DDIAgents: Mechanism-Conditioned Context Flow for Drug-Drug Interaction Prediction

Zhenqian Shen
szg22@mails.tsinghua.edu.cn
Department of Electronic Engineering,
Tsinghua University
Beijing, China

Yu Liu
yu.liu@eng.ox.ac.uk
Institute of Biomedical Engineering,
University of Oxford
Oxford, United Kingdom

Xiaoyi Fu
xiaoyif@alumni.cmu.edu
Division of Emerging Interdisciplinary Areas,
Hong Kong University of Science and Technology
Hong Kong, China

Quanming Yao*
qyaoaa@tsinghua.edu.cn
Department of Electronic Engineering
Tsinghua University
Beijing, China

Abstract

Drug-drug interaction (DDI) prediction is essential for medication safety, yet it requires reasoning over heterogeneous biomedical evidence whose relevance changes across interaction mechanisms. We propose DDIAgents, a mechanism-conditioned multi-agent framework that performs DDI prediction through dynamic knowledge orchestration. Given a drug pair, a planner agent instantiates specialized expert agents, routes mechanism-relevant knowledge sources to each agent, and aggregates their analyses through a conclusion agent. By adapting context flow to the inferred interaction mechanism, DDIAgents reduces irrelevant information, supports complementary expert reasoning, and produces interpretable agent-level rationales. Extensive experiments on realistic DDI prediction benchmarks show that DDIAgents consistently outperforms existing feature-based, graph-based, LLM-based, and agent-based baselines. Beyond prediction performance, DDIAgents demonstrates how multi-agent systems can organize heterogeneous scientific knowledge for adaptive and interpretable AI4Science reasoning.

CCS Concepts

• **Applied computing** → **Bioinformatics**; • **Computing methodologies** → **Multi-agent systems**; *Machine learning approaches.*

Keywords

Drug-Drug Interaction Prediction, Multi-Agent System, Dynamic Context Flow

ACM Reference Format:
Zhenqian Shen, Yu Liu, Xiaoyi Fu, and Quanming Yao. 2026. DDIAgents: Mechanism-Conditioned Context Flow for Drug-Drug Interaction Prediction. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3770855.3818959

*Corresponding author

KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818959

1 Introduction

Modern drug discovery is defined by prohibitive costs and protracted timelines, frequently taking 10–15 years and over 2 billion dollars per approved drug [5, 21]. It is therefore crucial to identify drug-drug interactions (DDIs) that could lead to adverse drug reactions and serious health risks for patients. However, DDI prediction remains a challenging scientific problem due to intricate biomedical mechanisms, including altered pharmacokinetics and pharmacodynamics [20], alongside clinical treatment factors [24]. Moreover, candidate DDI verification is expensive and time-consuming, requiring extensive experiments from preclinical studies to large-scale clinical trials. This setting reflects a broader data-efficient AI4Science challenge: direct labels and experimental feedback are costly, so models must make better use of prior knowledge, task structure, and limited feedback rather than relying only on scaling supervision [36, 37]. In this work, we focus on a complementary agentic question: how should a reasoning system decide which biomedical evidence each expert should use when the underlying interaction mechanism changes across drug pairs?

To accelerate DDI discovery, a variety of computational methods have been developed for prediction [14, 22]. These methods broadly fall into three categories: feature-based, graph-based, and large language model (LLM)-based approaches. Feature-based methods [16, 26] typically encode drugs with molecular features and train neural networks for interaction prediction. Graph-based methods broadly utilize graph embedding methods, graph neural networks (GNNs) or graph transformers to extract predictive features from diverse data sources, ranging from individual drug molecular structures [19, 42] to complex biomedical networks [46, 48]. LLM-based methods primarily utilize the powerful reasoning capabilities based on textual descriptions of drugs and interactions [41, 47].

Despite these advances, DDI prediction is not merely a single-pattern learning problem; it is a mechanism-conditioned knowledge orchestration problem. First, different drug pairs can be governed by **diverse mechanisms**, from physiological modulation and drug–target binding events to chemically mediated reactions. Second, the decisive evidence exhibits **knowledge heterogeneity**, spanning

natural-language descriptions, structured biomedical relations, and molecular substructures. A metabolism-driven interaction may require enzyme or transporter evidence, whereas a target-overlap interaction may depend more on pathway or adverse-effect evidence. However, most existing approaches remain inflexible: they follow fixed reasoning patterns and rely on static evidence integration, which can introduce irrelevant evidence or miss the modality that is decisive for a specific mechanism.

