PKAG-DDI: Pairwise Knowledge-Augmented Language Model for Drug-Drug Interaction Event Text Generation

Ziyan Wang[1]*, Zhankun Xiong[1]*, Feng Huang[1], Wen Zhang[123]†,
[1]College of Informatics, Huazhong Agricultural University, Wuhan, China
[2]Hubei Key Laboratory of Agricultural Bioinformatics,
Huazhong Agricultural University, Wuhan, China
[3]Engineering Research Center of Intelligent Technology for Agriculture,
Ministry of Education, Wuhan, China
{wangziyan,xiongzk,fhuang233}@webmail.hzau.edu.cn, zhangwen@mail.hzau.edu.cn

Abstract

Drug-drug interactions (DDIs) arise when multiple drugs are administered concurrently. Accurately predicting the specific mechanisms underlying DDIs (named DDI events or DDIEs) is critical for the safe clinical use of drugs. DDIEs are typically represented as textual descriptions. However, most computational methods focus more on predicting the DDIE class label over generating human-readable natural language increasing clinicians' interpretation costs. Furthermore, current methods overlook the fact that each drug assumes distinct biological functions in a DDI, which, when used as input context, can enhance the understanding of the DDIE process and benefit DDIE generation by the language model (LM). In this work, we propose a novel pairwise knowledge-augmented generative method (termed PKAG-DDI) for DDIE text generation. It consists of a pairwise knowledge selector efficiently injecting structural information between drugs bidirectionally and simultaneously to select pairwise biological functions from the knowledge set, and a pairwise knowledge integration strategy that matches and integrates the selected biological functions into the LM. Experiments on two professional datasets show that PKAG-DDI outperforms existing methods in DDIE text generation, especially in challenging inductive scenarios, indicating its practicality and generalization[1].

1 Introduction

Unexpected drug-drug interactions (DDIs) may arise when people take multiple drugs simultaneously to treat complex diseases and potentially induce diverse pharmacokinetic and pharmacodynamic consequences, named DDI events (DDIEs). Predicting DDIEs holds significant importance for public health security and clinical research (Ryu et al., 2018).

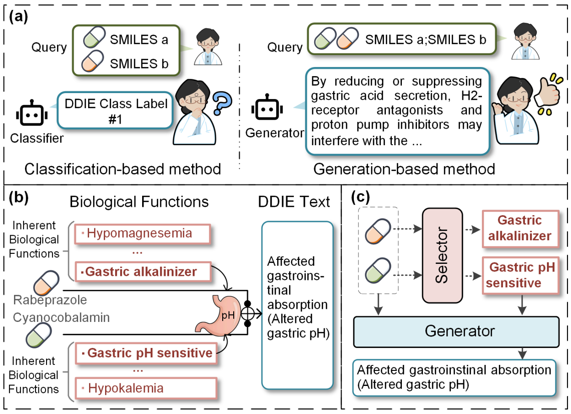

(a)
Query
SMILES a
SMILES b
DDIE Class Label #1
Classifier
Classification-based method
Query
SMILES a;SMILES b
By reducing or suppressing gastric acid secretion, H2-receptor antagonists and proton pump inhibitors may interfere with the ...
Generator
Generation-based method
(b)
Biological Functions
DDIE Text
Inherent Biological Functions
· Hypomagnesemia
...
·Gastric alkalinizer
Rabeprazole
Cyanocobalamin
pH
Affected gastroinstinal absorption (Altered gastric pH)
· Gastric pH sensitive
...
· Hypokalemia
Inherent Biological Functions
(c)
Selector
Gastric alkalinizer
Gastric pH sensitive
Generator
Affected gastroinstinal absorption (Altered gastric pH)

Figure 1: (a) The difference between classification-based and generation-based methods. (b) An example of a DDI. Every drug has several inherent biological functions. In a DDI, it has the most DDIE-relevant biological function. (c) When there are no prior biological functions, our method can use the selector to gather relevant biological functions to augment the DDIE prediction.

Currently, most computational-based DDIE prediction methods (Ryu et al., 2018; Xiong et al., 2023; Wang et al., 2024a) view the DDIE prediction as a classification task, which entails categorizing various DDIE texts into a finite number of predefined classes. On the one hand, predicting labels lacks intuitiveness, and predicting labels needs an extra label-to-text list to obtain corresponding DDIE information. On the other hand, the latest DDI databases, such as DDInter2.0 (Tian et al., 2024), provide more detailed DDIE texts, which are not readily classifiable. Therefore, as shown in Figure 1 (a), in DDIE prediction, directly generating DDIE texts is a more natural way, highlighting the critical need to develop generation-based methods.

Language models (LMs) have shown notable success in general text generation tasks (Brown

*Equal contribution.

†Corresponding author.

[1]Our data and source code are available at https://github.com/wzy-Sarah/PKAG-DDI

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10996–11010
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

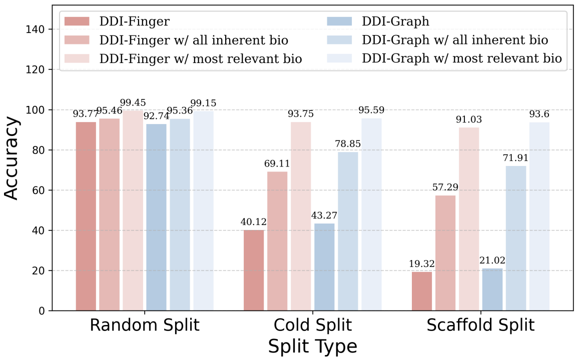


Figure 2: The multi-class classification performance of two basic models (DDI-Finger and DDI-Graph) on the MecDDI dataset. These models with the biological function as additional information (w/ all inherent bio, w/ most relevant bio) have remarkable improvement under three common DDIE prediction scenarios. More details are shown in Appendix A.

et al., 2020; Luo et al., 2022). A recent method MolTC (Fang et al., 2024) introduces an LM framework for molecular relational learning from molecular structure and tries to tackle the DDIE generation task. The strategy that infers the DDIE from the molecular structures of drugs may overlook the fact that each drug in a DDI has distinct biological functions (including drug mechanisms, activities, and categories) (Hu et al., 2023). The biological functions can serve as a context to explicitly supplement and explain biological logical inference processes when DDI occurs. For example, as depicted in Figure 1 (b), Rabeprazole, as *gastric alkalinizer*[2] can influence the activity of Cyanocobalamin, which is *gastric pH sensitive*, thereby leading to gastrointestinal malabsorption caused by changes in gastric pH. The quantitative analysis shown in Figure 2 confirmed that the biological function information can significantly enhance the accuracy of DDIE prediction. Thus, incorporating biological functions holds promise for improving the capability of LM for DDIE text generation. However, biological functions are specialized knowledge that may not be readily accessible, potentially limiting widespread practical applications.

To tackle the above problem, we noticed retrieval-augmented generation (RAG), which has emerged as a popular knowledge augment technique for LM (Gao et al., 2023; Fan et al., 2024). RAG leverages a retriever to extract several pieces of query-relevant knowledge from external databases when specialized knowledge is lacking and integrates the knowledge and original input to enhance the generation of LM. However, applying it to our task faces two challenges. (1) Every single drug has multiple inherent biological functions, and when interacting with another drug, it will have at least one most DDIE-relevant biological function (see Figure 1 (b)). As Figure 2 shows that the most DDIE-relevant biological function is the key to the performance gains on DDIE prediction. According to this, the first challenge is how to model the mutual-conditioned DDIE-relevant biological function (named pairwise biological function) selection while maintaining efficiency. (2) Simply integrating all selected potential pairwise biological functions may introduce additional interference to LM generation, such as the noise from unmatched or irrelevant biological functions. How to design an effective biological function integration strategy is the second challenge.

In this work, we propose a novel pairwise knowledge-augmented generative method (PKAG-DDI) for DDIE generation, which selects the pairwise biological functions from the external knowledge set and takes them as context to enhance the DDIE generation of LM. Specifically, PKAG-DDI first designed a pairwise knowledge selector (PKS), which enables the mutual injection of molecular structural information between two drugs. Within the PKS, a reuse strategy is also introduced, sharing the weight of most components to reduce the computational burden. Then PKAG-DDI proposed a tailored pairwise knowledge integration strategy, which matches the selected biological functions and integrates them selectively into LM, boosting the LM's capability in generating DDIE texts.

