ExDDI: Explaining Drug-Drug Interaction Predictions with Natural Language

Zhaoyue Sun[1], **Jiazheng Li**[2], **Gabriele Pergola**[1], **Yulan He**[1,2,3]

[1]Department of Computer Science, University of Warwick
[2]Department of Informatics, King's College London
[3]The Alan Turing Institute
Zhaoyue.Sun@warwick.ac.uk, Jiazheng.Li@kcl.ac.uk,
Gabriele.Pergola.1@warwick.ac.uk, Yulan.He@kcl.ac.uk

Abstract

Predicting unknown drug-drug interactions (DDIs) is crucial for improving medication safety. Previous efforts in DDI prediction have typically focused on binary classification or predicting DDI categories, with the absence of explanatory insights that could enhance trust in these predictions. In this work, we propose to generate natural language explanations for DDI predictions, enabling the model to reveal the underlying pharmacodynamics and pharmacokinetics mechanisms simultaneously as making the prediction. To do this, we have collected DDI explanations from DDInter and DrugBank and developed various models for extensive experiments and analysis. Our models can provide accurate explanations for unknown DDIs between known drugs. This paper contributes new tools to the field of DDI prediction and lays a solid foundation for further research on generating explanations for DDI predictions.

Code and Data — https://github.com/ZhaoyueSun/ExDDI

Introduction

Drug-drug interaction (DDI) refers to the alteration of the effects of one or more drugs when drugs are taken simultaneously (Zhang et al. 2023). Such changes may lead to loss of therapeutic effect or occurrence of toxicity, threatening patient safety (Zhang et al. 2023). With the increasing number of approved drugs in recent years, the likelihood of interactions between drugs has also increased (Khori, Semnani, and Roshandel 2011; Han et al. 2022). Although wet lab experiments are available for validating DDIs, they are hindered by strict experimental conditions and high costs (Safdari et al. 2016), making it unfeasible to explore all potential interaction combinations. Therefore, computational methods for predicting DDIs have been extensively researched, and numerous models have demonstrated strong predictive capabilities. However, as predictive capabilities advance, models tend to become more complex and opaque, obstructing users' understanding of the predicted results (Vo et al. 2022).

Specifically, the majority of previous methods have focused only on binary classification, i.e., predicting whether there is an interaction between two drugs (Figure 1(b)), yet overlooking the mechanisms and outcomes of DDIs (Zhang et al. 2023). To help users better grasp DDI knowledge from predictions, some studies have proposed DDI-type prediction (Ryu, Kim, and Lee 2018; Deng et al. 2020; Lin et al. 2022), which is defined as a multi-classification problem that categorises DDIs into various subtypes according to their effects. For example, Deng et al. (2020) used NLP techniques to extract quadruples (`drugA, drugB, mechanism, action`) from DDI descriptions collected from DrugBank, where '`mechanism`' refers to the drugs' effects on metabolism, serum concentration, therapeutic efficacy, etc., and '`action`' indicates an increase or decrease. They summarised DDIs into 65 types based on the extracted quadruples used for classification (Figure 1(c)).

While DDI-type prediction reveals the outcomes of DDI events, the granularity is coarse, lacking attention to the underlying causes of DDIs. As a valuable resource, Xiong et al. (2022) constructed the DDInter database, gathering information on 1.8k approved drugs and 0.24M associated DDIs, along with detailed explanations. The explanations were collected from scientific literature in PubMed and medication guides of drugs, and reviewed by a clinical pharmacist team.

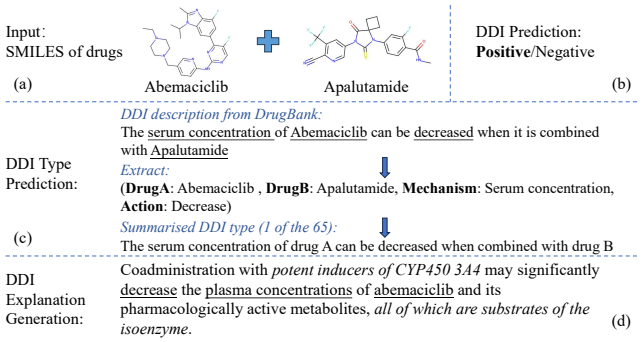

Figure 1: Examples of different DDI prediction tasks. (a) Model inputs, i.e., SMILES representations of the drug pairs; (b) Traditional DDI prediction: binary classification task; (c) DDI-type prediction: multiple classification task; (d) DDI explanation generation: our proposed task, formulated as text generation. The underlined content represents the annotations involved in DDI type prediction, while the italicized text denotes unique content provided by DDInter's explanations.

Compared to DDI types defined by previous research, the DDI explanations provided in DDInter are more informative, encompassing not only the consequences of DDIs but also the cause mechanisms contributing to their occurrence, as illustrated in Figure 1(d).

In this work, we propose a novel task of generating natural language explanations for DDI predictions. **Our goals** are : 1) to explore methods that generate explanations of the underlying pharmacodynamic or pharmacokinetic mechanisms when predicting DDIs. These explanations could help researchers evaluate the plausibility of the model's predictions based on their expertise; 2) to investigate how the explanation generation process influences the prediction task.

For this task, we collected long and short explanations of DDIs from DDInter and DrugBank, respectively. We conducted extensive experiments in both *transductive* and *inductive* settings to meet the needs of application scenarios. We propose and evaluate the performance of the **ExDDI** family methods for DDI explanation generation, which includes *three different fine-tuning paradigms*—namely, seq-to-seq, multi-task training, and multi-task training with staged inference —along with a *retrieval-based unsupervised model* and an *LLM-based (i.e., ChatGPT) in-context demonstration prompting model*.