Meanwhile, the recent emergence of LLM-based multi-agent systems in other scientific domains has become a promising solution for providing adaptable mechanism analysis of DDIs. Analogous to a multidisciplinary research team, multi-agent frameworks coordinate multiple role-specialized "expert" agents to solve complex problems [3, 31]. This perspective also aligns with data-efficient agentic learning, where performance is improved by organizing roles, feedback, and interactions rather than simply increasing labeled data or model scale [36]. In such systems, the reasoning performance critically depends on the **context flow**, namely how knowledge sources are routed to the agents that need it. However, existing multi-agent systems often rely on a fixed context flow, where agents are restricted to a predefined knowledge scope. This rigidity forces expert agents to handle a fixed scope of knowledge regardless of the specific task, which can introduce irrelevant information into the reasoning process and obscure the critical insights.

To address the challenges above, we introduce DDIAgents, a mechanism-conditioned multi-agent framework for DDI prediction that leverages dynamic context flow for specialized DDI analysis. Specifically, DDIAgents conducts the following three stages iteratively. Firstly, expert agent instantiation gathers a team of specialized expert agents. Secondly, dynamic context flow routes suitable knowledge sources to each expert agent according to the mechanism-specific evidence needs of the current drug pair. Finally, expert agents generate domain-specific analyses that are synthesized by a conclusion agent to either yield a final prediction or provide strategic guidance for iterative refinement. Experiments on two benchmark datasets demonstrate the superior performance of DDIAgents compared to existing methods. Furthermore, the agent-based analysis demonstrates that DDIAgents yields interpretable mechanistic explanations of predicted DDIs. The main contributions are summarized as follows:

- We formulate DDI prediction as a mechanism-conditioned knowledge orchestration problem and introduce DDIAgents, a multi-agent framework that explicitly accounts for diverse mechanisms and heterogeneous biomedical evidence.
- We design a dynamic context flow mechanism that strategically routes heterogeneous knowledge sources to specialized agents, thereby enhancing knowledge utility and reducing noisy information in the DDI reasoning process.
- We perform extensive experiments on two benchmark datasets, demonstrating that DDIAgents consistently outperforms existing feature-based, graph-based, LLM-based, and agent-based approaches. Comprehensive case studies further show that DDIAgents provides interpretable agent-level insights into the underlying mechanisms of DDIs.

2 Related Works

2.1 Drug-Drug Interaction Prediction

DDI prediction has recently been a critical topic in computational pharmacology, aiming to identify potential drug interactions that may lead to adverse effects or altered therapeutic efficacy [11, 14]. Existing approaches are commonly grouped into feature-based, graph-based, and LLM-based methods.

Feature-based methods encode drugs with molecular descriptors and train supervised predictors [16, 25, 26, 40]. They work well when interactions are driven by chemical properties, but struggle to incorporate broader biological mechanisms. Graph-based methods exploit structured relations from molecular graphs and biomedical networks through knowledge graph embedding methods [43], GNNs [46, 48] or heterogeneous graph transformers [30]. While they better capture relational evidence, their pipelines are typically fixed, limiting adaptability across drug pairs. Recently, LLM-based methods prompt models with drug or interaction descriptions [23, 41, 47], retrieve graph-path evidence [1], or apply case-based reasoning [15]. CBR-DDI shows that historical cases can provide useful reasoning examples for DDI prediction. Unlike case-based approaches, DDIAgents focuses on routing heterogeneous biomedical evidence to specialized agents according to mechanism hypotheses. Thus, DDIAgents targets mechanism-conditioned knowledge orchestration rather than rationale transfer.

Overall, existing DDI prediction methods provide complementary strengths, but most still follow uniform reasoning patterns and static knowledge utilization. This rigidity is a key limitation for DDI prediction, where different drug pairs require different explanatory routes and different forms of supporting evidence.

2.2 Multi-Agent Reasoning for Scientific Tasks

Powered by strong natural language understanding and generation, LLMs have recently enabled a new paradigm for tackling complex scientific tasks via multi-agent collaboration [3, 13, 31]. Compared with conventional machine learning, multi-agent systems decompose a task for role-specialized agents, each equipped with distinct domain expertise and responsibilities. This design principle is closely related to data-efficient agentic learning, which studies how agent systems can adapt under limited supervision, limited feedback, and costly interactions [36].

Coscientist [3] is an early framework that formulates an end-to-end workflow for autonomous chemical research, coordinating agents such as planner agent, document searchers and experiment executors. SciAgents [9] builds a materials-science knowledge graph and organizes a team to generate hypotheses by following evidence paths extracted from the knowledge graph. Virtual Lab [31] instantiates a group of LLM scientist agents for interdisciplinary research and has been applied to novel nanobody design. Related directions also explore multi-agent systems for clinical decision-making, assembling expert agents with distinct specialties [13, 32].

Despite these advances, many multi-agent systems rely on fixed agent configurations and static information-access patterns. In scientific problems where different sub-problems require different knowledge sources, the way information is allocated across agents can substantially influence reasoning quality and interpretability.