Generally speaking, the main contributions of this paper are described as follows:

- We are the first to take the biological functions of drugs as the input context to augment LM for DDIE generation and propose a novel pairwise knowledge-augmented generative model (PKAG-DDI) for DDIE generation in real application scenarios, which can explicitly reveal the logical process underlying DDI occurrence.

- We introduce a pairwise knowledge selector to efficiently select the pairwise DDIE-relevant biological functions from a knowledge set and a pairwise knowledge integration strategy to

[2] Herein, the italics refer to biological functions, which can explain, to some extent, why concurrent administration of Rabeprazole and Cyanocobalamin may impact gastrointestinal absorption.

match and inject pairwise knowledge into LM for accurate DDIE generation.

- The extensive experiments on two professional datasets show that PKAG-DDI outperforms existing methods in DDIE generation, especially in challenging inductive scenarios, indicating its practicality and generalization.

2 Related Works

2.1 Drug-Drug Interaction Event Prediction

Current DDIE prediction methods generally focus on classifying drug pair instances through label prediction, collectively called classification-based methods. Some of them construct a drug association network and subsequently employ GNNs to aggregate interaction information for DDIE prediction (Xiong et al., 2023; Chen et al., 2021; Wang et al., 2024a), while others focus on developing efficient feature encoders (e.g., DNNs, GNNs) to learn drug pair representations from molecular identity information, including SMILES (Simplified Molecular Input Line Entry System) strings, fingerprints, 2D structures, and 3D structures for DDIE prediction (Nyamabo et al., 2021, 2022; Li et al., 2023; Zhu et al., 2022; He et al., 2022). Given that extracting identity information does not need global interaction information, the latter methods are more suitable for generalizing to the challenging inductive DDIE prediction scenario, where all test drugs are absent from the training set. Overall, the predictions of all classification-based methods lack intuitiveness and are restricted by predefined class boundaries.

Recently, language models (LMs) have gained widespread adoption and offer novel perspectives for diverse biomedical applications, such as MolT5 (Edwards et al., 2022) and MolTC (Fang et al., 2024), which are devoted to translating biochemical language into readable natural language. In particular, MolTC leverages a cross-modal adapter combined with a pre-trained language model (LM) for molecular interaction prediction. However, these methods do not explicitly capture the individual drug's biological function within DDI.

2.2 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) (Lewis et al., 2020) invokes a retriever to search and extract input query-relevant domain knowledge (such as facts or documents in a corpus) from external sets, and a generator that leverages the query alongside the knowledge to augment the generation (Gao et al., 2023; Fan et al., 2024). Dense retrieval (Karpukhin et al., 2020; Lin et al., 2023a; Izacard et al., 2024) embeds queries and external knowledge into continuous vector spaces and calculates relevant scores between them to get a ranked retrieved knowledge list, and can meet our needs for cross-modal retrieval (i.e., molecules to biological functions). However, since biological functions are domain-specific phrases and their scale is relatively small compared to large corpora, we streamline the retrieval process to a biological function selection/prediction task in our work, thereby reducing the computational burden, which voids the need for knowledge encoding knowledge by Bag of Words or BERT (Singh et al., 2021). Moreover, existing retrieval methods are not directly applicable to the pairwise retrieval objects in our task, necessitating further design.

In addition, there are two commonly used input integration strategies for injecting the top-K knowledge texts into generation. The first is concatenating the input query and all retrieved knowledge into a single prompt for the generation model (Ram et al., 2023). This strategy may confuse LMs with irrelevant information (Xu et al., 2024). The second is concatenating each top-K knowledge to the input, respectively, and ensembling output probabilities from all K knowledge (Lin et al., 2023b; Shi et al., 2024). However, they cannot tackle our pairwise knowledge task.

3 Methodology

3.1 Problem Formulation

The problem we aim to solve consists of stages: pairwise knowledge selector (PKS) and pairwise knowledge-augmented LM (PKA-LM). For selecting, given a drug pair a and b and a biological function set $\mathcal{C}$, the selector with parameters θ is to model two distribution $p_\theta(c^a|(a,b))$ and $p_\theta(c^b|(b,a))$ over the biological functions for drug a and drug b, respectively. Here, we denote the primary drug a interacting with another drug b as (a,b) and vice versa (b,a). Note that (a,b) and (b,a) have their respective biological functions but lead to the same DDIE. $c^a, c^b \in \mathcal{C}$. For DDIE generation, given the input pairwise drugs and their potential pairwise biological functions, the generator $p_\eta(y|a,b,c^a,c^b)$ with parameters η is to generate DDIE text y.

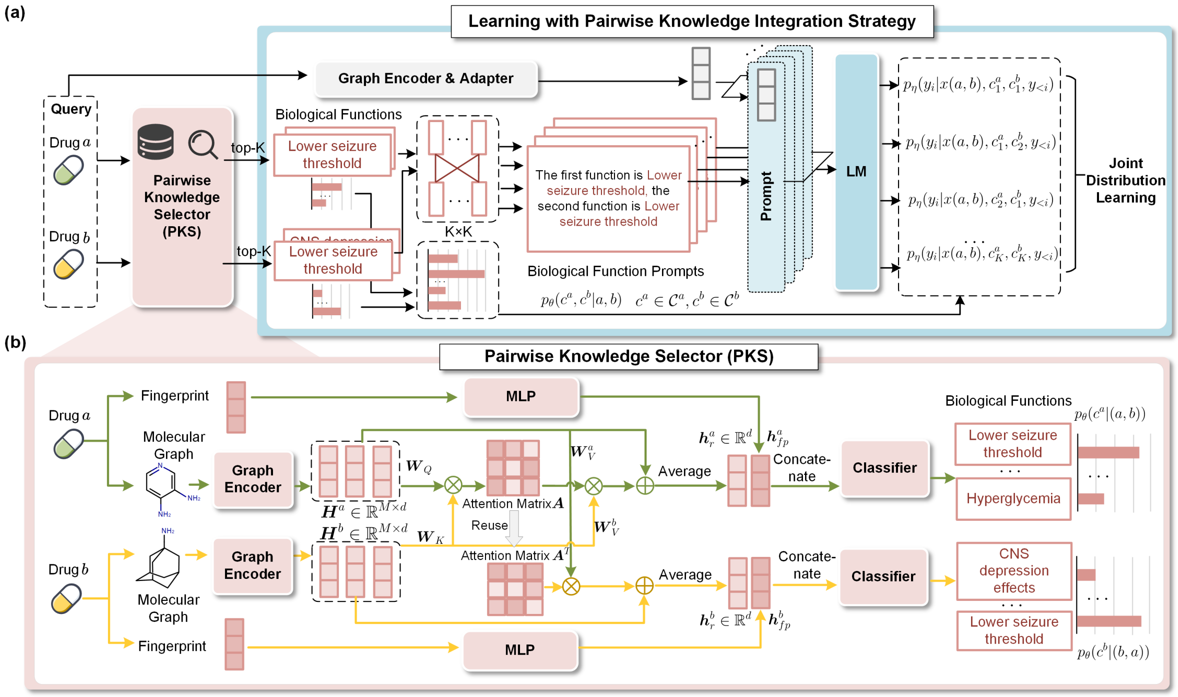


Figure 3: The overall framework of PKAG-DDI. (a) is the overall framework of PKAG-DDI. It firstly uses a PKS to select top-K potential biological functions of each drug from the knowledge base and then match all possible pairwise biological functions and corresponding probability scores. Afterward, the matched pairwise biological functions together with the query information are put into LM for joint distribution learning by marginalizing all $K \times K$ pairwise biological functions. (b) PKS predicts K biological functions of each drug.

3.2 Pairwise Knowledge Selector (PKS)

Model Architecture. PKS with the parameters θ aims to model two distributions $p_\theta(c^a|(a,b))$ and $p_\theta(c^b|(b,a))$, which first learn drug representations of a and b, respectively, and then use two classifiers to predict the pairwise biological functions. The process is shown in Figure 3 (b), with detailed descriptions provided below.

