frontiers | Frontiers in Pharmacology

TYPE Original Research
PUBLISHED 02 June 2025
DOI 10.3389/fphar.2025.1589788

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology, China

REVIEWED BY
Bo-Wei Zhao,
Chinese Academy of Sciences (CAS), China
Zhiyuan Chen,
University of Nottingham Malaysia Campus, Malaysia

*CORRESPONDENCE
Chengcheng Zhang,
zcc1203_hit@163.com
Tianyi Zhao,
zty2009@hit.edu.cn

RECEIVED 08 March 2025
ACCEPTED 20 May 2025
PUBLISHED 02 June 2025

CITATION
Qi H, Li X, Zhang C and Zhao T (2025) Improving drug-drug interaction prediction via in-context learning and judging with large language models.
Front. Pharmacol. 16:1589788.
doi: 10.3389/fphar.2025.1589788

Improving drug-drug interaction prediction via in-context learning and judging with large language models

He Qi[1,2], Xiaoqiang Li[3], Chengcheng Zhang[4]* and Tianyi Zhao[1,5]*

[1]School of Medicine and Health, Harbin Institute of Technology, Harbin, China, [2]Center for Drug Evaluation and Inspection for Heilongjiang Province, Harbin, China, [3]Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, China, [4]Faculty of Computing, Harbin Institute of Technology, Harbin, China, [5]Harbin Institute of Technology Zhengzhou Research Institute, Zhengzhou, China

Introduction: Large Language Models (LLMs), recognized for their advanced capabilities in natural language processing, have been successfully employed across various domains. However, their effectiveness in addressing challenges related to drug discovery has yet to be fully elucidated.

Methods: In this paper, we propose a novel LLM based method for drug-drug interaction (DDI) prediction, named DDI-JUDGE, achieved through the integration of judging and ICL prompts. The proposed method outperforms existing LLM approaches, demonstrating the potential of LLMs for predicting DDIs. We introduce a novel in-context learning (ICL) prompt paradigm that selects high-similarity samples as positive and negative prompts, enabling the model to effectively learn and generalize knowledge. Additionally, we present an ICL-based prompt template that structures inputs, prediction tasks, relevant factors, and examples, leveraging the pre-trained knowledge and contextual understanding of LLMs to enhance DDI prediction capabilities. To further refine predictions, we employ GPT-4 as a discriminator to assess the relevance of predictions generated by multiple LLMs.

Results: DDI-JUDGE achieves the best performance among all models in both zero-shot and few-shot settings, with an AUC of 0.642/0.788 and AUPR of 0.629/0.801, respectively. These results demonstrate its superior predictive capability and robustness across different learning scenarios.

Development: These findings highlight the potential of LLMs in advancing drug discovery through more effective DDI prediction. The modular prompt structure, combined with ensemble reasoning, offers a scalable framework for knowledge-intensive biomedical applications. The code for DDI-JUDGE is available at https://github.com/zcc1203/ddi-judge.

KEYWORDS

large language models, drug-drug interactions, in-context learning, zero-shot, few-shot

1 Introduction

Polypharmacy, or the simultaneous use of multiple drugs, is common in the treatment of patients with various diseases (van Roon et al., 2005). However, it can lead to adverse drug reactions (DDIs) due to drug–drug interactions. DDIs are responsible for 30% of all reported adverse drug reactions, significantly impacting patient safety, morbidity, mortality,

and healthcare costs (Ryu et al., 2018). Given the complexity of diseases and the limitations of single-drug therapies, combination therapies have the potential to improve efficacy, but they also increase the risk of unintended interactions (Deng et al., 2020). Therefore, accurate DDI prediction is crucial for improving treatment outcomes and minimizing adverse effects. Although DDI research has become a major focus, the identification of DDIs remains challenging due to limited clinical trial resources and the rapid growth of biomedical data.

Current state-of-the-art DDI prediction methods include traditional machine learning and deep learning approaches. Among these, deep learning methods leverage technologies such as deep neural networks (DNNs) (Sze et al., 2017), convolutional neural networks (CNNs) (Alzubaidi et al., 2021), graph neural networks (GNNs) (Wu et al., 2020) and transformer (Vaswani, 2017), achieving remarkable performance. However, these methods often perform poorly in zero-shot scenarios and exhibit limited capability in learning from large-scale, multi-source data integration.

Large language models (LLMs), exemplified by architectures such as GPT-4 (Achiam et al., 2023), Claude (Ryu et al., 2018; Bai et al., 2022), llama (Touvron et al., 2023a), and Mistral (Jiang et al., 2023), have demonstrated transformative capabilities in general-domain tasks through their massive parameter spaces, self-supervised pretraining frameworks, and attention-based neural architectures. While LLMs demonstrate exceptional performance in general tasks, their capabilities in specialized application domains remain significantly constrained.

In the field of drug discovery, LLM have demonstrated significant potential in several directions, including the integration of multi-source data (Wan et al., 2024), the design of downstream tasks (Guo et al., 2023), and the optimization of prompting strategies for specific applications (Guo et al., 2024). These advancements have enabled LLMs to perform tasks such as molecular property prediction and molecular translation. However, critical challenges persist in applying LLMs to DDI prediction: 1) the scarcity of high-quality, annotated DDI datasets due to expensive experimental validation; 2) poor generalizability under zero-shot learning conditions, particularly for rare interaction types; 3) ineffective fusion of heterogeneous data modalities spanning molecular structures, pharmacological pathways, and clinical context.

To overcome these limitations, we introduced the DDI-JUDGE model, which employs in-Context Learning (ICL) to propose a prompt paradigm tailored for DDI tasks and leverages a judge to integrate the predictive capabilities of multiple LLMs.

The main contributions of this paper are as follows:

1) We propose a DDI prediction method based on large language models enhanced by judging and in-context learning, named DDI-JUDGE.
2) We propose a novel ICL prompt paradigm for DDI prediction, employing cosine similarity-based exemplar retrieval for in-context learning and coupling it with an ensemble discriminator module, such as GPT-4, that strategically aggregates predictions from heterogeneous LLMs through confidence-weighted voting, thereby improving robustness against model bias.
3) The effectiveness of our method has been demonstrated through comprehensive experiments in both zero-shot and few-shot scenarios, outperforming other LLM methods.

The structure of this paper is as follows: The Related Work section provides a brief review focusing on methods of DDI prediction. The Methods section offers a detailed description of the proposed DDI-JUDGE method. The Experiments and Results section presents the experimental setup and analyzes the results. Finally, the Conclusion section summarizes the key points and discusses potential directions for future research.

2 Related work

2.1 Methods of drug-drug interactions prediction

DDI prediction has been an essential area of research due to its critical implications for rational drug use, enhancing therapeutic efficacy, and minimizing adverse drug reactions. Numerous computational models, including traditional machine learning and deep learning approaches, have been developed for DDI prediction.