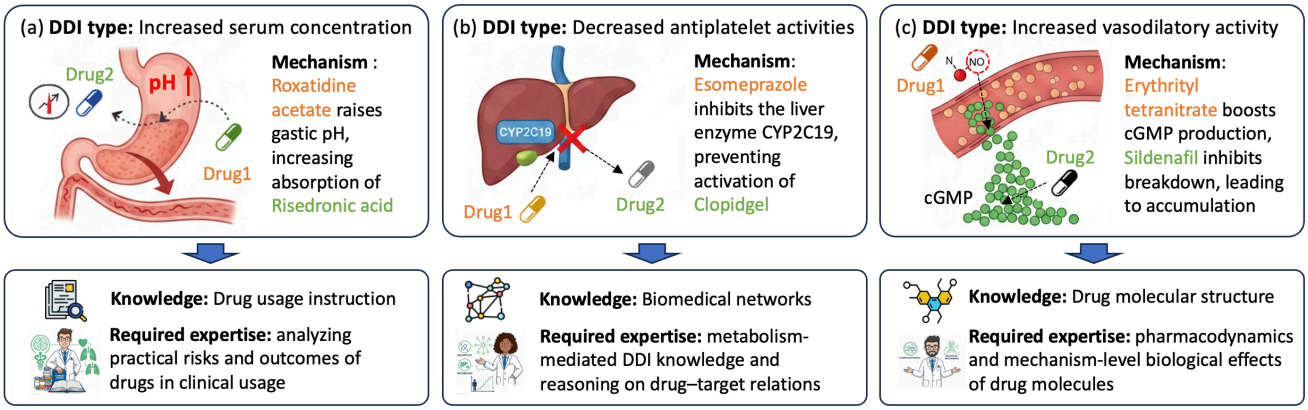


Figure 1: Three DDI examples with different mechanisms: (a) pharmacokinetic change, (b) pharmacodynamic interference, (c) chemically mediated synergy. The prediction of these DDIs requires different knowledge and expertise.

3 Motivation

3.1 DDI Prediction as a Mechanism-conditioned Reasoning Problem

Drug–drug interaction (DDI) prediction is fundamental to medication safety, yet it remains challenging in realistic settings because **it requires mechanism-conditioned reasoning rather than single-pattern learning**. Clinically meaningful DDIs arise from different mechanisms, and the correct interaction type for each drug pair depends on identifying the governing mechanism, rather than applying a universal decision rule.

Figure 1 highlights that clinically meaningful DDIs can be driven by diverse mechanisms, including pharmacokinetic change, pharmacodynamic interference and chemically mediated synergy. These mechanisms correspond to qualitatively different causal explanations and therefore require different reasoning paths. **Effective DDI prediction thus requires selecting an appropriate explanatory route before determining the interaction type.**

A closely related challenge is that **the decisive evidence is heterogeneous, mechanism-conditioned, and demands complementary expert perspectives**. No single modality or fixed expert viewpoint is uniformly informative across all DDIs. When knowledge is aggregated without expert-guided selection, critical evidence can be diluted by irrelevant context and cross-perspective distraction, ultimately undermining reliable prediction.

3.2 Limitations of Static Reasoning and the Need for Dynamic Context Flow

Despite substantial progress, most existing DDI prediction methods implicitly assume static reasoning. Feature-based and graph-based models rely on fixed pipelines and static feature fusion, while recent LLM-based methods typically employ fixed prompts and retrieval scopes. As a result, **they reuse the same reasoning pattern and evidence across drug pairs**, which can either omit the modality most relevant to the governing mechanism or overwhelm the model with irrelevant context.

Multi-agent reasoning provides a natural abstraction for incorporating multiple expert perspectives, but **naive multi-agent designs remain insufficient**. Most existing multi-agent systems for scientific tasks adopt a fixed context flow, i.e., heterogeneous knowledge is routed to expert agents according to a predetermined scheme. This rigidity constrains experts to static information scopes and prevents adaptation when mechanisms vary across instances. **Therefore, the primary bottleneck is not agent collaboration itself, but how knowledge is allocated to agents.**

These observations lead to a key insight: **mechanism-selective biomedical reasoning requires instance-adaptive context flow**. For each drug pair, the system should dynamically determine which experts to instantiate, which types of knowledge each expert should access, and whether these choices need to be revised as intermediate analyses reveal uncertainty or disagreement. This insight directly motivates **DDIAgents**, which models expert instantiation and knowledge routing as per-instance and iterative decisions, enabling targeted, interpretable and robust DDI prediction.