Given a drug a, we first convert its SMILES to two commonly used initial molecular modalities by an RDKit tool: a fingerprint (Glen et al., 2006) and a molecular graph (i.e., atoms as nodes and bonds as edges). The molecular graph is then input into a graph encoder (Wang et al., 2024b) to get the molecular node prototype representations $\boldsymbol{H}^a \in \mathbb{R}^{M\times d}$, where M is the predefined number of node prototype and d is the dimension. Similarly, we get the $\boldsymbol{H}^b \in \mathbb{R}^{M\times d}$ for drug b. Then, $\boldsymbol{H}^b$ is used as conditional information and injected into $\boldsymbol{H}^a$ as the following cross-attention:

$$\boldsymbol{A} = \frac{\boldsymbol{H}^a\boldsymbol{W}_Q \cdot \boldsymbol{H}^b\boldsymbol{W}_K{}^T}{\sqrt{d}}, \tag{1}$$

$$\boldsymbol{H}_r^a = \boldsymbol{H}^a + \lambda \cdot Softmax\left(\boldsymbol{A}\right)\left(\boldsymbol{H}^b\boldsymbol{W}_V^b\right), \tag{2}$$

where $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$ and $\boldsymbol{W}_V^b$ are the learnable projection matrices. In this work, $\boldsymbol{A}$ denotes the attention scores between the node prototype representations of drug a and drug b. $\boldsymbol{H}_r^a \in \mathbb{R}^{M\times d}$ is the information-injected node prototype representation of drug a, which is calculated by residual connection with attention-transferred $\boldsymbol{H}^b\boldsymbol{W}_V^b$ of drug b. $Softmax$ is the softmax operator, λ is a hyper-parameter controlling the information flow from drug b. After averaging the rows of $\boldsymbol{H}_r^a$, we get the drug representation $\boldsymbol{h}_r^a \in \mathbb{R}^d$ for drug a. Distinctively, considering that constructing the $p_\theta(c^a|(a,b))$ and $p_\theta(c^b|(b,a))$ respectively may lead to redundant computations of the graph encoder, we propose a reuse strategy: reusing node prototype representations $\boldsymbol{H}^a$ and the attention scores $\boldsymbol{A}$ for computing $\boldsymbol{H}_r^b$ simultaneously with $\boldsymbol{H}_r^a$. To be more specific, the $\boldsymbol{A}$ can model the importance of drug b, meanwhile the transposed $\boldsymbol{A}^T$ can model the importance of drug a, thus we inject the information of drug a for drug b by:

$$\boldsymbol{H}_r^b = \boldsymbol{H}^b + \lambda \cdot Softmax(\boldsymbol{A}^T)(\boldsymbol{H}^a\boldsymbol{W}_V^a), \tag{3}$$

where $\boldsymbol{W}_V^a$ is the learnable projection matrix. In a similar vein, we can get the drug representation $\boldsymbol{h}_r^b \in \mathbb{R}^d$ for drug b.

Subsequently, the $\boldsymbol{h}_r^a$ is concatenated with the fingerprint $\boldsymbol{h}_{fp}^a$, which has been transformed by a Multi-Layer Perceptron (MLP). Then, an MLP classifier is used to calculate the logits of (a, b) to c^a:

$$r(c^a|(a,b)) = MLP_{c^a}(\boldsymbol{h}_r^a||\boldsymbol{h}_{fp}^a). \tag{4}$$

Similarly, we get $r(c^b|(b,a))$, where r refers to the relevant scores between a drug pair and a biological function. After that, a $Softmax$ is used to obtain the probability, i.e., $p_\theta(c^a|(a,b))$ and $p_\theta(c^b|(b,a))$ among all biological functions in the set.

Training. We employ two negative log-likelihood losses for learning the distribution of $p_\theta(c^a|(a,b))$ and $p_\theta(c^b|(b,a))$, simultaneously. Note that since a drug pair may have more than one DDIE-relevant biological function, to simplify the learning object, we use the BM25 (Robertson et al., 2009) to select the most similar pairwise biological function to the corresponding DDIE text as the gold biological function labels for training. The reason is illustrated in Appendix B.

3.3 Learning with Pairwise Knowledge Integration Strategy

Model Architecture. Given drugs a and b as query input and their potential biological functions as condition, the generator $p_\eta(y|a,b,c^a,c^b)$ with parameters η is to generate DDIE text y. In particular, we propose a pairwise knowledge integration strategy to inject biological function information into the LM effectively. Hereafter, we detail the model architecture.

To construct the query input of LM, given drug pair a and b, we first use the molecular graph encoder and the graph-to-sequence adapter from MolTC (Fang et al., 2024) to encode the molecular graphs of drug pair a and b to molecular token embeddings in text space, i.e., $\boldsymbol{T}^a \in \mathbb{R}^{Q\times d_t}$ and $\boldsymbol{T}^a \in \mathbb{R}^{Q\times d_t}$ for drug a and b , respectively. Q is the number of tokens and d_t is the dimension. Consistent with (Fang et al., 2024), we also input the SMILES tokens of the drug pair $\boldsymbol{S}^a$ and $\boldsymbol{S}^b$ to LM. Thus, the query input x is formulated as:

$$x = x(a,b) = Prompt(\boldsymbol{S}^a, \boldsymbol{S}^b, \boldsymbol{T}^a, \boldsymbol{T}^b), \tag{5}$$

where $Prompt$ is the prompt text with the slots for $\boldsymbol{S}^a$, $\boldsymbol{S}^b$, $\boldsymbol{T}^a$ and $\boldsymbol{T}^b$.

To effectively and selectively integrate biological function information, we aim to model the target $p(y|a,b)$ by a joint probability distribution that marginalizes all latent pairwise biological functions c^a and c^b based on (Guu et al., 2020) rather than directly input all biological functions to LM:

$$p(y|a,b) = \sum_{c^a,c^b\in\mathcal{C}} p_\theta(c^a,c^b|a,b)p_\eta(y|a,b,c^a,c^b) \tag{6}$$

where $p_\theta(c^a,c^b|a,b)$ is the pairwise biological function distribution consisting of $p_\theta(c^a|(a,b))$ and $p_\theta(c^b|(b,a))$. $\mathcal{C}$ is the biological function set. We append a pairwise biological function c^a and c^b to the query input x to obtain the entire input of LM.

The Entire Input of LM

> The first drug is <$\boldsymbol{S}^a$> <$\boldsymbol{T}^a$>, the function is <c^a>. The second drug is <$\boldsymbol{S}^b$> <$\boldsymbol{T}^b$>, the function is <c^b>.

Given that marginalizing all latent pairwise biological functions is resource-intensive, we approximate the Equation 6 by assuming over pairwise top-K potential biological functions with the highest probability under $p_\theta(c^a|(a,b))$ and $p_\theta(c^b|(b,a))$, thus the Equation 6 is reformulated as:

$$p(y|a,b) \approx \sum_{\substack{c^a\in\mathcal{C}^a\\ c^b\in\mathcal{C}^b}} p_\theta(c^a,c^b|a,b)p_\eta(y|a,b,c^a,c^b) \tag{7}$$

where $\mathcal{C}^a \subset \mathcal{C}$ and $\mathcal{C}^b \subset \mathcal{C}$ are the top-K biological functions of drug a and drug b, respectively. We match all possible biological functions from $\mathcal{C}^a$ and $\mathcal{C}^b$ and obtain $K \times K$ pairwise biological functions. One pairwise biological function probability is defined as the joint $p_\theta^*(c^a|(a,b))$ among top-K and $p_\theta^*(c^b|(b,a))$ among top-K:

$$p_\theta(c^a,c^b|a,b) = p_\theta^*(c^a|(a,b))p_\theta^*(c^b|(b,a)) \tag{8}$$

$$p_\theta^*(c^a|(a,b)) = \frac{\exp(r(c^a|(a,b))}{\sum_{c^a\in\mathcal{C}^a}\exp(r(c^a|(a,b))}. \tag{9}$$

Similarly, we can get $p_\theta^*(c^b|(b,a))$. According to this, the $\sum_{\substack{c^a\in\mathcal{C}^a\\ c^b\in\mathcal{C}^b}} p_\theta(c^a,c^b|a,b) = 1$. Inspired by the token generation method (Lewis et al., 2020), the current generated token is not only based on previous $i-1$ tokens but also influenced by $p_\theta(c^a,c^b|a,b)$. Thus, finally $p(y|a,b)$ that generates DDIE with length L is turn as:

$$\prod_{i}^{L} \sum_{\substack{c^a \in \mathcal{C}^a \\ c^b \in \mathcal{C}^b}} p_\theta(c^a, c^b | a, b) p_\eta(y_i | x(a, b), c^a, c^b, y_{<i}) \tag{10}$$

Training. Given that the aim of our method is designed for a professional DDIE generation, we freeze the parameters of the graph encoder and utilize our DDIE data to fine-tune the adapter and LM (leverages the medium-sized Galactica$_{1.3B}$ (Taylor et al., 2022)) based on the pre-trained parameters in (Fang et al., 2024). We optimize the model by minimizing each target's negative marginal log-likelihood in Equation 10.