Traditional machine learning models predict DDIs by leveraging features such as drug similarity, protein-protein interaction networks, and drug phenotypic profiles. For instance, Bayesian models calculate interaction scores based on protein networks and drug phenotype similarity (Huang et al., 2013). Label propagation-based models (Zhang et al., 2015) integrate drug side effects and chemical structure data, while probabilistic frameworks, such as the collective soft logic model (Sridhar et al., 2016) rely on multi-source similarity features. Additionally, manifold regularization and matrix factorization approaches, like DDINMF (Yu et al., 2018) and TMFUF (Shi et al., 2018), enhance predictions by incorporating semi-nonnegative matrix decomposition and manifold structures.

Deep learning methods have significantly enhanced DDI prediction by enabling complex feature extraction and multi-source data integration. Models like DDIMDL (Deng et al., 2020) and CNN-DDI (Zhang et al., 2022) employ deep neural networks (DNNs) and CNNs, respectively, to calculate interaction probabilities using drug similarity matrices. Graph-based methods, such as SSI-DDI (Nyamabo et al., 2021), convert SMILES strings into molecular graphs and utilize graph attention networks (GATs) to extract substructure representations. Tensor-based approaches like STNN-DDI (Yu et al., 2022) employ tensor factorization to predict interaction types. Network-based methods have further refined DDI prediction by incorporating multi-relation and heterogeneous data. For instance, META-DDIE (Deng et al., 2022) combines frequent substructure mining and neural encoding for DDI type prediction, while DANN-DDI (Liu et al., 2022) employs attention mechanisms to generate comprehensive drug embeddings from heterogeneous networks. MRCGNN (Xiong et al., 2023) utilizes multi-relation DDI event graphs with relational graph convolutional networks for feature extraction. Similarly, SubGE-DDI (Shi et al., 2024) integrates substructure representations from molecular graphs with attention-based mechanisms to improve prediction accuracy. Furthermore, KGE-UNIT

(Zhang et al., 2024) enhances DDI prediction performance by multi-task learning. These network-driven and hybrid approaches offer significant improvements by combining molecular, structural, and contextual data in highly integrated frameworks. However, the current method has insufficient learning ability for massive multi-source data and cannot adapt well to the zero-shot scenario.

2.2 Large language models for drug discovery

LLM have shown significant potential in advancing molecular science by bridging textual information and molecular data, which has facilitated applications such as molecule retrieval, reaction prediction, and drug discovery. Recent studies, such as Text2Mol (Edwards et al., 2021), Molxpt (Liu Z. et al., 2023), and Mol-Instructions (Fang et al., 2023), have established connections between molecular structures and textual descriptions, enhancing tasks such as molecule editing, annotation, and retrosynthesis. In drug development, Y-Mol (Ma et al., 2024) and DrugReAlign (Wei et al., 2024) demonstrate the versatility of LLMs in tackling complex tasks. Y-Mol offers a biomedical knowledge-guided approach for virtual screening, property prediction, and drug interaction prediction, enhancing domain-specific reasoning. Meanwhile, DrugReAlign focuses on improving drug repurposing through a multisource prompt framework that integrates spatial interaction data and leverages LLMs for reliable drug-target analysis. In domains such as protein analysis and drug design, Protst (Xu et al., 2023) and Drugchat (Liang et al., 2023) employ in-context learning and interactive design to align with user-specific needs.

However, molecular interactions prediction tasks, such as DDIs, still face many challenges. Existing methods typically rely on high-quality fine-tuning data and computationally intensive fine-tuning algorithms (Hu et al., 2021), but the high cost of data acquisition and model training presents significant obstacles. Enhancing the ability of LLMs to predict DDIs under the constraints of limited data and training resources remains a key issue to be addressed.

2.3 Prompt engineering for LLM

The framework that combines pre-training and prompts has become a widely recognized best practice in natural language processing, particularly for addressing few-shot and zero-shot tasks (Liu P. et al., 2023). This approach is founded on the principle that LLM possess the capability for in-context learning by leveraging input contexts and instructions (Brown, 2020). Several studies have explored the use of LLM-based approaches for drug design by incorporating various prompting strategies. Li et al. (2024) propose a retrieval-based prompting approach for molecule-caption translation. Liu Y. et al. (2024) introduce MolecularGPT, which provides curated molecular instructions for over 1000 property prediction tasks. Chaves et al. (2024) present TxT-LLM, a method that combines free-text instructions with string representations of molecules throughout different stages of the drug discovery process. However, these methods may not fully capture the complexity of DDIs, which involve various factors including molecular, pharmacological, and clinical considerations.

ICL (Dong et al., 2022) can improve the model's ability to understand and adapt to different drug combinations by leveraging contextual information from multiple drug-related tasks. Additionally, it allows the model to flexibly adjust its responses based on new data or conditions, such as changes in drug formulations or patient-specific factors, thereby enhancing its predictive capability for DDIs. However, designing more effective ICL prompting paradigms for DDI prediction is an area that requires further study.

3 Methods

In this section, we will provide a detailed explanation of the DDI-JUDGE method. This method aims to explore how existing LLMs can be used for DDI prediction, with the overall framework illustrated in Figure 1. The method is primarily divided into three parts: 1) Selecting ICL samples, 2) Building prompts based on ICL, and 3) Generating an LLM-based discriminator to integrate multi-model results. First, DDI-JUDGE leverages drug similarity to select optimal prompt samples, performing positive sample selection and hard negative sample mining. Next, based on the selected prompt samples, we construct prompt templates specifically designed for DDI prediction. Finally, we use GPT to generate an LLM-based discriminator, which scores the predictions of multiple LLMs and integrates the results based on the scores.

3.1 Selecting better ICL samples of DDIs

ICL is a prompting paradigm applied to LLMs, which enhances the capabilities of LLM by using a small number of demonstration prompts. In DDI prediction, we need to study how to find more suitable prompt examples. In order to better select prompt examples, we propose an ICL positive and negative samples selection method for DDI based on drug similarity calculation. Three widely used similarity measures include Tanimoto similarity (Tanimoto, 1958), Cosine similarity, and Dice similarity. Tanimoto To more effectively assess the similarity of drug feature vectors, we examine the variations in the outcomes produced by these methods. In drug similarity calculations, Dice similarity emphasizes shared structural features, making it suitable for identifying common substructures. Cosine similarity focuses on the angular relationship of feature vectors, ideal for analyzing high-dimensional molecular data. Tanimoto similarity balances shared and unique molecular features, making it particularly effective for comparing molecular fingerprints in chem-informatics. Let x = $(x_1, x_2, \ldots, x_n)$ and y = $(y_1, y_2, \ldots, y_n)$ represent the binary molecular fingerprints of two drugs got from Rdkit (RDKit, 2013), where each element indicates the presence or absence of a specific substructure. The Tanimoto similarity is defined as shown in Equations 1:

$$Sim_T(x, y) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\|^2 \times \|\boldsymbol{y}\|^2} = \frac{\sum \boldsymbol{x_i y_i}}{\sqrt{\sum \boldsymbol{x_i^2}} + \sqrt{\sum \boldsymbol{y_i^2}} - \sum \boldsymbol{x_i y_i}} \quad (1)$$

where $x \cdot y$ denotes the dot product (inner product) of x and y, calculated as $\sum x_i y_i$, i.e., the summation of element-wise products.