Our contributions are:

- To the best of our knowledge, ***we are the first to explore the DDI explanation generation task***, which is crucial for trustworthy AI-driven drug safety research. We created the ExDDI model family for this task and carried out a comprehensive evaluation, offering tools and baselines for future studies.
- Our experiments reveal that ***top-performed fine-tuning methods can effectively capture molecular similarities and generate accurate explanations in the transductive setting***. However, their ability to generalise to unseen drugs during training is limited, likely due to the constraints of the linearised representation of SMILES. Fingerprint similarity-based retrieval methods can match the performance of fine-tuning approaches when both query drugs are unseen, even though they perform less effectively in settings that require less generalisation. On the other hand, general LLMs exhibit very limited capability in DDI prediction when given molecular representations of drugs.
- Additionally, we demonstrate that models trained on DDInter outperform those trained on DrugBank in prediction tasks, suggesting that ***rich, detailed explanations not only enhance human understanding but also improve model prediction capabilities***. Our experimental analysis provides valuable insights for advancing future DDI-related research.

Related Work

DDI Prediction and Interpretability Many efforts have been dedicated to DDI prediction over the years. Some of them are based on similarity measurements, which are grounded on the assumption that similar drugs may possess similar biological activity. Various similarity matrices - targeting molecule structure, side effect, protein targets, etc. - can be used for direct matching (Vilar et al. 2012; Ferdousi, Safdari, and Omidi 2017) or as features to train machine learning classifiers (Gottlieb et al. 2012; Cheng and Zhao 2014; Sridhar, Fakhraei, and Getoor 2016) and neural networks (Rohani and Eslahchi 2019; Lee, Park, and Ahn 2019; Zhang, Lu, and Zang 2022). Other approaches involve matrix decomposition of known DDI matrices combined with multiple relation matrices to predict unknown DDIs (Zhang et al. 2018; Rohani, Eslahchi, and Katanforoush 2020). Additionally, recent advancements have incorporated knowledge graphs (Asada, Miwa, and Sasaki 2023; Ren et al. 2022) and graph neural networks for learning single or paired molecular structures (Baitai et al. 2023; Li et al. 2023; Nyamabo et al. 2022) to enhance prediction accuracy.

In recent years, the transparency of DDI prediction models has gained significant attention. Some studies have employed matrix factorization (Zhu et al. 2022) or attention mechanisms (Ma and Lei 2023; Li et al. 2023) to identify representative features or substructures in DDI interactions, offering valuable insights into the underlying prediction mechanisms. However, generating natural language explanations that focus on elucidating the pharmacological principles of DDIs offers another promising direction for further exploration. Additionally, exploring whether introducing supervision signals from these explanations could enhance the prediction task itself is an intriguing question. Furthermore, natural language explanations are more user-friendly for human understanding and could be integrated with substructure-highlighting methods in future work.

Natural Language Explanation Generation Natural language explanation generation aims to create free-text explanations for model predictions to help users better understand model behaviour and make decisions. Previous work had explored various training paradigms over prediction and explanation generation, which were categorised into four types by Hase et al. (2020) based on whether the model is provided with labels (**RA**) or not (**RE**) during the generation of explanations and whether the generated explanations are used as part of the input for predicting (**ST**) or not (**MT**). Specifically, the **ST-RE** paradigm in the first stage trains the model to generate explanations based on the input text and, in the second stage, learns to predict labels based on explanations generated in the first stage (Rajani et al. 2019). The **ST-RA** paradigm first learns to produce explanations based on labels and input, then generates an explanation for each label in the second stage and trains the model to make predictions based on all explanations (Hase et al. 2020). The **MT-RE** paradigm refers to jointly training the model to generate both labels and explanations simultaneously (Narang et al. 2020; Yordanov et al. 2022). The **MT-RA** paradigm not only jointly trains the prediction and explanation generation, but also provides labels during the explanation generation process (Camburu et al. 2018). This is achieved by feeding the gold label to train the explanation generation model and using the label predicted by the model for inference. For the DDI explanation generation task, explanations for negative

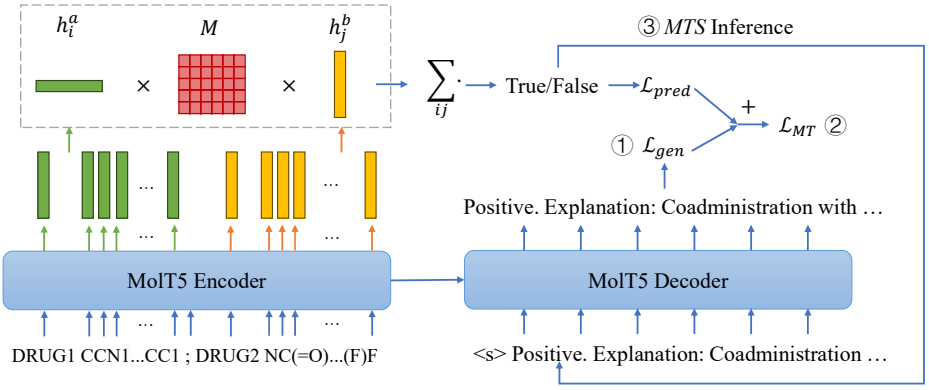

Figure 2: Illustration of the fine-tuning methods. ① the learning objective of the ExDDI-S2S model; ② the learning objective of the ExDDI-MT model; ③ the inference step of the ExDDI-MTS model.

cases are naturally absent. We use artificially constructed explanations for negative cases to train the model, but the relationship between such explanations and predictions is almost lexically distinguishable. Therefore, the ST paradigm is relatively less meaningful for our task. Our finetuning methods explore variants of MT-RE and MT-RA paradigms.