4 Methodology

4.1 Problem Setup

Let $\mathcal{D}$ be the set of all drugs and $\mathcal{R}$ be the set of possible interaction types. In the DDI prediction problem, effective utilization of heterogeneous knowledge sources is critical. We define a collection of knowledge sources as $\mathcal{S}$ that are relevant to the reasoning process in DDI prediction. Formally, the DDI prediction problem is to learn a predictor p that maps a pair of drugs and the integrated knowledge sources to a DDI type. For any drug pair $(u, v) \in \mathcal{D} \times \mathcal{D}$ and knowledge sources $\mathcal{S}$, the predicted DDI type $r \in \mathcal{R}$ is given by:

$$r = p(u, v; \mathcal{S}) \tag{1}$$

For agent-based methods, we implement a filtering stage to handle the context overflow caused by an exhaustive list of DDI candidates. A traditional model [30, 41] is trained to filter a predefined number of the most probable DDI candidates. The DDI prediction problem and these candidates are then formatted into a multiple-choice question q, serving as the input for agent-based methods.

4.2 Overall Framework

The framework of DDIAgents (as shown in Figure 2) incorporates the following three stages: **(a) Expert agent instantiation:** given a DDI query, a planner agent instantiates a tailored set of expert

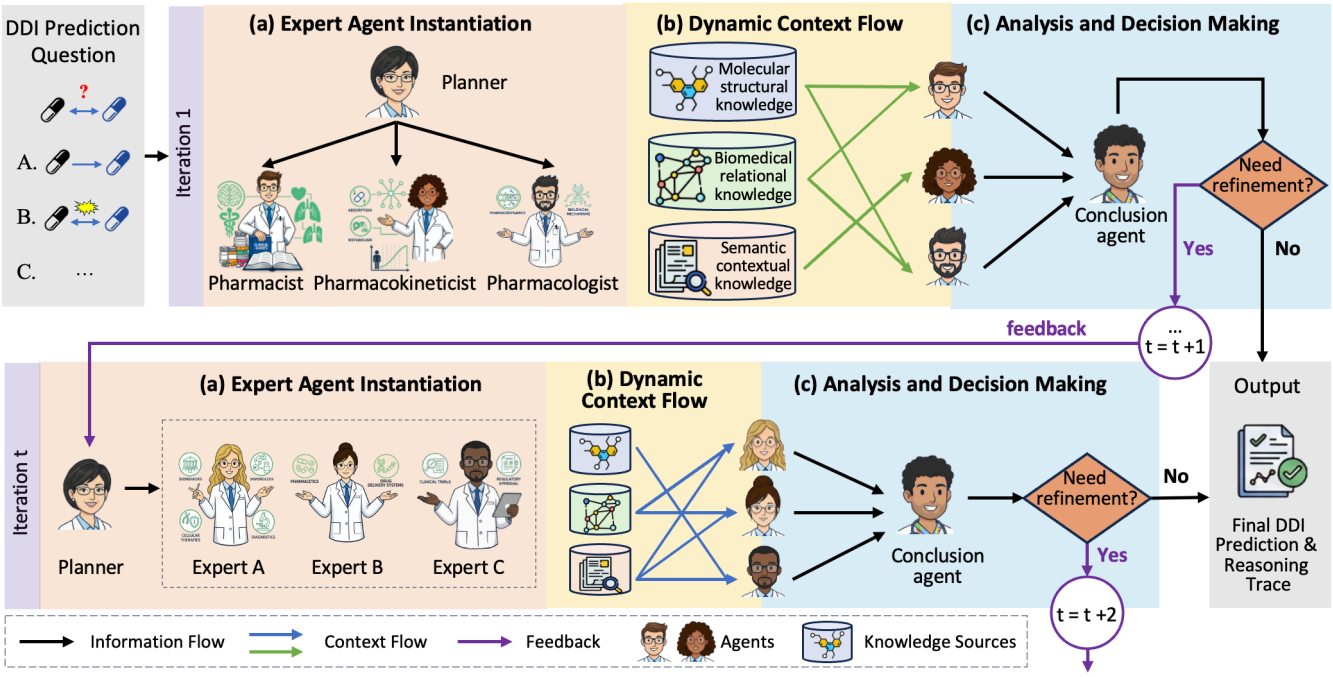


Figure 2: The overall framework of the proposed DDIAgents, which iteratively conducts three stages: (a) Expert agent instantiation, (b) Dynamic context flow, (c) Analysis and decision making.

agents with distinct expertise. This enables adaptive expertise composition, avoiding rigid agent configurations. **(b) Dynamic context flow:** the planner agent assigns each expert a mechanism-relevant subset of knowledge sources. This mitigates information overload and supports targeted and specialized reasoning. **(c) Analysis and decision making:** expert agents conduct specialized analyses, and a conclusion agent synthesizes their outputs to either deliver a final prediction or request another analysis iteration.

If the conclusion agent detects unresolved uncertainty or conflicting evidence, its feedback serves as an update signal for the next iteration, guiding expert agent instantiation and dynamic context flow to avoid repetition of the same reasoning pattern.