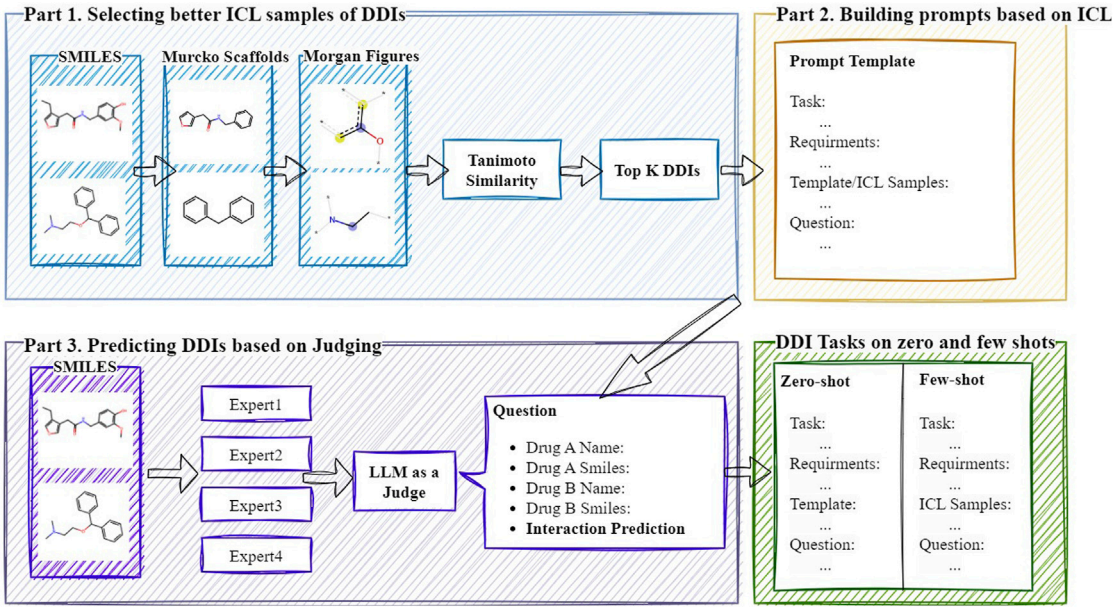

FIGURE 1
The workflow of DDI-JUDGE.

$\|x\|^2$ and $\|y\|^2$ represent the squared norms of vectors x and y, respectively. Here, the numerator represents the number of shared substructures, while the denominator captures the total number of unique substructures across both molecules.

In addition to Tanimoto similarity, we also evaluate Cosine similarity, which captures the angular distance between vectors of drugs. For the molecular fingerprints of two drugs x and y, the similarity can be calculated as shown in Equations 2:

$$Sim_C(x,y) = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|^2 \times \|\boldsymbol{y}\|^2} = \frac{\sum \boldsymbol{x_i y_i}}{\sqrt{\sum \boldsymbol{x_i^2}}\sqrt{\sum \boldsymbol{y_i^2}}} \quad (2)$$

Here, this formulation captures the relative orientation between molecular feature vectors. Further, Dice similarity can be calculated as shown in Equations 3:

$$Sim_D(x,y) = \frac{2\sum \boldsymbol{x_i y_i}}{\sum \boldsymbol{x_i^2} + \sum \boldsymbol{y_i^2}} \quad (3)$$

where the dice similarity ranges from 0 to 1. In addition to traditional fingerprint-based similarity measures, we further explore two additional categories of similarity metrics to enhance the selection of ICL examples: graph-based similarity and embedding-based similarity. To explore structural similarity at the graph level, we utilize the Weisfeiler-Lehman graph kernel. Let G_1 and G_2 denote two molecular graphs, and let $\varnothing(G)$ be the mapping of a graph to a high-dimensional feature space based on its structural patterns. The similarity between two embeddings is computed using their dot product, as defined in Equations 4:

$$Sim_G(G_1, G_2) = \langle \varnothing(G_1), \varnothing(G_2) \rangle \quad (4)$$

where $\langle * \rangle$ represents the dot product. This approach captures topological information beyond atom-level fingerprints, enabling graph-level matching when selecting prompts based on molecular structure.

For embedding-based similarity, we leverage pretrained deep learning models to extract SMILES-based embeddings. Given a pair of drugs x and y, their corresponding embedding vectors are denoted as e_x and e_y. The cosine similarity between the embedding vectors is used to compute their similarity, as shown in Equations 5:

$$Sim_E(x,y) = \frac{e_x \cdot e_y}{\|e_x\|^2 \times \|e_y\|^2} \quad (5)$$

This embedding-based metric captures both structural and functional properties encoded during pretraining, providing a complementary perspective to symbolic similarity. In our implementation, the embeddings are generated from SMILES sequences using the pretrained MolBERT (Fabian et al., 2020) model. Finally, We utilize the Tanimoto similarity based on 2048-bit Morgan fingerprints (Morgan, 1965) with a radius of two to calculate molecular scaffold similarity. The similarity score for each candidate drug pair is calculated as the product of the similarity scores of the two drugs. Among the known positive DDI samples, we identify the top-k most similar molecular SMILES pairs to construct positive sample prompts. Similarly, for negative sample prompts, we select the top-k most similar SMILES DDI pairs.

3.2 Building prompts based on ICL

In recent advances in language models, ICL has emerged as a method to enable models to learn tasks without explicit fine-tuning. ICL achieves this by providing examples within the input, allowing the model to understand the task through context and generate accurate outputs. Based on filtered positive and negative sample

You are an experienced pharmacologist with extensive knowledge of drug interactions. Your task is to determine whether there is an interaction between the two drugs based on their pharmacological profiles. Specifically, **you should consider the following factors:**

- **Pharmacodynamics**: How the drugs affect the body, including their effects on receptors and physiological systems.
- **Metabolic Pathways**: How the drugs are metabolized, including enzyme interactions and potential effects on drug metabolism.
- **Receptor Interactions**: Whether the drugs interact with the same or similar receptors.

Task: Given the names and SMILES structures of two drugs, predict if there is an **interaction** between them. You should answer **"yes"** if there is an adverse interaction and **"no"** if there is no adverse interaction.

Requirements:

- **Prediction and Explanation**: Provide a binary prediction ("yes" for an interaction or "no" for no interaction) and a concise explanation grounded in pharmacological evidence.
- **Evidence-Based**: Use reliable sources such as clinical trials, FDA labeling, drug interaction databases, and peer-reviewed literature to support your explanation.
- **Structured Explanation**: Clearly outline the reasoning for your prediction, addressing the pharmacological factors explicitly.