In recent years, inspired by the astonishing reasoning capabilities of LLMs, researchers have also explored generating explanations by prompting LLMs with different strategies (Wei et al. 2022; Lampinen et al. 2022; Wiegreffe et al. 2022). However, in our interactions with general LLMs, such as ChatGPT (OpenAI 2022), we find that while they can make reasonably sound judgments and explanations about whether known drugs have a DDI (Al-Ashwal et al. 2023), they often express incapability when asked with chemical molecular structures (e.g., with SMILES representation). Therefore, to compare the LLM performance with other methods proposed in this work, we prompt LLMs with in-context demonstrations of several similar drug pairs.

Method

Task Formulation Given a drug pair (d_1, d_2), one of our objectives is to predict DDI label $l \in \{\text{“}positive\text{”}, \text{“}negative\text{”}\}$, denoting the presence or absence of interactions between these drugs when administered together. Additionally, we aim to generate a textual explanation, s, elucidating the rationale behind the existence or non-existence of DDIs. For positive instances, we rely on DDI descriptions sourced from DDInter (Xiong et al. 2022) or DrugBank as the target explanation, while for negative instances lacking natural language explanations, we formulate target explanations using a predefined template: s = ‘<DRUG1_DEF>. <DRUG2_DEF>. There were no known direct interactions reported between them.’, where <DRUG1_DEF> and <DRUG2_DEF> represent the drug descriptions retrieved from the DDInter database for d_1 and d_2 respectively.

We explored three setups: *fine-tuning methods* with different paradigms, *retrieval-based* methods, and *LLM-based in-context demonstration prompting* methods on this task.

Fine-tuning Methods

For the fine-tuning methods, we constructed a Seq-to-Seq model (**ExDDI-S2S**), a Multi-Task training model with an additional classifier (**ExDDI-MT**) and a Multi-Task training model with Staged generation constrained by the classifier's prediction (**ExDDI-MTS**). We use MolT5 (Edwards et al. 2022) as the backbone encoder-decoder for these models as it has been pre-trained on molecule-text translation tasks that establish a connection that maps molecular features and natural language representations into a shared space, thereby enhancing the model's generalisability. Figure 2 shows the overall structure of the fine-tuning methods.

ExDDI-S2S For each query drug pair (d_1, d_2), we construct the model's input $\mathbf{x}$ as ‘DRUG1 <SMILES1>; DRUG2 <SMILES2>’, where <SMILES1> and <SMILES2> correspond to the SMILES representations of d_1 and d_2, respectively. For the target output, we first replace mentions of drug names in the target explanations with ‘DRUG1’ and ‘DRUG2’ through regularised expression matching, and then construct the generation target sequence $\mathbf{y}$ = ‘<s> <LABEL> Explanation: <EXP> </s>’, where <LABEL> represents ‘positive’ or ‘negative’ and <EXP> is the preprocessed explanation text. Then the model is trained by the following text generation loss:

$$\mathcal{L}_{gen} = -\sum_{i=1}^{N}\sum_{t=1}^{T} \log p(y_t^{(i)} | \mathbf{x}^{(i)}, \mathbf{y}_{<t}^{(i)}, \Theta), \qquad (1)$$

where N represents the size of the training set, T denotes the length of the target sequence, and Θ signifies the parameters of the encoder and decoder.

ExDDI-MT Simultaneously generating prediction labels during the target sequence generation process could potentially divert the model's attention from learning the classification task effectively. Hence, we attempt to introduce an extra classification module for multi-task training. The design of the classification module is inspired by Nyamabo et al.

(2022), where a linear transformation matrix $\mathbf{M}$ is learned to map the representations corresponding to d_1 and d_2 pairwise to a real-valued score, and the scores obtained for all pairs of representations are summed to make the prediction.

Specifically, suppose the encoder representations for the input corresponding to d_1 and d_2 are denoted as $\mathbf{H}_a = \{\mathbf{h}_1^a, \mathbf{h}_2^a, ..., \mathbf{h}_i^a ...\}$ and $\mathbf{H}_b = \{\mathbf{h}_1^b, \mathbf{h}_2^b, ..., \mathbf{h}_j^b ...\}$, respectively. Our objective is to learn the weights of $\mathbf{M}$ to predict the DDI label, with $\mathbf{M}$ having a dimension of (768×768) in our implementation:

$$\hat{l} = \text{Sigmoid}(\sum_{ij} \mathbf{h}_i^a \mathbf{M} (\mathbf{h}_j^b)^\top). \quad (2)$$

The classifier is then optimised by the binary cross-entropy loss, which is defined as:

$$\mathcal{L}_{pred} = -\frac{1}{N} \sum_{i=1}^{N} [l_i \log(\hat{l}_i) + (1 - l_i) \log(1 - \hat{l}_i)]. \quad (3)$$

The ExDDI-MT paradigm then jointly optimises the generation loss and classifier prediction loss. Thus, the overall loss function is:

$$\mathcal{L}_{MT} = \mathcal{L}_{gen} + \mathcal{L}_{pred} \quad (4)$$

ExDDI-MTS Inspired by Camburu et al. (2018), we are interested in exploring whether using the predictions of the multi-task trained classifier, which may have better classification performance than the decoder, as an additional constraint for decoding can improve the quality of explanation generation.