DDIE Prediction. Taking the query input x and top $K \times K$ pairwise biological functions from PKS, the generator utilizes the beam decoder to generate $K \times K$ DDIE texts. We take the one with highest generation scores exp$(p_\theta(c^a, c^b|a, b)p_\eta(y|a, b, c^a, c^b))$ as the prediction.

4 Experiments

4.1 Experimental Setup

Datasets. To extensively evaluate the predictive ability of models, we constructed two DDIE datasets from two professional DDI databases, MecDDI (Hu et al., 2023) and DDInter2.0 (Tian et al., 2024), respectively. ***MecDDI*** is a database that provides biological functions for all the collected DDIs[3]. We collected all DDIs from the MecDDI database and filtered out drugs lacking SMILES and finally obtained the MecDDI dataset, which contains 1,685 drugs, 1,061 types of biological functions, and 152,922 DDIs belonging to 103 types of DDIE. ***DDInter2.0*** is a comprehensive DDI database [4]. Different from MecDDI having highly-summarized and countable DDIE, the DDIE descriptions in DDInter2.0 are more detailed and hard to summarize to countable classes directly, as shown in Appendix C.1. After our collection, filtration, and ensuring every DDI has the MecDDI-provided biological functions of drugs, the DDInter dataset, in our work, contains 1,683 drugs and 152,887 DDIs.

Baselines. We compare our method with the following two types of baselines:

- ***Classification-Based Methods***: DeepDDI (Ryu et al., 2018) utilizes the structural similarity of drug pairs to predict the DDIE labels. GMPNN-CS (Nyamabo et al., 2022), SSI-DDI (Nyamabo et al., 2021), SA-DDI (Yang et al., 2022), MSAN (Zhu et al., 2022), and DSN-DDI (Li et al., 2023) design different GNN-based encoders to learn representations of 2D molecular structures of drug pairs for DDIE prediction. 3DGT-DDI (He et al., 2022) encodes 3D structure information of drug molecules through a molecular conformation encoder and follows a CNN for DDIE prediction.
- ***Generation-Based Methods***: MolTC (Fang et al., 2024) takes the SMILES and the 2D structure of drug pairs as input and is supervised by molecular descriptions for pretraining. Herein, to compare with our method, we fine-tune all its parameters in its fine-tuning stage with our dataset. MolT5 (Edwards et al., 2022) input the SMILES of a drug molecule, which is modified by adding another drug molecule for tackling our DDIE generation task.

Evaluation Protocols. In this work, we comprehensively measure methods under three DDI scenarios: ***Random Split*** simulating transductive scenarios means we randomly split the samples (drug pairs and corresponding DDIEs) in the dataset to training, validation, and testing sets by 7:1:2. Next are inductive scenarios: ***Cold Start Split*** means we split drugs into seen drugs and unseen drugs in the ratio of 2:1, the drug pairs in the training set only involve seen drugs, the drug pair in the validation set is composed of seen drug and unseen drug, the drug pairs in the testing set only involve unseen drugs. ***Scaffold Split*** is the same as Cold Start Split except for the difference that the seen and unseen drugs are split by molecular scaffold. All results are the means of 3 independent runs. We conduct separate training and evaluation processes for both datasets.

Metrics For generation, we employ BLEU-2, BLEU-4, (Papineni et al., 2002), ROUGE-L (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) for quantitative analysis. For classification, we

[3] https://mecddi.idrblab.net/

[4] https://ddinter2.scbdd.com/

Model	Dataset	Random Split				Cold Split				Scaffold Split			
		B-2	B-4	Metero	R-L	B-2	B-4	Metero	R-L	B-2	B-4	Metero	R-L
MolT5	MecDDI	93.51	92.71	94.09	93.59	62.82	58.05	64.29	61.95	47.78	41.97	50.29	46.45
MolTC		96.02	95.59	96.12	96.02	59.84	55.16	61.51	58.85	44.88	39.33	46.65	42.39
PKAG-DDI		96.75	96.39	96.85	96.72	62.96	58.71	64.34	61.87	48.30	42.89	49.53	45.78
PKAG-DDI*		**99.63**	**99.58**	**99.68**	**99.89**	**99.19**	**99.35**	**99.21**	**99.58**	**99.17**	**99.29**	**99.19**	**99.46**
MolT5	DDInter 2.0	81.30	80.30	86.02	86.38	29.01	24.21	34.38	31.99	10.94	5.84	18.40	15.96
MolTC		83.18	81.71	86.50	85.84	37.34	32.36	40.71	39.47	20.40	14.20	28.31	28.39
PKAG-DDI		92.39	**91.54**	92.78	92.10	44.18	39.42	46.13	43.66	22.85	16.23	30.45	28.66
PKAG-DDI*		**92.43**	91.49	**92.88**	**92.17**	**56.35**	**52.10**	**57.49**	**56.59**	**39.15**	**33.00**	**48.74**	**48.62**

Table 1: Results (in %) of our method with generation-based baselines for DDIE generation on the MecDDI dataset and DDInter2.0 dataset under three different data split settings. The abbreviations are BLEU-2, BLEU-4, and ROUGE-L, respectively. The best and suboptimal results are highlighted in **bold** and underline, respectively.

Model	Minor		Moderate		Major	
	B-2	R-L	B-2	R-L	B-2	R-L
MolTC	75.15	79.09	83.89	86.98	82.39	83.89
PKAG-DDI	**85.86**	**87.08**	**93.66**	**93.12**	**88.80**	**89.61**

Table 2: The quality of generated text in different clinical risk levels.

employ Accuracy and Macro-F1 to measure the results. Note that to qualitatively measure the performance of generation-based methods, we convert them to text classification, i.e., we first vectorize (Harris, 1954) the predicted text and the texts of all DDIE classes, then calculate the cosine similarity between them. The most similar label is the prediction. The implementation details and hyperparameters are shown in Appendix C.2

4.2 Comparison with Baselines

In this section, we evaluate the performance of our method using both textual generation and multi-class classification metrics. To measure the effectiveness of biological function in LM, we propose an upper-bound model PKAG-DDI*, which directly adds the pairwise gold biological function to the input prompt of the generator for the prediction of DDIE.

Evaluate the Generative Capacity. We compare our methods with the generation-based methods on the MecDDI and DDInter2.0 datasets under three data splitting scenarios. The results are shown in Table 1. Except for our PKAG-DDI*, PKAG-DDI achieves optimal performance in almost all evaluation metrics in both datasets, indicating the remarkable and consistent superiority of PKAG-DDI in the generation of DDIE text. Moreover, we have the following observations: (1) Performance of all methods decreases when replacing the experiment data from the MecDDI to the DDInter2.0, highlighting the difficulty of generating more detailed and complex DDIE text on DDInter2.0. In particular, PKAG-DDI shows prominent advantages in DDInter2.0, which indicates that PKAG-DDI exhibits robust capabilities for long and complex DDIE text generation. (2) The superior performance of the upper-bound model PKAG-DDI* underscores the significant contribution of capturing biological functions to DDI text generation. Moreover, PKAG-DDI achieves comparable results with PKAG-DDI* in some scenarios, validating the effectiveness of our pairwise knowledge integration strategy.

To further evaluate the real-world applicability of the generated text, we categorize all DDIs into three clinical risk levels (i.e., Minor, Moderate, and Major). We compare our method with the suboptimal baseline (MolTC) on the Random Split set of DDInter 2.0, and the results are shown in Table 2. The results indicate that in different clinical risk level scenarios, the generated text of our method has a significant advantage. In addition, performance at the Major level is higher than that at the Minor level, indicating our method's high value in clinical major risk assessment.