Template:

- Drug A Name: [Name of Drug A]
- Drug A Smiles: [SMILES structure of Drug A]
- Drug B Name: [Name of Drug B]
- Drug B Smiles: [SMILES structure of Drug B]
- **Interaction Prediction**: [yes/no]
- **Explanation**: [Provide a detailed explanation referencing pharmacodynamics, metabolic pathways, receptor interactions, and clinical data.]

Question:

- Drug A Name: Desipramine
- Drug A Smiles: CNCCCN1C2=CC=CC=C2CCC3=CC=CC=C31
- Drug B Name: Entacapone
- Drug B Smiles: CCN(CC)C(=O)C(=CC1=CC(=C(C(=C1)O)O)[N+](=O)[O-])C#N
- **Interaction Prediction**:
- **Explanation:**

FIGURE 2
The zero-shot prompt of DDIs prediction.

examples, we constructed prompts for DDI prediction, which are categorized into zero-shot and few-shot scenarios.

In the zero-shot scenario, the model makes predictions based purely on its pre-trained knowledge, without relying on specific examples. This approach is suited for predicting interactions between novel or previously unseen drug combinations as shown in Figure 2. In contrast, the few-shot scenario provides a small set of examples to help guide the model's predictions, particularly when limited data or related examples are available as shown in Figure 3. The prompt follows a structured format, consisting of several key components: input requirements, prediction task, consideration factors, and examples. The input requirements specify the drug names and their corresponding SMILES structures. The prediction task involves predicting whether an interaction exists between the two drugs, with the outcome being "yes" or "no." Consideration factors include an analysis of pharmacodynamics, metabolic pathways, receptor interactions, and relevant clinical data, including FDA labels and peer-reviewed literature. Finally, the examples section provides a practical demonstration of the input format and expected prediction output, ensuring clarity in applying the model.

3.3 Predicting DDIs based on judging

In this study, we use GPT-4 as a judge to evaluate DDIs prediction generated by multiple LLMs. The discriminator assesses the quality of the explanations provided for each DDI prediction based on four key criteria: scientific accuracy, clarity and coherence, evidence support, and

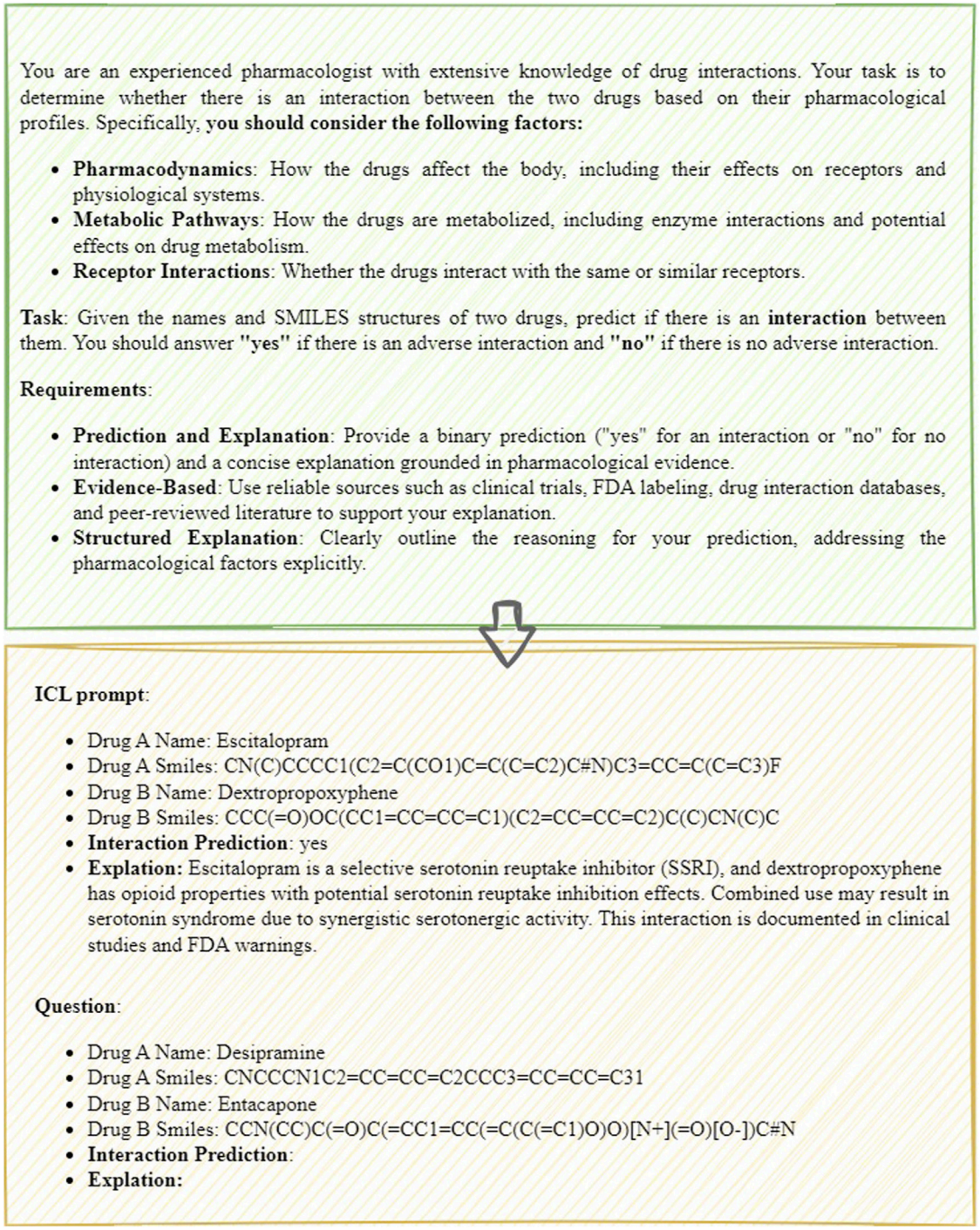

FIGURE 3
The few-shot prompt of DDIs prediction.

relevance. A detailed prompt is designed for both zero-shot and few-shot scenarios as shown in Figure 4. In the zero-shot scenario, the prompt clearly outlines the evaluation criteria and instructions for GPT to assess each prediction and explanation. In the few-shot scenario, the prompt includes several examples of high-quality evaluations to help the model understand how to assign scores. Each explanation is scored on a scale from one to five for each criterion, and an overall score is assigned based on the evaluation. After scoring the results from all models, the predictions are combined using a weighted fusion approach, where each model's score is multiplied by a predetermined weight reflecting its reliability or performance, and the weighted scores are summed to generate the final DDI prediction.