As shown in Figure 2, during the inference stage, we first utilise the fine-tuned encoder and classifier weights M from ExDDI-MT to predict $\hat{l}$. If $\hat{l}$ is 1, then the decoding prefix Pr is set to "`<s> positive`"; otherwise, it is set to "`<s> negative`". The prefix is prepended during decoding before generating the explanation, specifically:

$$\mathbf{y} = \text{Decoder}(Pr; \mathbf{x}; \Theta) \quad (5)$$

Retrieval-based Method

For the retrieval-based method (**ExDDI-RV**), we retrieve the most similar drug pair in the training set to the query drugs, then use the DDI label and explanation of the retrieved case as the response to the query. This is based on the assumption that similar drugs often share similar pharmacological properties, and indeed, in the data we have collected, a large number of DDIs share the same explanation.

We retrieve the nearest drug pairs based on the similarity of the drugs' chemical molecular structures. Initially, we retrieve the top-K ($K = 50$) most similar drugs for each query drug from the training set's drug list. The similarity score between two drugs is calculated by the Tanimoto coefficient (also known as Jaccard similarity) of their fingerprints, which are binary vectors indicating the presence or absence of specific chemical substructures. We employ RDKit to extract MACCS keys (Durant et al. 2002) as the fingerprints. Subsequently, we pair the top-K nearest neighbours of the two drugs, resulting in K^2 candidate drug pairs. Each candidate drug pair's similarity score is computed as the product of the similarity scores for each retrieved drug and its corresponding query drug. Following this, we filter out drug pairs that do not exist in the training set and re-rank the remaining pairs. Ultimately, the top drug pair is obtained as the retrieval result, and its label and explanation are used as the response.

LLM-based In-Context Prompting

To assess the capabilities of general LLMs in predicting and generating explanations for DDIs based on molecular representations, we constructed an in-context demonstration prompting (**ExDDI-IC**) method. Based on the retrieval process described in the previous subsection, we retrieve the five most similar drug pairs for each test case. We then use their input, i.e., the SMILES representations, and the DDI label and explanation, as demonstrations to prompt ChatGPT to generate responses. The instructions used to prompt ChatGPT are shown in Appendix A (Sun et al. 2024).

Experiments

Experimental Setup

For hyper-parameter selection and training details, please refer to Appendix B (Sun et al. 2024).

Datasets We evaluate the model performance based on two databases: DDInter (Xiong et al. 2022) and DrugBank (v5.1.10). DrugBank is a widely used resource for training and evaluating DDI prediction models, but it only provides brief DDI explanations for open download. On the other hand, DDInter offers more extensive and detailed explanations involving pharmacodynamics and pharmacokinetics principles. The selection of these two datasets is motivated by their coverage of explanations with varying lengths, enabling us to gain insights into the model's performance when generating explanations of different complexities. We collected data on drug SMILES representations, drug descriptions, annotations of DDIs and relevant explanations to construct the datasets. The detailed description and statistics of the data are reported in Appendix C (Sun et al. 2024).

Settings for Model Generalisation Evaluation To examine the model's generalisation ability to new drugs, we followed previous work (Nyamabo et al. 2022; Li et al. 2023) to evaluate the model under both transductive and inductive settings. For ***transductive setting***, we evaluate the models' performance on unknown DDI pairs, allowing drugs from the training set to also appear in the test set. We randomly divided all positive and negative samples into training/validation/test sets with a ratio of 0.7/0.1/0.2. For ***inductive setting***, we evaluate the model's performance not only on unknown DDIs but also on unknown drugs. Specifically, the test set is split into *inductive S1* and *inductive S2* subsets according to whether both drugs are unavailable in the training set or only one drug is unavailable in the training set. We first divided drugs into three sets, M1, M2, and M3, with proportions of 0.75/0.05/0.2. Then, the training set consists of DDI samples where both drugs in the queried drug pair are from M1; The validation set includes samples where both drugs are from M2, or one is from M2 and the other is from M1;

	DrugBank											
	Transductive				Inductive Test S2				Inductive Test S1			
	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
ExDDI-IC	0.1187	0.2870	0.1192	0.2257	0.1174	0.2891	0.1161	0.2268	0.0674	0.2678	0.0808	0.2006
ExDDI-RV	0.5037	0.6780	0.5492	0.6250	0.4555	0.6261	0.4866	0.5717	0.2069	0.4557	0.2708	0.3906
ExDDI-S2S	0.9352	0.9410	**0.9109**	**0.9321**	0.5209	0.6470	0.5321	0.6179	**0.2197**	0.4451	0.2704	0.3915
ExDDI-MT	**0.9447**	0.9419	0.9076	0.9319	0.5157	0.6519	0.5344	0.6218	0.2071	0.4448	0.2701	0.3903
ExDDI-MTS	0.9441	**0.9421**	0.9073	0.9319	**0.5301**	**0.6590**	**0.5390**	**0.6281**	0.2145	**0.4578**	**0.2791**	**0.4023**
	DDInter											
	Transductive				Inductive Test S2				Inductive Test S1			
	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
ExDDI-IC	0.1367	0.2773	0.1301	0.2182	0.1257	0.2687	0.1184	0.2097	0.0657	0.2351	0.0715	0.1719
ExDDI-RV	0.5701	0.6456	0.5612	0.6080	0.4934	0.5780	0.4825	0.5376	**0.2500**	0.3825	0.2462	0.3305
ExDDI-S2S	**0.9392**	**0.9489**	**0.9371**	**0.9443**	**0.5261**	**0.6106**	**0.5249**	**0.5868**	0.2403	**0.3877**	**0.2481**	**0.3412**
ExDDI-MT	0.9383	0.9477	0.9352	0.9428	0.5161	0.6041	0.5160	0.5791	0.2309	0.3787	0.2365	0.3304
ExDDI-MTS	0.9384	0.9474	0.9350	0.9425	0.5172	0.6035	0.5151	0.5783	0.2309	0.3778	0.2353	0.3294

Table 1: Explanation generation results. Mean values over 5-fold cross-validation are presented for all models except ExDDI-IC, which was run only once due to the high cost of API calls and its low performance. The best results for each dataset are highlighted in bold. A complete table with standard deviations is provided in Appendix D to save page space.