4.3 Expert Agent Instantiation

Selecting appropriate experts is essential for reliable and mechanism-aware DDI interpretation. In the first iteration ($t = 1$), we instantiate three core experts to cover complementary perspectives on clinical risk, pharmacokinetics, and biological mechanism:

- Pharmacist [38]: Focuses on practical, patient-oriented risks and outcomes of drugs in clinical usage.
- Pharmacokineticist [44]: Analyzes absorption, distribution, metabolism and excretion to identify exposure-altering interactions.
- Pharmacologist [18]: Reasons about pharmacodynamic properties and mechanism-level biological effects.

For subsequent iterations ($t > 1$), the framework moves beyond this fixed configuration. Instead of reusing the same roles, the planner analyzes the original query together with the conclusion agent's feedback from iteration $t - 1$ to instantiate a new set of experts. Concretely, the planner agent identifies unresolved gaps or conflicts in the last iteration, and then specifies expertise profiles via targeted role instructions (see Appendix A.8). This adaptive instantiation allows DDIAgents to revise its reasoning trajectory to match the case-specific difficulty, rather than applying a fixed analytic pattern across all drug pairs.

4.4 Dynamic Context Flow

To support specialized and rigorous expert analyses, we introduce a dynamic context flow that routes knowledge sources to experts in an expertise-aware and mechanism-dependent manner. Concretely, we organize domain knowledge into a set of heterogeneous knowledge sources, each capturing a complementary view:

- **Molecular Structural Knowledge**: We incorporate the intrinsic chemical topology of drugs. Specifically, we utilize SMILES strings to represent the atom-bond connectivity and functional groups of drug molecules, which are fundamental to understanding the steric and electronic basis of interactions.
- **Biomedical Relational Knowledge**: We leverage the systemic connectivity of drugs within the biological network. This encompasses known DDI training data, drug-target associations and drug side-effect profiles. This view places the drug within the broader context of the biomedical system.
- **Semantic Contextual Knowledge**: We utilize natural language descriptions to capture high-level pharmacological concepts. This includes concise summaries of drug mechanisms and interaction types generated by LLMs [2], providing a semantic layer that complements raw structural data.

Given these knowledge sources, the dynamic context flow controls how they are assigned to expert agents. Specifically, the planner agent assigns each expert agent a subset of knowledge sources that aligns with the agent's expertise and the current reasoning

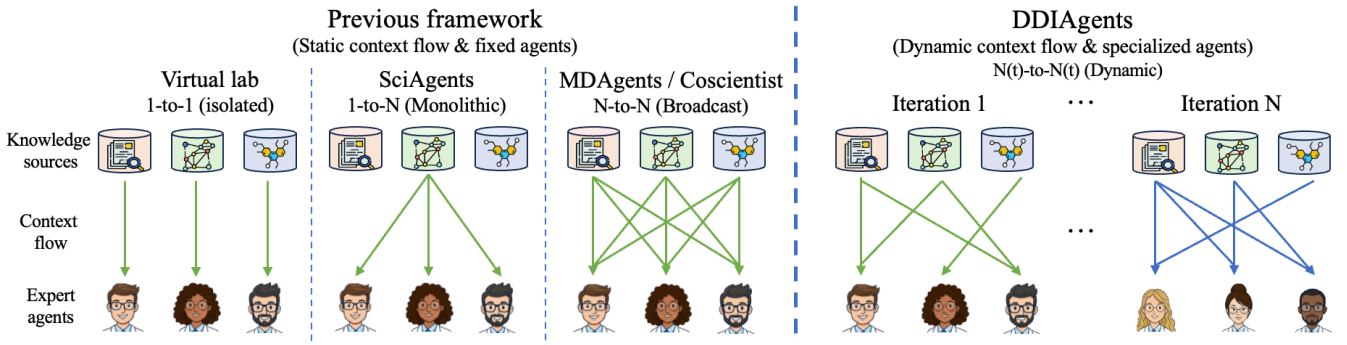


Figure 3: Comparison between DDIAgents framework and previous multi-agent methods.

demands. As a result, agents may receive different combinations of evidence within the same iteration. For instance, a chemistry-oriented expert may be assigned primarily structural evidence, while a biology-oriented expert may focus on relational evidence, and a clinical expert may rely more on semantic descriptions and safety-oriented relational cues. This allocation is updated across iterations using the intermediate feedback from the conclusion agent, enabling the system to shift attention as the hypothesized mechanism sharpens or new uncertainties emerge.

By turning knowledge allocation into a targeted, expertise-aware pipeline, dynamic context flow mitigates two common failure modes in DDI reasoning. Firstly, it reduces information dilution [28], where mechanism-critical evidence is buried in irrelevant context. Secondly, it alleviates cross-domain distraction [27], where experts are forced to process outside their functional role. This design prepares for the next stage of the framework: since each expert receives a role-aligned subset of knowledge sources, the subsequent analysis step can construct a tailored evidence context per expert and elicit deeper, more focused reasoning.