After scoring the results from all models, the predictions are combined using a weighted fusion approach, where the weight w_i

You are an expert pharmacologist tasked with evaluating the quality of an explanation for a drug-drug interaction (DDI) prediction. Below are the results from multiple language models (LLMs) that predict whether there is an interaction between two drugs, including their explanations. Your job is to assess the quality of the explanation based on the following criteria:

1. **Scientific Accuracy**: Does the explanation align with established pharmacological knowledge, including drug mechanisms, metabolic pathways, and receptor interactions?
2. **Clarity and Coherence**: Is the explanation clearly written and easy to understand, without unnecessary complexity or ambiguity?
3. **Evidence Support**: Does the explanation reference known clinical data, pharmacological studies, or reliable sources to justify the prediction?
4. **Relevance**: Does the explanation stay focused on the key pharmacological aspects that could lead to a drug interaction, without introducing irrelevant information?

Please evaluate the explanation for the given DDI prediction by assigning a score from 1 to 5 (1 = very poor, 5 = excellent) for each of the above criteria and provide a brief justification for each score. Afterward, give an overall score (1-5) for the quality of the explanation.

Input:

- Drug A Name: [Name of Drug A]
- Drug A Smiles: [SMILES structure of Drug A]
- Drug B Name: [Name of Drug B]
- Drug B Smiles: [SMILES structure of Drug B]
- **Interaction Prediction**: [yes/no]
- **Explanation**: [Provide a detailed explanation referencing pharmacodynamics, metabolic pathways, receptor interactions, and clinical data.

Question:

Score Criteria:

- **Scientific Accuracy**: [Score] (Justification)
- **Clarity and Coherence**: [Score] (Justification)
- **Evidence Support**: [Score] (Justification)
- **Relevance**: [Score] (Justification)
-
- **Final Evaluation**: Provide a final score from 1 to 5 based on the overall quality of the explanation.

FIGURE 4
The prompt of the DDIs prediction judge.

for each model i is determined by the score given by the discriminator to that model's output S_{model} is calculated as shown in Equations 6:

$$S_{final} = \sum_{i=1}^{N} w_i S_{model} \tag{6}$$

Compared to other ensemble learning techniques such as stacking or boosting, we adopt a weighted fusion strategy to maintain a streamlined, inference-oriented workflow. This approach eliminates the need to train extra models and fits well with LLM workflows that rely mainly on inference rather than supervised training.

4 Results

4.1 Datasets

In the paper, we use the Luo's dataset (Luo et al., 2017) contains the following information, as shown in Table 1. It integrates information from multiple authoritative biomedical sources. Specifically, drug-related information was obtained from DrugBank 3.0 (Craig et al., 2010), protein data from Human Protein Reference Database (Suraj et al., 2004), disease associations from Comparative Toxicogenomics Database (Mattingly et al., 2003), and side-effect information from SIDER (Michael et al., 2016). These heterogeneous entities—including drugs, proteins, diseases, and side effects—were incorporated into a unified heterogeneous

TABLE 1 The detail of LUO's datasets.

Node types	Num	Edge types	Num
Drug	708	Drug-drug	10,036
Protein	1,512	Drug-protein	1,923
Disease	5,603	Protein-protein	7,363
Side Effect	4,192	Drug-Disease	199,214
		Protein-Disease	1,596,745

TABLE 2 The results of three similarity measure.

Similarity	AUC	AUPR
Cosine	0.763	0.799
Dice	0.779	0.787
Tanimoto	0.788	0.801
Graph-based	0.785	0.798
Embedding-based	0.794	0.815

network. The dataset includes 12,015 nodes (708 drugs, 1,512 proteins, 5,603 diseases, and 4,191 side effects) and over 1.89 million edges, with 10,036 known drug-drug interactions. This large-scale, multi-relational structure allows for comprehensive modeling of biomedical interactions.

In our study, we employ cross-validation to evaluate the effectiveness of our proposed method. Specifically, we use a 10-fold cross-validation approach. The dataset is randomly partitioned into ten subsets, from which nine subsets are used for training and the remaining one for testing. This process is repeated ten times, with each subset serving as the test set once. The final performance result is computed as the average of the outcomes from all ten iterations. The final performance is reported as the average of the results across all ten folds, which helps reduce variance due to random partitioning and enables reliable comparison of different models. In the zero-shot scenario, the model is directly tested using the test set. In the few-shot scenario, positive and negative samples are selected from the training set for use in context learning prompts.

4.2 Evaluation criteria

DDI prediction is a classification task where the outcomes are categorized into four types: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Based on these classifications, AUPR (Area Under the Precision-Recall Curve) and AUC (Area Under the ROC Curve) are widely used evaluation metrics. 1) AUC assesses the model's ability to rank true DDIs higher than non-DDIs across all possible thresholds. It reflects the trade-off between the true positive rate (TPR = TP/(TP + FN)) and the false positive rate (FPR = FP/(FP + TN)), providing a comprehensive view of classification performance. 2) AUPR focuses on the balance between precision (TP/(TP + FP)) and recall (TP/(TP + FN)), which is particularly informative in imbalanced datasets such as DDI, where positive examples are much rarer than negatives.

In summary, higher AUC and AUPR values indicate that the model is better at identifying true interactions while minimizing false positives, which is critical in real-world pharmaceutical applications where missing or wrongly predicting DDIs can have serious consequences.

4.3 Comparison models

Given our focus on zero-shot and few-shot scenarios, we primarily selected models based on LLMs. These models include GPT-4 (Achiam et al., 2023), GPT-3.5 (Brown, 2020), Davinci-003, and llama 2 (Touvron et al., 2023b). In addition to these well-established models, we also included more recent state-of-the-art models such as llama 3 (Dubey et al., 2024), GPT-4o, DeepSeek V3 (Liu A. et al., 2024), and Claude 3.5 (Bae et al., 2024). All these models leverage the Transformer architecture, utilizing self-attention mechanisms and large-scale pretraining to achieve efficient generation and understanding of natural language processing tasks through deep learning techniques. Specifically, GPT-4 and GPT-3.5 are known for their advanced reasoning and language understanding capabilities, while Davinci-003 provides a robust foundation for few-shot learning. The inclusion of GPT-4o, DeepSeek V3, and Claude 3.5 ensures that our benchmark is up-to-date with the latest advancements in the field.

4.4 Comparison experiments

First, to analyze the impact of drug similarity on the selection of positive and negative drug pairs as ICL prompts in DDI-JUDGE, we discuss the effects of Cosine similarity, Dice similarity, and Tanimoto similarity on the final results. As shown in Table 2, Tanimoto similarity achieves the best performance, although the differences among the three are minimal. Tanimoto similarity is particularly suitable for drug similarity calculation as it effectively balances shared and unique features, accurately capturing the chemical relationships between drug molecules. To provide a more comprehensive discussion on the role of similarity metrics in DDI-JUDGE, we further evaluate two additional approaches: graph-based similarity and embedding-based similarity. Specifically, we apply the WL graph kernel to compute graph similarity, and use SMILES-based embeddings generated by MolBERT to measure embedding similarity. According to Table 2, the embedding-based method achieves the best overall performance, with an AUC of 0.794 and AUPR of 0.815, surpassing all other methods. The graph-based approach also performs competitively, with results close to those of Tanimoto, indicating that incorporating molecular topology can be beneficial. These results suggest that embedding-based similarity is particularly effective for capturing deeper structural and semantic information.

