203.06125*.
- [50] C. Edwards, T. Lai, K. Ros, G. Honke, K. Cho, H. Ji, Translation between molecules and natural language, 2022, *arXiv preprint arXiv:2204.11817*.
- [51] M. Xu, X. Yuan, S. Miret, J. Tang, Protst: Multi-modality learning of protein sequences and biomedical texts, in: *International Conference on Machine Learning*, PMLR, 2023, pp. 38749–38767.
- [52] E. Krissinel, On the relationship between sequence and structure similarities in proteomics, *Bioinformatics* 23 (6) (2007) 717–723.
- [53] P.A. Alexander, Y. He, Y. Chen, J. Orban, P.N. Bryan, The design and characterization of two proteins with 88% sequence identity but different structure and function, *Proc. Natl. Acad. Sci.* 104 (29) (2007) 11963–11968.
- [54] N. Zhang, Z. Bi, X. Liang, S. Cheng, H. Hong, S. Deng, J. Lian, Q. Zhang, H. Chen, Ontoprotein: Protein pretraining with gene ontology embedding, 2022, *arXiv preprint arXiv:2201.11147*.
- [55] H.-Y. Zhou, Y. Fu, Z. Zhang, B. Cheng, Y. Yu, Protein representation learning via knowledge enhanced primary structure reasoning, in: *The Eleventh International Conference on Learning Representations*, 2023.
- [56] A.A. Edera, D.H. Milone, G. Stegmayer, Anc2vec: Embedding gene ontology terms by preserving ancestors relationships, *Brief. Bioinform.* 23 (2) (2022) bbac003.
- [57] W. Li, B. Wang, J. Dai, Y. Kou, X. Chen, Y. Pan, S. Hu, Z.Z. Xu, Partial order relation-based gene ontology embedding improves protein function prediction, *Brief. Bioinform.* 25 (2) (2024) bbae077.
- [58] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, *arXiv preprint arXiv:1907.11692*.
- [59] J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, *arXiv preprint arXiv:1810.04805*.
- [60] S. Wada, T. Takeda, S. Manabe, S. Konishi, J. Kamohara, Y. Matsumura, Pre-training technique to localize medical bert and enhance biomedical bert, 2020, *arXiv preprint arXiv:2005.07202*.
- [61] L.F. Ribeiro, P.H. Saverese, D.R. Figueiredo, Struc2vec: Learning node representations from structural identity, in: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 385–394.
- [62] B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, in: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 701–710.
- [63] A. Grover, J. Leskovec, Node2vec: Scalable feature learning for networks, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 855–864.
- [64] N. Shazeer, Glu variants improve transformer, 2020, *arXiv preprint arXiv:2002.05202*.
- [65] K. Ikeda, Y. Maezawa, T. Yonezawa, Y. Shimizu, T. Tashiro, S. Kanai, N. Sugaya, Y. Masuda, N. Inoue, T. Niimi, et al., Dlip-PPI library: An integrated chemical database of small-to-medium-sized molecules targeting protein-protein interactions, *Front. Chem.* 10 (2023) 1090643.
- [66] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A.J. Ballard, J. Bambrick, et al., Accurate structure prediction of biomolecular interactions with AlphaFold 3, *Nature* 630 (8016) (2024) 493–500.
- [67] Y. Wang, Y. Zhai, Y. Ding, Q. Zou, SBSM-pro: Support bio-sequence machine for proteins, *Sci. China Inf. Sci.* 67 (11) (2024) 212106.
- [68] P. Kumar Meher, S. Hati, T.K. Sahu, U. Pradhan, A. Gupta, S.N. Rath, SVM-root: Identification of root-associated proteins in plants by employing the support vector machine with sequence-derived features, *Curr. Bioinform.* 19 (1) (2024) 91–102.
- [69] H. Feng, W. Yang, J. Chen, On optimal streaming kernelization algorithms, *Sci. China Inf. Sci.* 67 (8) (2024) 189101.
- [70] Y. Wang, X. Zhang, Y. Ju, Q. Liu, Q. Zou, Y. Zhang, Y. Ding, Y. Zhang, Identification of human microRNA-disease association via low-rank approximation-based link propagation and multiple kernel learning, *Front. Comput. Sci.* 18 (2) (2024) 182903.
- [71] S. Balaji, R. Magar, Y. Jadhav, A.B. Farimani, Gpt-molberta: Gpt molecular features language model for molecular property prediction, 2023, *arXiv preprint arXiv:2310.03030*.
- [72] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al., Evolutionary-scale prediction of atomic-level protein structure with a language model, *Science* 379 (6637) (2023) 1123–1130.
- [73] Q. Pei, W. Zhang, J. Zhu, K. Wu, K. Gao, L. Wu, Y. Xia, R. Yan, Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations, 2023, *arXiv preprint arXiv:2310.07276*.
- [74] D. Szklarczyk, A.L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N.T. Doncheva, J.H. Morris, P. Bork, et al., STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, *Nucleic Acids Res.* 47 (D1) (2019) D607–D613.
- [75] F.R.P. Crisóstomo, Y. Feng, X. Zhu, K. Welsh, J. An, J.C. Reed, Z. Huang, Design and synthesis of a simplified inhibitor for XIAP-BIR3 domain, *Bioorganic & Med. Chem. Lett.* 19 (22) (2009) 6413–6418.
- [76] T.K. Oost, C. Sun, R.C. Armstrong, A.-S. Al-Asaad, S.F. Betz, T.L. Deckwerth, H. Ding, S.W. Elmore, R.P. Meadows, E.T. Olejniczak, et al., Discovery of potent antagonists of the antiapoptotic protein XIAP for the treatment of cancer, *J. Med. Chem.* 47 (18) (2004) 4417–4426.
- [77] Y. Yang, G. Li, D. Li, J. Zhang, P. Hu, L. Hu, Integrating fuzzy clustering and graph convolution network to accurately identify clusters from attributed graph, *IEEE Trans. Netw. Sci. Eng.* (2024).
- [78] C. Song, Y. Lin, S. Guo, H. Wan, Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 914–921, 01.
- [79] G. Bouritsas, F. Frasca, S. Zafeiriou, M.M. Bronstein, Improving graph neural network expressivity via subgraph isomorphism counting, *IEEE Trans. Pattern Anal. Mach. Intell.* 45 (1) (2022) 657–668.