Evaluate the Classification Capacity. We further assess the quality of our method's generated DDIE text using classification metrics and compare its performance with generation-based methods and classification-based DDIE prediction baselines on the MecDDI dataset. The results are shown in Table 3. Our proposed method, PKAG-DDI, still achieved competitive performances. On the one hand, this confirms that our proposed method does not sacrifice DDIE prediction accuracy for

Model	Model Type	Random Split		Cold Split		Scaffold Split	
		ACC.	F1	ACC.	F1	ACC.	F1
DeepDDI	Classification based	88.68	76.56	37.91	21.03	20.10	4.62
GMPNN-CS		68.89	40.62	26.96	12.50	18.89	**5.58**
SSI-DDI		90.04	80.19	36.00	22.04	16.54	4.24
DSN-DDI		91.03	81.09	39.83	23.63	19.55	4.77
MSAN		93.53	76.52	42.45	22.22	14.30	1.09
SA-DDI		**95.29**	**90.04**	38.32	25.46	16.48	4.11
3DGT-DDI		92.86	82.23	36.96	19.28	19.60	4.38
MolT5	Generation based	90.78	81.81	40.94	24.46	19.46	4.80
MolTC		94.12	87.51	40.41	27.46	18.11	4.80
PKAG-DDI		95.05	87.96	**44.39**	**27.80**	**21.97**	5.54

Table 3: Multi-class classification performance (in %). ACC. is the abbreviation for Accuracy.

Model	Random Split		Cold Split		Scaffold Split	
	A.@2↑	Time↓	A.@2↑	Time↓	A.@2↑	Time↓
PKR w/ BERT	90.38	170.8	46.28	90.3	26.12	95.7
PKR w/ BoW	97.50	476.2	47.06	261.6	**26.63**	255.4
PKS w/o Reuse	98.15	134.1	51.52	71.7	23.88	62.7
PKS	**98.51**	**65.0**	**53.98**	**41.9**	25.80	**35.2**

Table 4: The classification performance of the PKS and the constructed comparison methods on the MecDDI dataset. A.@2 refers to top 2 Accuracy, and the Time refers to the wall clock time of inference.

the sake of DDIE text generation. On the other hand, it demonstrates that DDIE text generation, as an emerging approach to DDIE prediction, holds significant potential and value for practical applications.

4.3 Efficiency Analysis of Pairwise Knowledge Selector (PKS)

In this paper, we simply the pairwise knowledge retriever to the selector as mentioned in Related Work and propose the reuse strategy. Thus, we assess the efficiency of PKS by comparing it with its variants, including dense retrieval variants with Bag of Word encoder and BERT encoder (dubbed PKR w/ BoW and PKR w/ BERT) and the variant without reuse strategy (PKS w/o Reuse). The results are shown in Table 4. PKS demonstrates a significant advantage in computational efficiency while maintaining great general performance. The higher performance of PKS compared with PKS w/o Reuse indicates that our reuse strategy not only effectively shortens the prediction time but also improves the performance. More details of variant models are shown in the Appendix D.1. Additionally, the ablation study of the influence of molecular fingerprints and graphs is shown in Appendix D.2.

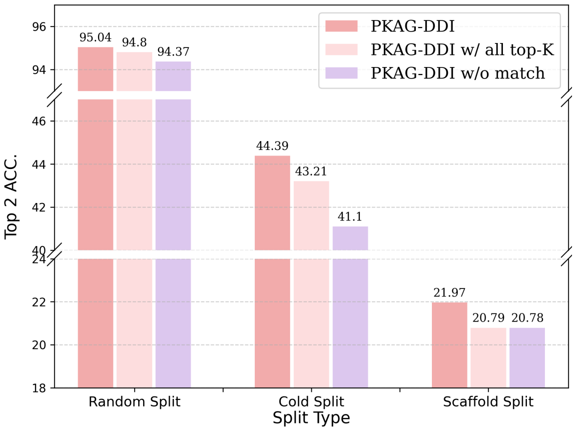


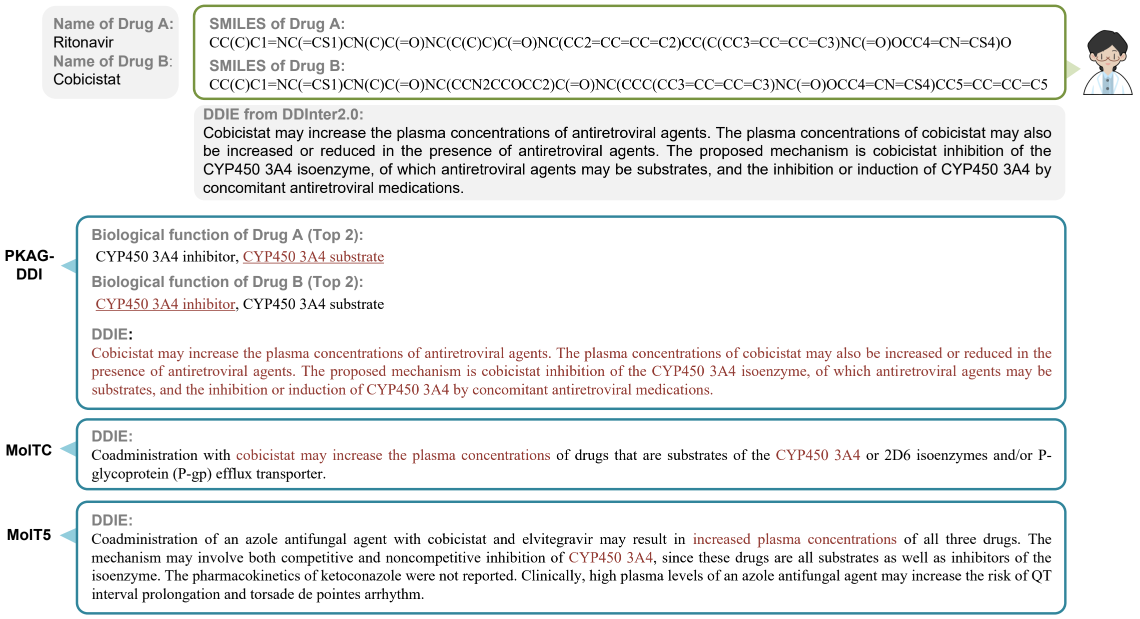
Figure 4: Results of different integration strategies.

4.4 Effectiveness of Pairwise Knowledge Integration Strategy.

In this section, we discuss the effectiveness of our proposed pairwise knowledge integration strategy in DDIE generation based on the selected biological function from PKS. We constructed two variants PKAG-DDI w/ all top-K and PKAG-DDI w/o match. PKAG-DDI w/ all top-K replaces our integration strategy to directly attach all top-$K \times K$ pairwise biological functions to the query input. PKAG-DDI w/o match simply matches pairwise biological functions according to their ranking and without using a joint probability distribution. The results are shown in Figure 4. It can be observed that PKAG-DDI achieves the best results in all three scenarios, which confirms that our integration strategy can effectively integrate the selected biological functions, thereby maximizing the enhancement of the LM's ability to generate DDIE text. The performance of PKAG-DDI w/ all top-K is inferior to PKAG-DDI, indicating that inputting all potential biological functions may bring noise to LM. PKAG-DDI w/o match achieves the worst performances, which demonstrates that simply matching based on rankings could lead to mismatched biological functions thereby introducing noise, and simply training all instances indiscriminately with the same label may confuse the model.

4.5 Case Study

To explicitly show the generation process and the quality of our method, we present a case study in Figure 5. We choose two CYP3A enzyme inhibitions, Ritonavir and Cobicistat, from (Zhong et al., 2024) as the example. When the user inputs the SMILES of the drug pair, our method can explic-



Name of Drug A:
Ritonavir
Name of Drug B:
Cobicistat

SMILES of Drug A:
CC(C)C1=NC(=CS1)CN(C)C(=O)NC(C(C)C)C(=O)NC(CC2=CC=CC=C2)CC(C(CC3=CC=CC=C3)NC(=O)OCC4=CN=CS4)O

SMILES of Drug B:
CC(C)C1=NC(=CS1)CN(C)C(=O)NC(CCN2CCOCC2)C(=O)NC(CCC(CC3=CC=CC=C3)NC(=O)OCC4=CN=CS4)CC5=CC=CC=C5

DDIE from DDInter2.0:
Cobicistat may increase the plasma concentrations of antiretroviral agents. The plasma concentrations of cobicistat may also be increased or reduced in the presence of antiretroviral agents. The proposed mechanism is cobicistat inhibition of the CYP450 3A4 isoenzyme, of which antiretroviral agents may be substrates, and the inhibition or induction of CYP450 3A4 by concomitant antiretroviral medications.