The *inductive S1* test set contains samples where both drugs are from M3, and the *inductive S2* test set contains samples where one drug is from M1 and the other is from M3. We conduct 5-fold cross-validation for all settings.

Settings for Prediction Evaluation Although our primary objective of this work is to study methods for generating explanations for DDI prediction, we are also interested in exploring how the generation of explanations impacts the DDI prediction tasks.

Therefore, for ***binary prediction***, we introduce an ablation study to investigate the model's prediction performance without generating explanations, where we remove the generation loss in Eq. 4 and train the model for binary classification prediction only. We also use the results reported by Nyamabo et al. (2022) (GMPNN-CS) and Li et al. (2023) (DSN-DDI) as additional baselines for comparison.

For ***DDI-type prediction***, since our model does not directly learn from multi-class tasks, we estimate its performance by mapping the generated explanations to mechanism categories. Specifically, for the DrugBank data, we processed the model-generated explanations using scripts from Xiong et al. (2022)'s work and mapped the extracted quadruples to known categories. For the DDInter data, we first mapped the model-generated explanations to preprocessed DDInter explanation templates based on the nearest Levenshtein distance, and then corresponded the generated explanations to DrugBank explanation categories based on the statistical relationship between DDInter explanation templates and DrugBank explanation categories. This aims to enable comparison of the results across both datasets. We report DDI-type prediction results reported by Yan et al. (2021) (NMDADNN) and Xiong et al. (2022) (DDIMDL) for reference. However, it's worth noting that although the data sources are all DrugBank, there are differences between our constructed data and theirs, which may lead to discrepancies in explanation categories. As a result, the DDI type categories covered by the data used by Yan et al. (2021) and Xiong et al. (2022) are 65, while the number of DDI type categories extracted using the same method on our data is 257. Additionally, NMDADNN and DDIMDL utilised additional information, such as drug targets and enzymes, in their inputs. However, considering the context of predicting DDIs for new drugs, where such knowledge might be unknown, our method utilises the chemical molecular representations of drugs as inputs only. In light of these factors, our model faces greater challenges in multi-classification tasks, which should be considered when comparing models.

Evaluation Metrics We evaluate DDI explanation generation using metrics including BLEU, ROUGE-1, ROUGE-2, and ROUGE-L scores. To assess the models' classification performance, we present accuracy and F1 scores in Figure 3 and Figure 4, while full numerical values, including precision and recall scores, are provided in the Appendix D (Sun et al. 2024). For multi-class classification, we employ macro-averaging for precision, recall, and F1 scores.

Main Result: Explanation Generation

Table 1 presents the evaluation results of the models in explanation generation. The results show that 1) For all methods except ExDDI-IC, performance in the *transductive setting* is significantly better than in the *inductive S2 setting*, which in turn outperforms the *inductive S1 setting*. This highlights that existing models face substantial challenges when dealing with unseen molecules. ExDDI-IC performs poorly in all three settings, indicating that general LLMs still struggle to handle molecular representations effectively. 2) In the *transductive setting*, fine-tuning methods demonstrate a clear advantage over the other two categories, achieving promising results with scores exceeding 0.9. However, in the *inductive S2 setting*, the advantage of fine-tuning methods over retrieval methods is greatly reduced. In the *inductive S1 setting*, their performance is nearly equivalent to that of retrieval methods. This suggests that fine-tuning methods can effectively learn the similarity between

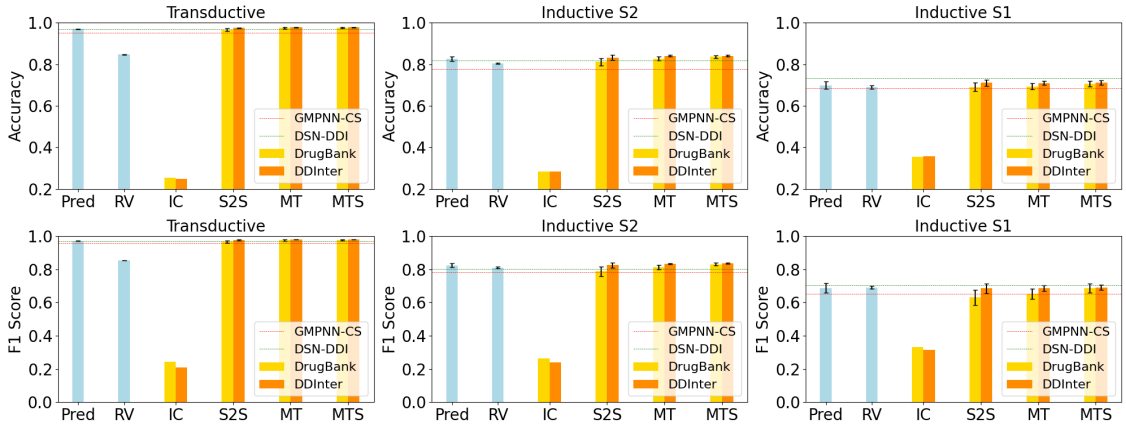

Figure 3: Binary classification results. Green and red dashed lines indicate results from the GMPNN-CS and DSN-DDI papers, respectively. Gold and orange bars represent models trained with DrugBank or DDInter explanations. Mean values over 5-fold cross-validation are shown for all models except ExDDI-IC. The error bars represent the standard deviation. Detailed numbers with precision and recall scores are available in the Appendix D.

drug molecules, but their generalisation ability to unseen molecules is still relatively poor. 3) There are no statistically significant differences in scores across the three fine-tuning paradigms. Even the simplest seq-to-seq scheme can produce competitive results. However, the standard deviation of results from multi-task training is relatively smaller (as reported in Appendix D), suggesting that they may offer greater stability. 4) Although the explanations in the DDInter data are richer and longer than those in DrugBank, there is no significant difference between their evaluation scores, indicating that text length is not the primary challenge in generating DDI explanations.