Overall, while Tanimoto similarity remains the most efficient and effective choice in our method, graph-based and embedding-based similarities present valuable alternatives that can be further explored or integrated in future improvements.

We mainly conducted experiments in two scenarios: zero-shot and few-shot. Zero-shot refers to the model predicting DDIs without any prior training examples or specific task prompts, relying solely on its pre-trained knowledge. Few-shot involves providing the

TABLE 3 The experimental results on the zero-shot scenario.

Methods	AUC	AUPR
GPT-4o	0.585	0.603
GPT-4	0.557	0.581
GPT-3.5	0.521	0.535
Davinci-003	0.443	0.416
llama 2	0.382	0.400
llama 3	0.573	0.551
Claude 3.5	0.536	0.577
DeepSeekV3	0.541	0.589
DDI-JUDGE	0.642	0.629

TABLE 4 The experimental results on the few-shot scenario.

Methods	AUC	AUPR
GPT-4o	0.681	0.643
GPT-4	0.656	0.637
GPT-3.5	0.632	0.622
Davinci-003	0.525	0.553
llama 2	0.417	0.488
llama 3	0.631	0.658
Claude 3.5	0.647	0.626
DeepSeekV3	0.679	0.631
DDI-JUDGE	0.788	0.801

model with a small number of examples, such as known interactions between drugs, to help it understand the task requirements before making predictions.

The experimental results under the zero-shot setting, as shown in Table 3, reveal that DDI-JUDGE demonstrates the best performance among all models, achieving the highest AUC (0.642) and AUPR (0.629). GPT-4o and DeepSeek V3 also perform well with AUC and AUPR values of 0.585/0.557 and 0.603/0.581, respectively. Llama two exhibits relatively weak performance, with an AUC of 0.382 and an AUPR of 0.400.

In the few-shot setting, as presented in Table 4, DDI-JUDGE once again achieves the highest performance, with an AUC of 0.788 and an AUPR of 0.801, showcasing its robustness when provided with a few examples. Davinci-003 and llama two show comparatively weaker performance, with AUC/AUPR values of 0.525/0.553 and 0.417/0.488, respectively. The results demonstrate that DDI-JUDGE effectively leverages few-shot examples to maintain its superior predictive capabilities.

Comparing the two settings, it is evident that all models benefit from the few-shot scenario, as providing a small number of examples improves their performance. DDI-JUDGE shows significant improvement, with its AUC increasing from 0.642 (zero-shot) to 0.768 (few-shot) and its AUPR rising from 0.629 to 0.760. Overall, few-shot learning enhances the models' predictive performance, with DDI-JUDGE maintaining its leading position across both settings.

TABLE 5 The results on different numbers of ICL prompt samples.

Methods	AUC	AUPR
DDI-JUDGE (zero-shot)	0.642	0.629
DDI-JUDGE (n = 1)	0.662	0.671
DDI-JUDGE (n = 2)	0.679	0.694
DDI-JUDGE (n = 4)	0.731	0.752
DDI-JUDGE (n = 8)	0.788	0.801

4.5 The impact of the number of ICL prompt samples

Next, we discussed the impact of different numbers of ICL prompt samples on the predictive performance, as shown in Table 5. As the number of ICL prompt samples increases, DDI-JUDGE's performance improves significantly. Starting from zero-shot, the AUC and AUPR steadily rise as more prompt samples are provided, with the best performance achieved when eight samples are used. These results suggest that increasing the number of ICL prompt samples provides more contextual information, allowing the model to better understand the task and make more accurate predictions.

4.6 Case study

Nowadays, an increasing number of studies are exploring whether methods can be directly translated into practical improvements in real-world drug discovery and are conducting relevant experiments (Zhao et al., 2022; Wang et al., 2025a; Wang et al., 2025b; Zhao et al., 2024; Zhao et al., 2025). To demonstrate the capability of our method in addressing real-world drug discovery issues, we conducted experiments and identified several DDIs that are not present in the DrugBank database.

1) When Rivaroxaban is used concomitantly with Dihydroxyaluminum Sodium Carbonate, the anticoagulant effect of Rivaroxaban may be compromised due to the potential for increased gastrointestinal bleeding in patients with gastroduodenal ulcers (Goldhaber, 2020).
2) When romidepsin is used concomitantly with quinidine, the risk or severity of QT interval prolongation may be increased. Romidepsin, a histone deacetylase inhibitor, is employed in the treatment of certain types of lymphoma; quinidine is an antiarrhythmic agent (Abu Rmilah et al., 2020).
3) Simvastatin is a cholesterol-lowering drug that works by inhibiting the enzyme HMG-CoA reductase. Fluconazole is a triazole antifungal agent. Studies have shown that the concurrent use of these two drugs may increase the risk of myopathy or rhabdomyolysis (Molden et al., 2008).

The case studies demonstrate the capacity of DDI-JUDGE to identify novel DDIs. Consequently, DDI-JUDGE exerts a beneficial influence on the design and development process of new drugs.

5 Conclusion

In this paper, we propose an LLM-based method for DDI prediction, which is achieved through the integration of judging and ICL prompts. The proposed method outperforms existing LLM approaches, demonstrating the potential of LLMs for predicting complex relationships in drug molecules.

First, we propose a novel ICL prompt paradigm for DDI prediction. This approach selects high-similarity samples as positive and negative prompts, enabling the LLM to effectively learn and generalize knowledge. Additionally, we introduce an ICL-based prompt template that organizes structured prompts, including input requirements, prediction tasks, relevant factors, and examples. By leveraging the pre-trained knowledge and contextual understanding of LLMs, this template enhances DDI prediction capabilities. Finally, we employ GPT-4 as a discriminator to assess the predictions of multiple LLMs based on scientific accuracy, clarity, evidence support, and relevance. These individual results are then combined through a weighted fusion method to improve prediction accuracy.

In addition, this study emphasizes zero-shot and few-shot prompting scenarios, which reflect the practical challenges of real-world DDI prediction, where labeled data are often scarce. As shown in our analysis, performance improves as the number of prompt examples increases, including the one-shot setting. Many-shot prompting, although potentially beneficial, was not explored further due to input length limitations and diminishing marginal gains. These findings highlight that zero-shot and few-shot prompting offer an effective and scalable approach to DDI prediction in settings with limited labeled data.