PKAG-DDI

Biological function of Drug A (Top 2):
CYP450 3A4 inhibitor, CYP450 3A4 substrate

Biological function of Drug B (Top 2):
CYP450 3A4 inhibitor, CYP450 3A4 substrate

DDIE:
Cobicistat may increase the plasma concentrations of antiretroviral agents. The plasma concentrations of cobicistat may also be increased or reduced in the presence of antiretroviral agents. The proposed mechanism is cobicistat inhibition of the CYP450 3A4 isoenzyme, of which antiretroviral agents may be substrates, and the inhibition or induction of CYP450 3A4 by concomitant antiretroviral medications.

MolTC

DDIE:
Coadministration with cobicistat may increase the plasma concentrations of drugs that are substrates of the CYP450 3A4 or 2D6 isoenzymes and/or P-glycoprotein (P-gp) efflux transporter.

MolT5

DDIE:
Coadministration of an azole antifungal agent with cobicistat and elvitegravir may result in increased plasma concentrations of all three drugs. The mechanism may involve both competitive and noncompetitive inhibition of CYP450 3A4, since these drugs are all substrates as well as inhibitors of the isoenzyme. The pharmacokinetics of ketoconazole were not reported. Clinically, high plasma levels of an azole antifungal agent may increase the risk of QT interval prolongation and torsade de pointes arrhythm.

Figure 5: The case study on DDInter2.0. Red text denotes content matching the reference labels. The underline indicates the gold biological function provided by MecDDI.

itly interpret the potential biological functions of each drug and further generate the DDIE text. We find that the gold biological functions are in our prediction, and the DDI text we provide is the same with the reference label, which demonstrates the effectiveness and practical application ability of our method. Alongside completely accurate predictions, we also showcase randomly selected examples containing partial prediction mismatches, which are shown in Appendix D.4.

5 Conclusion

In this paper, we emphasized the effectiveness of biological functions in DDIE text generation and introduced a novel pairwise knowledge-augmented generative method for DDIE generation, which can be applied to practical prediction scenarios where knowledge is absent. We also introduce a pairwise biological function sector to efficiently inject mutually conditional drug information and a pairwise knowledge integration strategy for matching and selectively integrating the knowledge to LM. Experiments demonstrate the superiority of our method over baselines. Our method provides a foundation for transitioning from classification to generation in future DDIE predictions.

Limitations

Although our work has researched the generalization and practical application scenarios, such as inductive sets where test drugs are unseen in the training set, and the case where biological functions are absent, it does not address the zero-shot scenarios where new drugs have novel biological functions that not included into the existing biological function set. Because the pairwise knowledge selector we provide depends on the fixed amount of knowledge set, and does not support dynamic updating of datasets. Though the variant of PKS (i.e., PKR w/ BERT and PKR w/ BoW) can address the issue, their accuracy and efficiency also need to be improved. Future endeavors will focus on more practical scenarios.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62372204, 62072206, 62102158); Huazhong Agricultural University Scientific & Technological Self-innovation Foundation; Fundamental Research Funds for the Central Universities (2662024SZ006). The funders have no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation* *measures for machine translation and/or summarization*, pages 65–72.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yujie Chen, Tengfei Ma, Xixi Yang, Jianmin Wang, Bosheng Song, and Xiangxiang Zeng. 2021. MUFFIN: multi-scale feature fusion for drug–drug interaction prediction. *Bioinformatics*.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Junfeng Fang, Shuai Zhang, Chang Wu, Zhengyi Yang, Zhiyuan Liu, Sihang Li, Kun Wang, Wenjie Du, and Xiang Wang. 2024. MolTC: Towards molecular relational modeling in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1943–1958, Bangkok, Thailand. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Robert C Glen, Andreas Bender, Catrin H Arnby, Lars Carlsson, Scott Boyer, and James Smith. 2006. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to adme. *IDrugs*, 9(3):199.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

ZS Harris. 1954. Distributional structure.

Haohuai He, Guanxing Chen, and Calvin Yu-Chian Chen. 2022. 3dgt-ddi: 3d graph and text based neural network for drug–drug interaction prediction. *Briefings in Bioinformatics*, 23(3):bbac134.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Wei Hu, Wei Zhang, Ying Zhou, Yongchao Luo, Xiuna Sun, Huimin Xu, Shuiyang Shi, Teng Li, Yichao Xu, Qianqian Yang, and others. 2023. MecDDI: Clarified Drug–Drug Interaction Mechanism Facilitating Rational Drug Use and Potential Drug–Drug Interaction Prediction. *Journal of Chemical Information and Modeling*, 63(5):1626–1636. Publisher: ACS Publications.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2024. Atlas: few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24(1). Publisher: JMLR.org.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zimeng Li, Shichao Zhu, Bin Shao, Xiangxiang Zeng, Tong Wang, and Tie-Yan Liu. 2023. Dsn-ddi: an accurate and generalized framework for drug–drug interaction prediction by dual-view representation learning. *Briefings in Bioinformatics*, 24(1):bbac597.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023a. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36:22820–22840.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. 2023b. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Arnold K Nyamabo, Hui Yu, Zun Liu, and Jian-Yu Shi. 2022. Drug–drug interaction prediction with learnable size-adaptive molecular substructures. *Briefings in Bioinformatics*, 23(1):bbab441.

Arnold K Nyamabo, Hui Yu, and Jian-Yu Shi. 2021. SSI–DDI: substructure–substructure interactions for drug–drug interaction prediction. *Briefings in Bioinformatics*. _eprint: https://academic.oup.com/bib/advance-article-pdf/doi/10.1093/bib/bbab133/37809433/bbab133.pdf.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. 2018. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the national academy of sciences*, 115(18):E4304–E4311.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8364–8377.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34:25968–25981.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Yao Tian, Jiacai Yi, Ningning Wang, Chengkun Wu, Jinfu Peng, Shao Liu, Guoping Yang, and Dongsheng Cao. 2024. DDInter 2.0: an enhanced drug interaction resource with expanded data coverage, new interaction types, and improved user interface. *Nucleic Acids Research*, page gkae726.

Yaqing Wang, Zaifei Yang, and Quanming Yao. 2024a. Accurate and interpretable drug-drug interaction prediction enabled by knowledge subgraph learning. *Communications Medicine*, 4(1):59.

Ziyan Wang, Zhankun Xiong, Feng Huang, Xuan Liu, and Wen Zhang. 2024b. Zeroddi: a zero-shot drug-drug interaction event prediction method with semantic enhanced learning and dual-modal uniform alignment. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 6071–6079.

Zhankun Xiong, Shichao Liu, Feng Huang, Ziyan Wang, Xuan Liu, Zhongfei Zhang, and Wen Zhang. 2023. Multi-relational contrastive learning graph neural network for drug-drug interaction event prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5339–5347.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. RECOMP: Improving Retrieval-Augmented LMs with Context Compression and Selective Augmentation. In *The Twelfth International Conference on Learning Representations*.

Ziduo Yang, Weihe Zhong, Qiujie Lv, and Calvin Yu-Chian Chen. 2022. Learning size-adaptive molecular substructures for explainable drug–drug interaction prediction by substructure-aware graph neural network. *Chemical science*, 13(29):8693–8703. Publisher: Royal Society of Chemistry.

Yi Zhong, Gaozheng Li, Ji Yang, Houbing Zheng, Yongqiang Yu, Jiheng Zhang, Heng Luo, Biao Wang, and Zuquan Weng. 2024. Learning motif-based graphs for drug–drug interaction prediction via local–global self-attention. *Nature Machine Intelligence*, pages 1–12.

Xinyu Zhu, Yongliang Shen, and Weiming Lu. 2022. Molecular substructure-aware network for drug-drug interaction prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4757–4761.