4.5 Analysis and Decision Making

At each iteration t, before expert agents begin reasoning, the system constructs a tailored knowledge context for them. This context is formed by retrieving and filtering the most pertinent evidence from the knowledge sources assigned in the current iteration (see Appendix A.3). The goal is to provide a focused but sufficient knowledge input, which is rich enough to support a mechanism-grounded analysis, while avoiding unnecessary information that can obscure critical cues.

Given the query and its curated context, each expert produces a DDI prediction and a structured rationale grounded in the provided evidence, reflecting its domain perspective. These expert reports are then passed to a conclusion agent, which synthesizes the set of analyses into a single decision. Rather than treating expert outputs as independent opinions, the conclusion agent aggregates them into a coherent explanation that resolves cross-expert dependencies.

The conclusion agent has two types of possible outputs. If the evidence and expert analyses are sufficient, it finalizes the DDI prediction along with a consolidated reasoning trace. Otherwise, it produces iteration feedback that explicitly identifies what remains uncertain and recommends how the next iteration should adjust—both in terms of which new experts to instantiate and which knowledge sources to prioritize. This feedback closes the loop, ensuring that subsequent iterations are guided by concrete suggestions rather than repeating the same reasoning pattern.

Table 1: Knowledge source utilization comparison.

Method \ Knowledge	Molecular Structure	Biomedical Network	Text Description
Feature-based [16, 26]	✓	-	-
Graph-based [19, 48]	✓	✓	-
LLM-based [41, 47]	-	-	✓
DDIAgents (Ours)	✓	✓	✓

4.6 Comparison with Existing Works

Table 1 reveals a key limitation of existing DDI prediction methods: they typically pre-commit to a narrow knowledge source scope. Feature-based methods mainly depend on molecular structure, graph-based methods incorporate biomedical networks but usually exclude unstructured descriptions, and LLM-based approaches usually depend on text alone. As a result, when the decisive cue for a drug pair lies outside the model's preferred modality, predictions can fail due to missing evidence. In contrast, DDIAgents simultaneously considers all of these knowledge sources, and routes them to experts in a mechanism-aware, instance-adaptive manner. Figure 3 further contrasts DDIAgents with representative multi-agent frameworks for scientific tasks. Unlike existing systems that rely on static execution, DDIAgents adopts an adaptive and iterative reasoning framework. Rather than maintaining fixed agent roles, DDIAgents leverages specialized agent instantiation to generate specialized experts tailored to specific sub-problems as the task evolves. The most significant distinction lies in the context flow, with different systems applying different mapping patterns:

- 1-to-1 (Isolated): Virtual Lab [31] assigns specific, pre-defined knowledge sources to specific agents in a fixed, parallel mapping.
- 1-to-N (Monolithic): SciAgents [9] relies on a single, monolithic knowledge graph distributed across multiple agents.
- N-to-N (Broadcast): Coscientist [3] and MDAgents [13] distribute all available information to every agent uniformly.
- N(t)-to-N(t) (Dynamic): DDIAgents treats the relation between knowledge sources and agents as variables that evolve over time. It adaptively maps pertinent, domain-specific data to the relevant experts based on the current reasoning state.

Table 2: General performance comparison between different DDI prediction methods. Results are reported as mean±std over 5 independent runs. The best and second-best results are highlighted in bold and underline, respectively.