The method currently has the following limitations: While it explores the potential of applying LLMs to DDI prediction, there is still a lack of domain-specific drug knowledge. For example, GPT-4, as a discriminator, may introduce potential biases due to its inability to fully understand domain-specific knowledge and scientific context. Future work will incorporate more drug-related data and perform fine-tuning to further optimize the performance.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

HQ: Writing – original draft, Data curation. XL: Writing – original draft. CZ: Writing – review and editing. TZ: Writing – review and editing, Project administration.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was funded by National Key R&D Program (2022YFC3321103).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Abu Rmilah, A. A., Lin, G., Begna, K. H., Friedman, P. A., and Herrmann, J. (2020). Risk of QTc prolongation among cancer patients treated with tyrosine kinase inhibitors. *Int. J. cancer* 147 (11), 3160–3167. doi:10.1002/ijc.33119

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). Gpt-4 technical report. doi:10.48550/arXiv.2303.08774

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J. big Data* 8, 53–74. doi:10.1186/s40537-021-00444-8

Bae, J., Kwon, S., and Myeong, S. (2024). Enhancing software code vulnerability detection using gpt-4o and claude-3.5 sonnet: a study on prompt engineering techniques. *Electronics* 13 (13), 2657. doi:10.3390/electronics13132657

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. (2022). Constitutional ai: harmlessness from ai feedback. doi:10.48550/arXiv.2212.08073

Brown, T. B. (2020). Language models are few-shot learners. doi:10.48550/ARXIV.2005.14165

Chaves, J. M. Z., Wang, E., Tu, T., Vaishnav, E. D., Lee, B., Mahdavi, S. S., et al. (2024). Tx-LLM: a large language model for therapeutics. doi:10.48550/arXiv.2406.06316

Craig, K., Vivian, L., Timothy, J., Philip, L., Son, L., Alex, F., et al. (2010). DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res.* 39 (Database issue), D1035–D1041. doi:10.1093/nar/gkq1126

Deng, Y., Qiu, Y., Xu, X., Liu, S., Zhang, Z., Zhu, S., et al. (2022). META-DDIE: predicting drug–drug interaction events with few-shot learning. *Briefings Bioinforma.* 23 (1), bbab514. doi:10.1093/bib/bbab514

Deng, Y., Xu, X., Qiu, Y., Xia, J., Zhang, W., and Liu, S. (2020). A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics* 36 (15), 4316–4322. doi:10.1093/bioinformatics/btaa501

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., et al. (2022). A survey on in-context learning. doi:10.48550/arXiv.2301.00234

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., et al. (2024). The llama 3 herd of models. doi:10.48550/arXiv.2407.21783

Edwards, C., Zhai, C., and Ji, H. (2021). "Text2mol: cross-modal molecule retrieval with natural language queries," in *Proceedings of the 2021 conference on empirical methods in natural language processing*, 595–607. doi:10.18653/v1/2021.emnlp-main.47

Fabian, B., Edlich, T., Gaspar, H., Segler, M., and Ahmed, M. (2020). Molecular representation learning with language models and domain-relevant auxiliary tasks. doi:10.48550/arXiv.2011.13230

Fang, Y., Liang, X., Zhang, N., Liu, K., Huang, R., Chen, Z., et al. (2023). Mol-instructions: a large-scale biomolecular instruction dataset for large language models. 2023. doi:10.48550/arXiv.2306.08018

Goldhaber, S. Z. (2020). Thromboembolism prophylaxis for patients discharged from the hospital: easier said than done. *Am. Coll. Cardiol. Found. Wash. D.C.* 75, 3148–3150. doi:10.1016/j.jacc.2020.05.023

Guo, B., Wang, H., Xiao, W., Chen, H., Lee, Z., Han, S., et al. (2024). Sample design engineering: an empirical study of what makes good downstream fine-tuning samples for LLMs. doi:10.48550/arXiv.2404.13033

Guo, T., Nan, B., Liang, Z., Guo, Z., Chawla, N., Wiest, O., et al. (2023). What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Adv. Neural Inf. Process. Syst.* 36, 59662–59688. doi:10.48550/arXiv.2305.18365

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al. (2021). Lora: low-rank adaptation of large language models. *arXiv Prepr.* doi:10.48550/arXiv.2106.09685

Huang, J., Niu, C., Green, C. D., Yang, L., Mei, H., and Han, J.-D. J. (2013). Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network. *PLoS Comput. Biol.* 9 (3), e1002998. doi:10.1371/journal.pcbi.1002998

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D., et al. (2023). Mistral 7B. arXiv preprint arXiv:231006825. doi:10.48550/arXiv.2310.06825

Li, J., Liu, Y., Fan, W., Wei, X.-Y., Liu, H., Tang, J., et al. (2024). Empowering molecule discovery for molecule-caption translation with large language models: a chatgpt perspective. *IEEE Trans. Knowl. Data Eng.* 36, 6071–6083. doi:10.1109/tkde.2024.3393356

Liang, Y., Zhang, R., Zhang, L., and Xie, P. (2023). Drugchat: towards enabling chatgpt-like capabilities on drug molecule graphs. doi:10.48550/arXiv.2309.03907

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., et al. (2024b). Deepseek-v3 technical report. doi:10.48550/arXiv.2412.19437

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023b). Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55 (9), 1–35. doi:10.1145/3560815

Liu, S., Zhang, Y., Cui, Y., Qiu, Y., Deng, Y., Zhang, Z., et al. (2022). Enhancing drug-drug interaction prediction using deep attention neural networks. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 20 (2), 976–985. doi:10.1109/TCBB.2022.3172421

Liu, Y., Ding, S., Zhou, S., Fan, W., and Tan, Q. (2024a). MolecularGPT: open large language model (LLM) for few-shot molecular property prediction. doi:10.48550/arXiv.2406.12950

Liu, Z., Zhang, W., Xia, Y., Wu, L., Xie, S., Qin, T., et al. (2023a). Molxpt: wrapping molecules with text for generative pre-training. *arXiv Prepr. arXiv:230510688*, 1606–1616. doi:10.18653/v1/2023.acl-short.138

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* 8 (1), 573. doi:10.1038/s41467-017-00680-8

Ma, T., Lin, X., Li, T., Li, C., Chen, L., Zhou, P., et al. (2024). Y-mol: a multiscale biomedical knowledge-guided large language model for drug development. doi:10.48550/arXiv.2410.11550

Mattingly, C. J., Colby, G. T., Forrest, J. N., and Boyer, J. L. (2003). The comparative Toxicogenomics database (CTD). *Environ. Health Perspect.* 111 (6), 793–795. doi:10.1289/ehp.6028

Michael, K., Ivica, L., Juhl, J. L., and Peer, B. (2016). The SIDER database of drugs and side effects. *Nucleic Acids Res.* 44 (D1), D1075–D1079. doi:10.1093/nar/gkv1075

Molden, E., Skovlund, E., and Braathen, P. (2008). Risk management of simvastatin or atorvastatin interactions with CYP3A4 inhibitors. *Drug Saf.* 31, 587–596. doi:10.2165/00002018-200831070-00004

Morgan, H. L. (1965). The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chem. documentation* 5 (2), 107–113. doi:10.1021/c160017a018

Nyamabo, A. K., Yu, H., and Shi, J.-Y. (2021). SSI–DDI: substructure–substructure interactions for drug–drug interaction prediction. *Briefings Bioinforma.* 22 (6), bbab133. doi:10.1093/bib/bbab133

Rdkit, L. G. (2013). A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* 8 (31.10), 5281.