A Preliminary Experiment

To evaluate the effectiveness of the biological function in the DDIE prediction task, we experiment two basic DDIE classification models (**DDI-Finger** and **DDI-Graph**) and their variants that added the biological function using MecDDI dataset. These two models utilize the popular feature of drug molecules (i.e., molecular 2D graphs and fingerprints [5]) that SMILES can easily convert. Specifically, for DDI-Finger, we employ the RDKit tool to convert SMILES representations into Extended Connectivity Fingerprints (ECFP) with 1024 dimensions for drug molecules. The fingerprints of drug pairs are then concatenated and fed into a 3-layer Multi-Layer Perceptron (MLP) classifier. For DDI-Graph, we utilize the RDKit tool to convert SMILES into 2D molecular graphs. Subsequently,

[5] A widely utilized binary bit vector representing molecular substructures in drug discovery.

we apply a 3-layer Graph Isomorphism Network (GIN) to obtain the representations of drug pairs, which are concatenated and then input into the 3-layer MLP for classification. Moreover, the variants of the two models are adding different kinds of biological functions. **DDI-Finger w/ all inherent bio** refers to the DDI-Finger incorporating the representations of drug pairs concatenated with the corresponding one-hot vector of all inherent biological functions of drugs. **DDI-Finger w/ most relevant bio** means that the model injects the most DDIE-relevant biological function. **DDI-Graph w/ all inherent bio** and **Graph w/ most relevant bio** are the same.

As Figure 2 shows, the accuracy score of both DDI-Finger and DDI-Graph improves dramatically when injecting the biological function information, suggesting that we can preliminarily conclude that biological functions have the potential to enhance DDI prediction performance and hold value for further research. In addition, we find that the accuracy of DDI-Finger w/ all inherent bio and DDI-Graph w/ all inherent bio is lower than that of DDI-Finger w/ most relevant bio and DDI-Graph w/ most relevant bio, respectively, indicating that the irrelevant biological function in the current DDI may bring noise and hamper the correct prediction. According to this, in this work, we are dedicated to predicting the most DDIE-relevant biological function of drugs in different DDI for better DDIE generation.

B The Discussion of Choosing Gold Biological Function for PKS Training

Herein, we discuss the reasons that we chose the drug's gold biological function for multi-class classification prediction in PKS rather than multi-label classification, and the reason that we use BM25 to pick the most similar biological function to DDIE as the gold one. (1) In the MecDDI dataset, only 19% of drug pairs have multiple biological functions, and most of these functions are highly similar (e.g., "hyperglycemia" and "hyperglycemic effects"). Adopting multi-label classification solely to accommodate these rare cases could degrade overall prediction accuracy. Since errors in the first stage (biological function prediction) may propagate and amplify in the second stage (DDI generation), we prioritize single-label classification (i.e., choosing a gold label) to maximize prediction accuracy. (2) Given the inherent context sensitivity of autoregressive language models, the more token-similar the input and output text, the stronger the guidance of the input text for the label prediction, thereby improving the accuracy of prediction. Accordingly, for instance, when evaluating a drug's biological functions ("Antihypertensive agent" versus "Hypotensive effects") against its DDIE description ("Pharmacodynamic additive effects (Additive **hypotensive effects**)"), BM25 scoring demonstrates higher similarity for "**Hypotensive effects**" due to its direct lexical correspondence. Thus, although the "Antihypertensive agent" also shows the semantic relevance with the DDIE, according to the BM25 score, we chose the "Hypotensive effects" as the gold biological function.

C Experiment Set

C.1 Dataset

DDIE Text Descriptions. Considering the increased demand for distinct and high-quality DDI mechanism descriptions, our work focuses on generating more detailed pharmacokinetic and pharmacodynamic event descriptions for DDI. Therefore, we use the MecDDI database, which provides drug biological function information, and the DDInter2.0 database, which offers detailed DDI descriptions, for our experimental analysis. Several examples of DDIE text descriptions are shown in Table 5. Compared with general DDI databases, such as DrugBank, the DDIE from databases MecDDI and DDInter2.0 is more specific. Moreover, the examples illustrate that MecDDI shares more keywords with DDInter2.0 in DDI event descriptions than DrugBank does, which demonstrates that the biological functions from MecDDI may also be applicable to the DDIE prediction in DDInter2.0.

Task Setting In constructing the MecDDI dataset, we find that 99.71% of drug pairs have only a single DDIE. Consequently, we formulate the DDIE classification in the MecDDI dataset as a multi-class classification task rather than a multi-label classification task.

C.2 Model Configuration

PKAG-DDI consists of two stages: biological function selecting and language model learning, which have significant differences in model size. Thus, we used different devices to train them, separately. For the first stage, we developed our model on the machine with a 15 vCPU Intel(R) Xeon(R) Platinum 8362 CPU @ 2.80GHz (CPU) and an

Drug Pair	Mirtazapine & Ivosidenib
MecDDI	Pharmacodynamic additive effects (Increased risk of prolong QT interval).
DDInter2.0	Ivosidenib can cause prolongation of the QT interval. Theoretically, coadministration with other agents that can prolong the QT interval may result in additive effects and increased risk of ventricular arrhythmias including torsade de pointes and sudden death.
DrugBank	The metabolism of Mirtazapine can be increased when combined with Ivosidenib.
Drug Pair	Tenecteplase & Treprostinil
MecDDI	Pharmacodynamic additive effects (Increased risk of bleeding).
DDInter2.0	Drugs that inhibit platelet function may increase the risk of bleeding when administered prior to, during, or after thrombolytic therapy.
DrugBank	The risk or severity of adverse effects can be increased when Tenecteplase is combined with Treprostinil.
Drug Pair	Tizanidine & Opicapone
MecDDI	Pharmacodynamic additive effects (Additive CNS depression effects).
DDInter2.0	The sedative effect of tizanidine may be potentiated by concomitant use of other agents with central nervous system (CNS) depressant effects. In addition, tizanidine and many of these agents (e.g., alcohol, anxiolytics, sedatives, hypnotics, antidepressants, antipsychotics, opioids, muscle relaxants) also can exhibit hypotensive effects, which may be additive during coadministration and may increase the risk of symptomatic hypotension and orthostasis, particularly during initiation of therapy or dose escalation. Tizanidine itself is a central alpha-2 adrenergic agonist. Pharmacologic studies have found tizanidine to possess between 1/10 to 1/50 of the potency of clonidine, a structurally similar agent, in lowering blood pressure.
DrugBank	The risk or severity of adverse effects can be increased when Tizanidine is combined with Opicapone.

Table 5: Examples of the ground-truth DDIE provided by MecDDI, DDInter2.0 and DrugBank, respectively.

NVIDIA GeForce RTX 3090 (GPU). For the second stage, we developed our model on a machine with two A800s with 80GB of video memory. Our model is implemented with PyTorch (2.1.0+cu121), PyTorch-geometric (2.6.1), RDkit (2024.03.5), and pytorch_lightning (1.9.0). The size of Galactica is 1.3 B. The pre-trained model parameter is the "stage2/last.ckpt" from MolTC.

C.3 Training Strategy

In stage two, we employed the AdamW optimizer with a weight decay of 0.05 and a learning rate of 0.001. We implemented an early stopping strategy in the training process to conserve computational resources with "patience" in 5 epochs, and "min_delta" of training loss is 0.0002. To fair comparison, the MolTC and MolT5 use the same training strategy with our model. Moreover, we fine-tune our model using Low-Rank Adaptation (LoRA) (Hu et al., 2021), which is one of the parameter-efficient fine-tuning (PEFT) technologies.

C.4 Hyper-parameters

The primary hyper-parameters, such as the learning rate, weight decay, dropout rate, and the parameter λ that controls the information flow from another drug, etc., are searched by using hyper-parameter tuning technology Optuna. The hyper-parameters in PKAG-DDI are represented in the code. Although the biological function pairs grow quadratically with K in theory, in practical applications, most of the biological functions of a drug (around 97% in the current dataset) are less than three. That is, the larger the K, the more noise it will bring and may hinder the DDIE prediction. Thus, we suggest using $K = 2$.

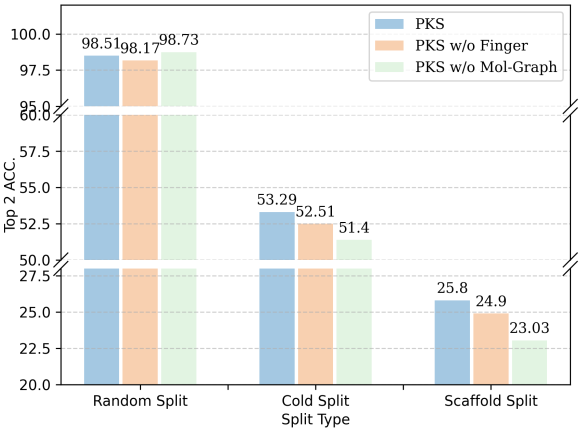


Figure 6: The ablation study about the information of Fingerprint and Molecular Graph.