Type	Method	DrugBank				TWOSIDES			
		S1		S2		S1		S2	
		Acc	F1	Acc	F1	Hit@5	NDCG@5	Hit@5	NDCG@5
Feature-based	MLP	34.25±1.15	11.00±0.81	20.93±1.29	2.80±0.41	40.08±2.19	14.29±0.72	26.79±0.35	7.24±0.47
Graph-based	Decagon	20.07±1.59	3.49±0.54	6.87±0.22	1.96±0.11	32.75±1.13	10.71±0.57	24.22±0.65	6.59±0.29
	EmerGNN	29.57±1.43	13.32±0.37	15.33±0.49	3.54±0.07	35.84±1.75	11.07±0.68	28.35±1.34	8.56±0.72
	MSTE	31.60±0.37	12.87±0.12	12.73±0.14	1.80±0.07	30.94±0.62	10.69±0.14	25.00±1.13	6.67±0.45
	TIGER	45.98±2.35	26.78±1.33	20.64±1.44	3.32±0.37	42.24±0.38	15.00±0.21	25.26±0.17	6.65±0.09
LLM-based	Single LLM	8.71±0.37	4.10±0.32	7.30±0.29	3.94±0.23	27.30±0.31	7.91±0.17	23.78±0.74	6.58±0.36
	K-Path	35.34±1.08	21.41±0.59	19.48±0.48	3.71±0.19	31.15±0.43	9.57±0.18	24.01±0.38	5.82±0.07
	CBR-DDI	41.38±1.68	31.47±0.76	23.42±0.41	4.48±0.47	35.07±1.51	13.09±0.47	32.77±0.34	10.31±0.18
	TextDDI	49.53±2.14	31.72±1.27	28.75±0.91	8.26±0.37	42.04±1.86	15.38±0.79	29.35±1.18	9.77±0.75
	DDI-GPT	50.68±0.79	34.71±1.07	33.47±1.21	15.01±0.61	43.90±2.61	17.31±1.44	33.51±1.34	10.24±0.68
Agent-based	Reflexion	47.31±2.34	35.11±2.45	31.74±1.76	14.07±1.22	41.47±1.92	17.04±1.21	30.71±1.44	9.55±0.51
	Debate	52.18±1.46	40.14±1.17	35.03±1.64	15.84±0.78	50.61±2.07	19.03±0.74	38.66±1.41	12.78±0.62
	AgentVerse	52.34±2.08	41.17±1.21	36.48±0.56	14.01±0.38	49.33±1.46	19.44±0.59	40.38±0.76	13.14±0.24
	MDAgents	54.69±0.74	39.52±0.54	36.29±1.17	15.97±0.42	51.42±2.14	20.07±1.36	40.09±1.61	12.64±0.75
	DDIAgents	**55.75±0.74**	**45.52±0.89**	**38.21±1.03**	**18.17±0.51**	**53.94±1.24**	**21.21±0.48**	**42.26±0.92**	**14.08±0.15**

5 Experiments

5.1 Experimental Setup

Datasets. Following [45, 46], we conduct experiments on two widely used public DDI prediction datasets: DrugBank [39], a multi-class DDI prediction dataset where each drug pair is associated with one of 86 possible interaction types; TWOSIDES [33], a multilabel DDI prediction dataset that records 209 possible side effects.

Experimental Setting. To incorporate realistic factors in evaluation, we partition the drug set $\mathcal{D}$ into known drugs $\mathcal{D}_{\text{known}}$ and new drugs $\mathcal{D}_{\text{new}}$ based on approval time. We consider three prediction tasks: S0 (known–known), S1 (known–new), and S2 (new–new). In this work, we focus on the more realistic and challenging S1/S2 tasks, and report S0 results in Appendix B.1.

Evaluation Metrics. For the DrugBank dataset that features a single interaction type per drug pair, we employed Accuracy and F1 Score. The TWOSIDES dataset often includes multiple interaction types per drug pair; thus, we framed the analysis as a recommendation task, using Hit@5 and NDCG@5 for evaluation. This evaluation method differs from most existing works that use TWOSIDES for binary classification (results shown in Appendix B.2). We argue that binary classification tasks are often redundant and include uninformative negative samples. Our proposed ranking metrics accurately reflect realistic scenarios where a drug pair may exhibit multiple interactions, and the objective is to identify the most probable risks.

Implementation Details. We use Qwen2.5 series [35] as the backbone LLM for DDIAgents and other types of agent-based baseline methods in quantitative experiments. We set the maximum iteration round $T = 3$ and agent number $m = 3$. For DrugBank, each DDI question takes 12.83 seconds on average, and for TWOSIDES it takes 15.66 seconds—both within an acceptable range. More implementation details are shown in Appendix A. Our code and data are available at https://github.com/zgs0314/DDIAgents.

Baseline Methods. In this work, we compare our proposed DDIAgents framework with four types of baseline methods: (1) *Feature-based methods*: MLP [25]; (2) *Graph-based methods*: MSTE [43], Decagon [48], EmerGNN [46] and TIGER [30]; (3) *LLM-based methods*: Single LLM [12], TextDDI [47], DDI-GPT [41], CBR-DDI [15], K-Path [1]; and (4) *Agent-based methods*: Reflexion [29], Debate [6], AgentVerse [4], MDAgents [13].

5.2 General Performance Comparison

Table 2 summarizes the overall comparison across S1 and S2 tasks on DrugBank and TWOSIDES. We can see that our proposed DDIAgents framework consistently outperforms all baseline methods. Compared with the best-performing baseline, DDIAgents improves F1 by 13.77% in S2 tasks on DrugBank, and increases NDCG@5 by 7.15% in S2 tasks on TWOSIDES. These gains validate the core design of DDIAgents: orchestrating multiple expert agents and routing specialized knowledge sources to each expert to enable complementary and targeted reasoning.