Prediction Performance of ExDDI Models

Binary Classification Figure 3 presents the performance of different models in the DDI binary prediction task. The results suggest that: 1) For all fine-tuning methods, the prediction results of models trained on DDInter data outperform those trained on DrugBank across all settings ($p < 0.05$, paired t-test), and the standard deviations of the scores are also smaller. This indicates that training models to generate more detailed DDI explanations can indeed be beneficial for the prediction task. As for ExDDI-IC, in most cases, the prediction performance is actually better when DrugBank explanations are provided as demonstrations compared to DDInter demonstrations. This may reveal the limited reasoning ability of general LLMs when dealing with complex explanations in the DDI task. 2) Compared to the **Pred** model trained without generation loss, i.e., by removing the decoder in Figure 2, the ExDDI-MT and ExDDI-MTS models trained on DDInter data generally perform better. However, the models trained on DrugBank data sometimes perform worse than the **Pred** model, indicating that overly brief explanations may not contribute to the prediction task. The ExDDI-S2S model, when trained on DrugBank data, may suffer more performance loss and exhibit greater fluctuations, suggesting that introducing a component-aware interaction module in multi-task training is beneficial. 3) Compared to previous state-of-the-art methods, our top-performing model is generally on par. Specifically, all our fine-tuning methods perform well in both the *transductive setting* and the *inductive S2 setting*, but they slightly underperform compared to DSN-DDI in the more demanding *inductive S1 setting*. This drop in performance suggests that, compared to prediction methods based on graph neural networks that learn 2D molecular features, using linear SMILES representation exhibits relatively poorer generalisation when dealing with unknown drugs, despite the potential benefits of supervision signals from explanations. However, it is important to note that effectively mapping 2D molecular representations learned through graph structures into a shared space with language representations to generate well-generalised explanations has not been thoroughly explored yet, making it a promising research direction for future work.

Multiple Classification Figure 4 shows the results of the model's DDI-type prediction. The results reveal that: 1) Retrieval and fine-tuning methods based on DDInter explanations continue to outperform models based on DrugBank in DDI-type prediction ($p < 0.05$, paired t-test), further demonstrating the value of more detailed explanations in enhancing prediction accuracy. In addition, the performance gap between ExDDI-IC, when using DDInter demonstrations versus DrugBank demonstrations, is more significant. This suggests that ChatGPT faces greater challenges in reasoning through complex explanations, particularly for more demanding tasks. 2) Fine-tuning methods perform better than retrieval methods in both the *transductive setting* and the *inductive S2 setting*. However, in the *inductive S1 setting*, their performance degrades to a level similar to that of retrieval methods, with the F1 score potentially even lower. Among the fine-tuning methods, ExDDI-S2S shows a slight edge in terms of mean values, albeit with a larger standard

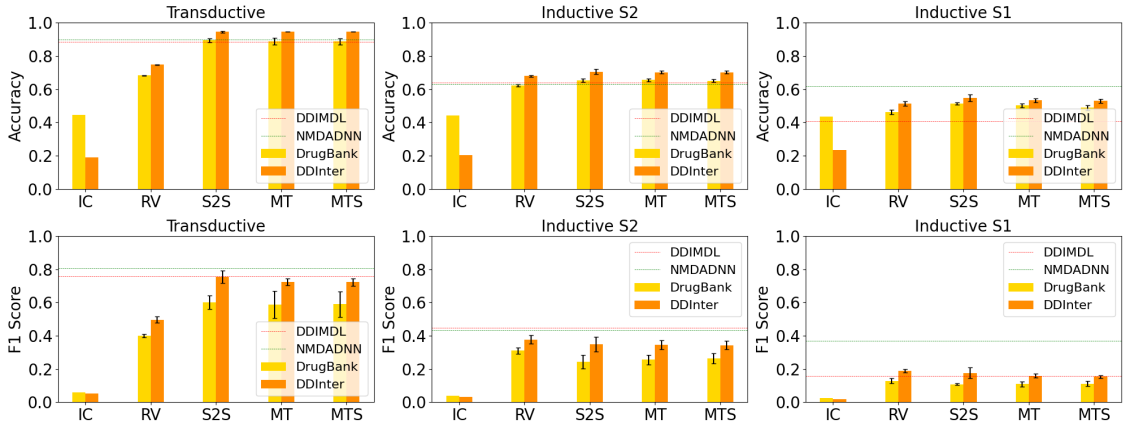

Figure 4: Multiple classification results. Green and red dashed lines indicate results from the DDIMDL and NMDADNN papers, respectively. Gold and orange bars represent models trained with DrugBank or DDInter explanations. Mean values over 5-fold cross-validation are shown for all models except ExDDI-IC. The error bars represent the standard deviation. Detailed numbers with precision and recall scores are available in the Appendix D.

deviation in most cases. These results are consistent with our analysis of the explanation generation outcomes and suggest a degree of interdependence between the two tasks. 3) Numerically, our method does not always achieve higher scores in the multi-classification task compared to previous methods. However, this may be due to our data encompassing a broader range of DDI-type labels when applying the same pre-processing script. Furthermore, the DDIMDL and NMDADNN methods utilise additional information beyond molecular expressions, such as target and enzyme data, which complicates direct comparisons.