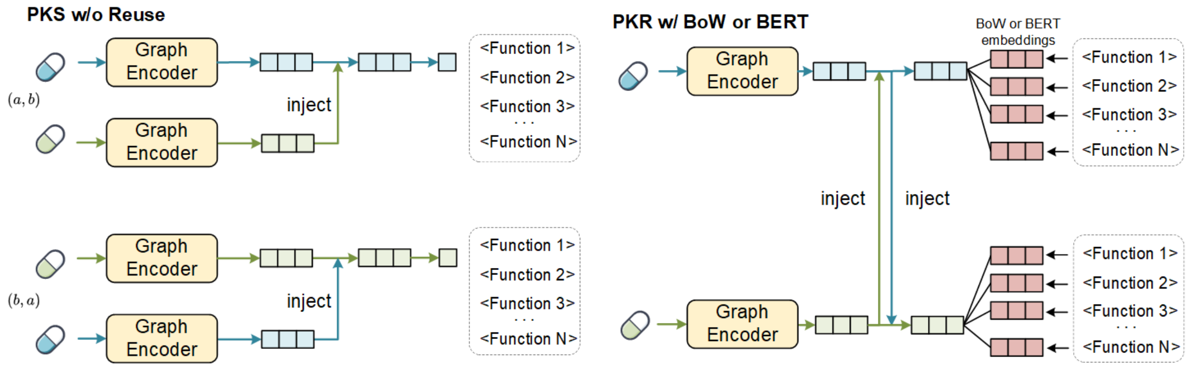


Figure 7: The variants of PKS.

SMILES	Molecular Graph	Random Split ACC.	Cold Split ACC.	Scaffold Split ACC.
✓	×	94.94	43.95	21.55
×	✓	94.95	43.86	19.98
✓	✓	**95.05**	**44.39**	**21.97**

Table 6: The ablation study about evaluating the impact of SMILES and 2D molecular graph in our method.

D Experiments

D.1 Supplement of Efficiency Analysis of PKS

We conduct two dense retrieval variants of PKS by removing the classifier by a knowledge encoder and dense inner product model (dubbed pairwise knowledge retriever, PKR). Considering that knowledge (i.e., biological functions) are phrases with an average length of approximately four, we use a common word encoder technology Bag of Word (BoW), and a semantic encoder technology BERT (with the pretrained weight *scibert_scivocab_uncased*). Moreover, we also conduct the variant of PKS without the reuse strategy (dubbed PKS w/o Reuse). That is training the instance (a, b) and (b, a) separately. The models' architectures are illustrated in the Figure 7. The results shown in Figure 4 still demonstrate that PKS still outperforms the PKR w/ BoW and PKR w/ BERT in Random Split and Cold Split, demonstrating that directly predicting the label of biological function is enough for most biological function selection scenarios. Additionally, the superior performance of the pairwise knowledge retriever in the Scaffold Split set indicates that dense retrieval is better suited for scenarios where the test instance distribution differs from the training instance distribution.

D.2 Ablation Study of PKS

To evaluate the influence of molecular fingerprints and graphs on PKS. We construct two variants of PKS, PKS w/o Finger and PKS w/o Mol-Graph by removing molecular fingerprints and graphs, respectively, and compare them with our PKS in three data split scenarios. The results are illustrated in Figure 6. PKS fusing fingerprints and graphs simultaneously achieves the best performance in most scenarios, indicating they help the model comprehensively learn drug molecules and boost the accuracy of biological function selection in more challenging cold and scaffold scenarios. The slightly inferior performance of PKS compared to PKS w/o Mol-Graph in the random split scenario may stem from the fact that when all drugs are seen and samples are abundant, using two types of molecular structural data simultaneously could slightly constrain model optimization.

D.3 Ablation Study of the Multi-Modality of LM

In Section 3.3, we use two modalities (that is, the SMILES and the 2D molecular graph) to represent the feature of drugs for LM. Herein, we conduct an ablation study to evaluate the impact of these two modalities on our method. The results shown in Table 6 indicate that the use of both the SMILES information and the molecular graph is beneficial for the generation of DDIE.

D.4 Supplement of Case Study

Given PKAG-DDI's superior predictive accuracy (e.g., achieving a 92.39% BLEU-2), we randomly selected a set of not fully correct test samples from DDInter2.0 to show our method's predictions. As depicted in Figure 8, we present the Top-2 predictions for biological functions and the predictions for DDIE, along with their corresponding labels. Taking the first DDI sample as an example, PKAG-DDI predicts Glycerol phenylbutyrate

Drug Pair Name	Biological Functions				DDIE	
	Prediction of Drug1	Gold Label of Drug1	Prediction of Drug2	Gold Label of Drug2	Prediction	Label
Ethinylestradiol & Glycerol phenylbutyrate	CYP450 3A4 substrate, CYP450 2D6 substrate	CYP450 3A4 substrate	Hyperammonemia, CYP450 2D6 inhibitor	CYP450 3A4 inducer	Coadministration of glycerol phenylbutyrate may lead to decreased concentrations of CYP450 3A4 substrates. The proposed mechanism is inhibition of CYP450 3A4 by phenylbutyrate, the active moiety of glycerol phenylbutyrate, which has been shown to be an in vitro inhibitor of this isoenzyme.	Coadministration with glycerol phenylbutyrate may decrease the systemic exposure to and therapeutic efficacy of drugs that are substrates of CYP450 3A4. Glycerol phenylbutyrate, its active moiety phenylbutyrate, and active metabolite phenylacetic acid are considered weak CYP450 3A4 inducers in vivo.
Lomitapide & Crizotinib	CYP450 3A4 substrate, Hepatotoxicity	CYP450 3A4 substrate	CYP450 3A4 inhibitor, Hepatotoxicity	CYP450 3A4 inhibitor	Coadministration with inhibitors of CYP450 3A4 may significantly increase the plasma concentrations of lomitapide, which is primarily metabolized by the isoenzyme.	Coadministration with potent and moderate inhibitors of CYP450 3A4 may significantly increase the plasma concentrations of lomitapide, which is primarily metabolized by the isoenzyme.
Diiodohydroxyquinoline & Iodide I-123	Interfere with the therapeutic effect and/or diagnostic result of radioiodides, Peripheral neuropathy	Interfere with the therapeutic effect and/or diagnostic result of radioiodides	Radioiodide, Anion exchange resin/polymer	Radioiodide	INTERVAL: Certain drugs may affect the pharmacokinetics, pharmacodynamics, and/or diagnostic results of radioiodides.	INTERVAL: Iodine-containing products may affect the pharmacokinetics, pharmacodynamics, and/or diagnostic results of radioiodides. These effects may last up to 8 weeks.
Levothyroxine & Testolactone	Thyroid hormone, Binds to cation exchange resin/polymer	Thyroid hormone	Increase the risk of hyperthyroidism in patients receiving thyroid hormone replacement therapy, Increase serum thyroid-binding globulin concentration	Increase the risk of hyperthyroidism in patients receiving thyroid hormone replacement therapy	Theoretically, exemestane may potentiate the risk of hyperthyroidism in patients receiving thyroid hormone replacement therapy. Exemestane has been reported to induce hyperthyroidism in patients receiving thyroid hormone therapy.	Androgens may induce reversible clinical hyperthyroidism in patients receiving thyroid hormone replacement therapy. The proposed mechanism is androgen-induced decrease in T4 binding globulin resulting in decreased serum T4, increased T3 uptake resin and free T4, and decreased TSH levels.

Figure 8: The examples of predictions. The red text indicates matches, while the blue text indicates mismatches.

as an inhibitor, consequently leading to the DDIE prediction identifying the drug as an inhibitor as well, which is the major difference between this DDIE prediction and its label. This indicates that biological functions have a strong propensity for guiding the generation of DDIEs. Moreover, when the biological functions generated by our method are correct, the resulting DDIE closely matches the labels, such as the second and third examples in Figure 8.