Ryu, J. Y., Kim, H. U., and Lee, S. Y. (2018). Deep learning improves prediction of drug–drug and drug–food interactions. *Proc. Natl. Acad. Sci.* 115 (18), E4304-E4311–E4311. doi:10.1073/pnas.1803294115

Shi, J.-Y., Huang, H., Li, J.-X., Lei, P., Zhang, Y.-N., Dong, K., et al. (2018). TMFUF: a triple matrix factorization-based unified framework for predicting comprehensive drug-drug interactions of new drugs. *BMC Bioinforma.* 19, 411–437. doi:10.1186/s12859-018-2379-8

Shi, Y., He, M., Chen, J., Han, F., and Cai, Y. (2024). SubGE-DDI: a new prediction model for drug-drug interaction established through biomedical texts and drug-pairs knowledge subgraph enhancement. *PLOS Comput. Biol.* 20 (4), e1011989. doi:10.1371/journal.pcbi.1011989

Sridhar, D., Fakhraei, S., and Getoor, L. (2016). A probabilistic approach for collective similarity-based drug–drug interaction prediction. *Bioinformatics* 32 (20), 3175–3182. doi:10.1093/bioinformatics/btw342

Suraj, P., Daniel, N. J., Kristiansen, T. Z., Ramars, A., Vineeth, S., Babylakshmi, M., et al. (2004). Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.* 32 (Suppl. l_1), D497–D501. doi:10.1093/nar/gkh070

Sze, V., Chen, Y.-H., Yang, T.-J., and Emer, J. S. (2017). Efficient processing of deep neural networks: a tutorial and survey. *Proc. IEEE* 105 (12), 2295–2329. doi:10.1109/jproc.2017.2761740

Tanimoto, T. T. (1958). Elementary mathematical theory of classification and prediction.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023a). LLaMA: open and efficient foundation language models. doi:10.48550/arXiv.2302.13971

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023b). Llama 2: open foundation and fine-tuned chat models. doi:10.48550/arXiv.2307.09288

van Roon, E. N., Flikweert, S., le Comte, M., Langendijk, P. N., Kwee-Zuiderwijk, W. J., Smits, P., et al. (2005). Clinical relevance of drug-drug interactions: a structured assessment procedure. *Drug Saf.* 28, 1131–1139. doi:10.2165/00002018-200528120-00007

Vaswani, A. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* doi:10.48550/arXiv.1706.03762

Wan, F., Huang, X., Cai, D., Quan, X., Bi, W., and Shi, S. (2024). Knowledge fusion of large language models. doi:10.48550/arXiv.2401.10491

Wang, J., Feng, J., Kang, Y., Pan, P., Ge, J., Wang, Y., et al. (2025b). Discovery of antimicrobial peptides with notable antibacterial potency by an LLM-based foundation model. *Sci. Adv.* 11 (10), eads8932. doi:10.1126/sciadv.ads8932

Wang, J., Luo, H., Qin, R., Wang, M., Wan, X., Fang, M., et al. (2025a). 3DSMILES-GPT: 3D molecular pocket-based generation with token-only large language model. *Chem. Sci.* 16 (2), 637–648. doi:10.1039/d4sc06864e

Wei, J., Zhuo, L., Fu, X., Zeng, X., Wang, L., Zou, Q., et al. (2024). DrugReAlign: a multisource prompt framework for drug repurposing based on large language models. *BMC Biol.* 22 (1), 226. doi:10.1186/s12915-024-02028-3

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Trans. neural Netw. Learn. Syst.* 32 (1), 4–24. doi:10.1109/TNNLS.2020.2978386

Xiong, Z., Liu, S., Huang, F., Wang, Z., Liu, X., Zhang, Z., et al. (2023). Multi-relational contrastive learning graph neural network for drug-drug interaction event prediction. *Proc. AAAI Conf. Artif. Intell.* 37 (4), 5339–5347. doi:10.1609/aaai.v37i4.25665

Xu, M., Yuan, X., Miret, S., and Tang, J. (2023). "Protst: multi-modality learning of protein sequences and biomedical texts," in *International conference on machine learning*. Honolulu, HI: PMLR, 38749–38767.

Yu, H., Mao, K.-T., Shi, J.-Y., Huang, H., Chen, Z., Dong, K., et al. (2018). Predicting and understanding comprehensive drug-drug interactions via semi-nonnegative matrix factorization. *BMC Syst. Biol.* 12, 14–110. doi:10.1186/s12918-018-0532-7

Yu, H., Zhao, S., and Shi, J. (2022). Stnn-ddi: a substructure-aware tensor neural network to predict drug–drug interactions. *Briefings Bioinforma.* 23 (4), bbac209. doi:10.1093/bib/bbac209

Zhang, C., Lu, Y., and Zang, T. (2022). CNN-DDI: a learning-based method for predicting drug–drug interactions using convolution neural networks. *BMC Bioinforma.* 23 (Suppl. 1), 88. doi:10.1186/s12859-022-04612-2

Zhang, C., Zang, T., and Zhao, T. (2024). KGE-UNIT: toward the unification of molecular interactions prediction based on knowledge graph and multi-task learning on drug discovery. *Briefings Bioinforma.* 25 (2), bbae043. doi:10.1093/bib/bbae043

Zhang, P., Wang, F., Hu, J., and Sorrentino, R. (2015). Label propagation prediction of drug-drug interactions based on clinical side effects. *Sci. Rep.* 5 (1), 12339. doi:10.1038/srep12339

Zhao, B.-W., Su, X.-R., Hu, P.-W., Ma, Y.-P., Zhou, X., and Hu, L. (2022). A geometric deep learning framework for drug repositioning over heterogeneous information networks. *Briefings Bioinforma.* 23 (6), bbac384. doi:10.1093/bib/bbac384

Zhao, B.-W., Su, X.-R., Yang, Y., Li, D.-X., Li, G.-D., Hu, P.-W., et al. (2024). A heterogeneous information network learning model with neighborhood-level structural representation for predicting lncRNA-miRNA interactions. *Comput. Struct. Biotechnol. J.* 23, 2924–2933. doi:10.1016/j.csbj.2024.06.032

Zhao, B.-W., Su, X.-R., Yang, Y., Li, D.-X., Li, G.-D., Hu, P.-W., et al. (2025). Regulation-aware graph learning for drug repositioning over heterogeneous biological network. *Inf. Sci.* 686, 121360. doi:10.1016/j.ins.2024.121360