Qualitative Analysis

To evaluate the quality of the model-generated explanations from multiple perspectives, we manually evaluated 20 data points from one fold of each setting in the DDInter dataset. The examples and results are detailed in Appendix E (Sun et al. 2024).

In summary, in the *transductive setting*, the fine-tuning models not only achieve good prediction results, but also generate explanations that are largely consistent with the annotations. The retrieval model, while generally accurate in its predictions, occasionally produces explanations that diverge from the annotations but remain relevant. The in-context prompting model, however, shows poor prediction outcomes and generates less relevant explanations.

In the *inductive S2 setting*, both the fine-tuning and retrieval models exhibit an increased likelihood of not predicting positive cases. According to their generated explanation, this is attributed to the models only correctly learning the characteristics of one of the drugs involved. Even among correctly predicted positive cases, the number of explanations that do not fully match the annotations increases, although they still retain some degree of relevance. For negative case explanations, although most models predict correctly, they tend to generate descriptions that are accurate for just one of drugs. Nevertheless, models trained with multi-tasking are more likely to generate descriptions that are both accurate and relevant for both drugs. Unlike in the *transductive setting*, the ExDDI-IC model shows a slight improvement in correct predictions, and for false positive predictions, a significant proportion remains relevant to the actual drug interactions, whereas false negative cases often fail to provide any useful information.

In the *inductive S1 setting*, the proportion of irrelevant explanations generated by the models is higher. Here, when predictions are correct, explanations generated by seq-to-seq models are more relevant than those from multi-task models, which is the opposite of what is observed in the *inductive S2 setting*. When predictions are incorrect, nearly all models fail to generate relevant explanations or only generate explanations related to one of the drugs.

Conclusion

This work introduces the task of generating natural language explanations for DDI predictions, advancing DDI computational methods towards a more trustworthy AI direction. We developed the ExDDI family of models for this task and conducted a thorough evaluation, providing tools and baselines for future research. The experimental results reveal that the top-performing methods can effectively predict and explain new DDI relationships for known drugs, but their ability to predict and explain DDIs involving new drugs still requires significant improvement. Additionally, our experiments indicate that training models to generate more detailed DDI explanations can enhance the prediction task itself. We believe that generalising to molecular structures unseen during training is a significant challenge for current DDI prediction and explanation generation models. Future research should consider incorporating the graph structure of chemical molecules and utilising multi-dimensional similarity information to learn more informative drug representations.

Ethical Statement

Our development of the DDI explanation generation model aims to aid researchers in identifying new DDIs. However, practitioners should be aware that the model may sometimes produce erroneous outputs and should use their professional expertise to assess the reliability of the model-generated content. It is important to note that the models and data presented in this work are not intended for use as medical advice.

Acknowledgements

This work was supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) through a Turing AI Fellowship (grant no. EP/V020579/1, EP/V020579/2).

References

Al-Ashwal, F. Y.; Zawiah, M.; Gharaibeh, L.; Abu-Farha, R.; and Bitar, A. N. 2023. Evaluating the sensitivity, specificity, and Accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and bard against conventional drug-drug interactions clinical tools. *Drug, Healthcare and Patient Safety*, 137–147.

Asada, M.; Miwa, M.; and Sasaki, Y. 2023. Integrating heterogeneous knowledge graphs into drug–drug interaction extraction from the literature. *Bioinformatics*, 39(1): btac754.

Baitai, C.; Peng, J.; Zhang, Y.; and Liu, Y. 2023. Molecular Structure-Based Double-Central Drug-Drug Interaction Prediction. In *International Conference on Artificial Neural Networks*, 127–138. Springer.

Camburu, O.-M.; Rocktäschel, T.; Lukasiewicz, T.; and Blunsom, P. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Cheng, F.; and Zhao, Z. 2014. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association*, 21(e2): e278–e286.

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.

Durant, J. L.; Leland, B. A.; Henry, D. R.; and Nourse, J. G. 2002. Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6): 1273–1280.

Edwards, C.; Lai, T.; Ros, K.; Honke, G.; Cho, K.; and Ji, H. 2022. Translation between Molecules and Natural Language. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 375–413. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Ferdousi, R.; Safdari, R.; and Omidi, Y. 2017. Computational prediction of drug-drug interactions based on drugs functional similarities. *Journal of biomedical informatics*, 70: 54–64.

Gottlieb, A.; Stein, G. Y.; Oron, Y.; Ruppin, E.; and Sharan, R. 2012. INDI: a computational framework for inferring drug interactions and their associated recommendations. *Molecular systems biology*, 8(1): 592.

Han, K.; Cao, P.; Wang, Y.; Xie, F.; Ma, J.; Yu, M.; Wang, J.; Xu, Y.; Zhang, Y.; and Wan, J. 2022. A review of approaches for predicting drug–drug interactions based on machine learning. *Frontiers in pharmacology*, 12: 814858.

Hase, P.; Zhang, S.; Xie, H.; and Bansal, M. 2020. Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language? In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4351–4367. Online: Association for Computational Linguistics.

Khori, V.; Semnani, S.; and Roshandel, G. 2011. Frequency distribution of drug interactions and some of related factors in prescriptions. *Medical Journal of Tabriz University of Medical Sciences*, 27(4): 29–32.

Lampinen, A.; Dasgupta, I.; Chan, S.; Mathewson, K.; Tessler, M.; Creswell, A.; McClelland, J.; Wang, J.; and Hill, F. 2022. Can language models learn from explanations in context? In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 537–563. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Lee, G.; Park, C.; and Ahn, J. 2019. Novel deep learning model for more accurate prediction of drug-drug interaction effects. *BMC bioinformatics*, 20: 1–8.

Li, Z.; Zhu, S.; Shao, B.; Zeng, X.; Wang, T.; and Liu, T.-Y. 2023. DSN-DDI: an accurate and generalized framework for drug–drug interaction prediction by dual-view representation learning. *Briefings in Bioinformatics*, 24(1): bbac597.

Lin, S.; Wang, Y.; Zhang, L.; Chu, Y.; Liu, Y.; Fang, Y.; Jiang, M.; Wang, Q.; Zhao, B.; Xiong, Y.; et al. 2022. MDF-SA-DDI: predicting drug–drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism. *Briefings in Bioinformatics*, 23(1): bbab421.

Ma, M.; and Lei, X. 2023. A dual graph neural network for drug–drug interactions prediction based on molecular structure and interactions. *PLOS Computational Biology*, 19(1): e1010812.

Narang, S.; Raffel, C.; Lee, K.; Roberts, A.; Fiedel, N.; and Malkan, K. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Nyamabo, A. K.; Yu, H.; Liu, Z.; and Shi, J.-Y. 2022. Drug–drug interaction prediction with learnable size-adaptive molecular substructures. *Briefings in Bioinformatics*, 23(1): bbab441.

OpenAI. 2022. ChatGPT.

Rajani, N. F.; McCann, B.; Xiong, C.; and Socher, R. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In Korhonen, A.; Traum, D.; and

Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4932–4942. Florence, Italy: Association for Computational Linguistics.

Ren, Z.-H.; Yu, C.-Q.; Li, L.-P.; You, Z.-H.; Guan, Y.-J.; Wang, X.-F.; and Pan, J. 2022. BioDKG–DDI: predicting drug–drug interactions based on drug knowledge graph fusing biochemical information. *Briefings in Functional Genomics*, 21(3): 216–229.

Rohani, N.; and Eslahchi, C. 2019. Drug-drug interaction predicting by neural network using integrated similarity. *Scientific reports*, 9(1): 13645.

Rohani, N.; Eslahchi, C.; and Katanforoush, A. 2020. ISCMF: Integrated similarity-constrained matrix factorization for drug–drug interaction prediction. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 9: 1–8.

Ryu, J. Y.; Kim, H. U.; and Lee, S. Y. 2018. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the national academy of sciences*, 115(18): E4304–E4311.

Safdari, R.; Ferdousi, R.; Aziziheris, K.; Niakan-Kalhori, S. R.; and Omidi, Y. 2016. Computerized techniques pave the way for drug-drug interaction prediction and interpretation. *BioImpacts: BI*, 6(2): 71.

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

Sun, Z.; Li, J.; Pergola, G.; and He, Y. 2024. ExDDI: Explaining Drug-Drug Interaction Predictions with Natural Language. arXiv:2409.05592.

Vilar, S.; Harpaz, R.; Uriarte, E.; Santana, L.; Rabadan, R.; and Friedman, C. 2012. Drug—drug interaction through molecular structure similarity analysis. *Journal of the American Medical Informatics Association*, 19(6): 1066–1074.

Vo, T. H.; Nguyen, N. T. K.; Kha, Q. H.; and Le, N. Q. K. 2022. On the road to explainable AI in drug-drug interactions prediction: A systematic review. *Computational and Structural Biotechnology Journal*, 20: 2112–2123.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wiegreffe, S.; Hessel, J.; Swayamdipta, S.; Riedl, M.; and Choi, Y. 2022. Reframing Human-AI Collaboration for Generating Free-Text Explanations. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 632–658. Seattle, United States: Association for Computational Linguistics.

Xiong, G.; Yang, Z.; Yi, J.; Wang, N.; Wang, L.; Zhu, H.; Wu, C.; Lu, A.; Chen, X.; Liu, S.; et al. 2022. DDInter: an online drug–drug interaction database towards improving clinical decision-making and patient safety. *Nucleic acids research*, 50(D1): D1200–D1207.

Yan, X.-Y.; Yin, P.-W.; Wu, X.-M.; and Han, J.-X. 2021. Prediction of the drug–drug interaction types with the unified embedding features from drug similarity networks. *frontiers in Pharmacology*, 12: 794205.

Yordanov, Y.; Kocijan, V.; Lukasiewicz, T.; and Camburu, O.-M. 2022. Few-Shot Out-of-Domain Transfer Learning of Natural Language Explanations in a Label-Abundant Setup. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 3486–3501. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Zhang, C.; Lu, Y.; and Zang, T. 2022. CNN-DDI: a learning-based method for predicting drug–drug interactions using convolution neural networks. *BMC bioinformatics*, 23(Suppl 1): 88.

Zhang, W.; Chen, Y.; Li, D.; and Yue, X. 2018. Manifold regularized matrix factorization for drug-drug interaction prediction. *Journal of biomedical informatics*, 88: 90–97.

Zhang, Y.; Deng, Z.; Xu, X.; Feng, Y.; and Junliang, S. 2023. Application of Artificial Intelligence in Drug–Drug Interactions Prediction: A Review. *Journal of Chemical Information and Modeling*.

Zhu, J.; Liu, Y.; Zhang, Y.; Chen, Z.; and Wu, X. 2022. Multi-Attribute Discriminative Representation Learning for Prediction of Adverse Drug-Drug Interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 10129–10144.