Among baselines, feature-based and graph-based models show comparatively weak performance. A key reason is that the evaluation protocol adopts an approval-time-based drug split to better reflect real-world deployment, which substantially increases task difficulty. Under this more realistic setting, methods that rely on molecular features or limited graph structure cannot effectively leverage heterogeneous evidence, and thus struggle to capture the diverse mechanisms underlying DDIs. The performance of LLM-based methods also remains limited, primarily because they utilize only the textual descriptions, which fails to comprehensively model the multi-faceted evidence required for robust DDI prediction. In particular, the Single LLM baseline performs poorly, suggesting that memorization or knowledge leakage in current LLMs is not the primary driver of strong performance in this task.

In contrast, agent-based methods generally achieve the best performance among different types of methods. This is because their

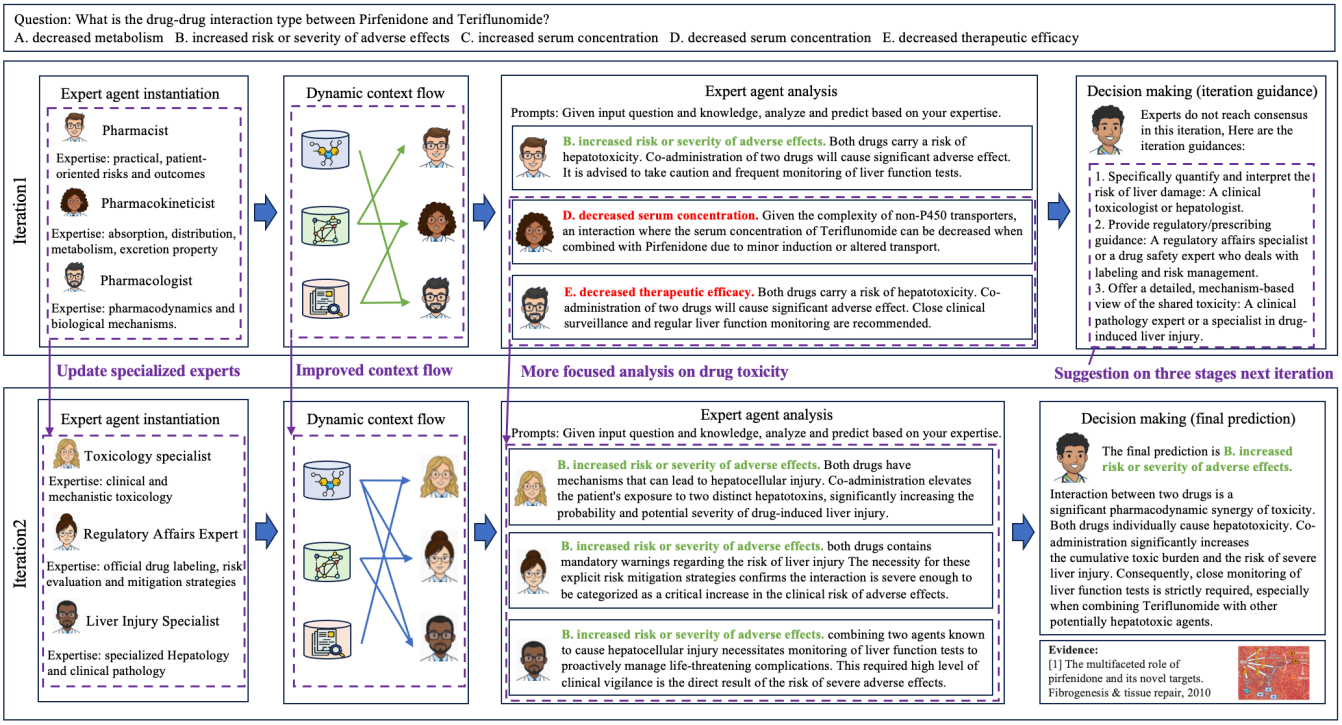


Figure 4: A prediction case of DDIAgents. Bold red text indicates incorrect answers and bold green text indicates correct answers.

strong capability in handling heterogeneous knowledge sources and the shared high-recall candidate retriever provides a reasonable starting point for the prediction (its performance is shown in Appendix B.3). Notably, methods that incorporate multiple expert agents (e.g., MDAgents and AgentVerse) consistently outperform single-agent methods (e.g., Reflexion). This comparison highlights the critical need for integrating specialized analyses from multiple complementary perspectives.

5.3 Ablation Study

To validate the effectiveness of the proposed components in DDIAgents, we compare it with the following variants: (a) w/o DynExp, which utilizes three predefined expert agents across iterations; (b) w/o DynKnow, which provides all of the knowledge sources instead of using dynamic context flow; (c) w/o Iter, which only uses a single iteration to obtain the DDI prediction.

From results in Table 3, we can see that removing any proposed component results in a performance drop, validating the necessity of the designs in DDIAgents. Specifically, the largest performance degradation is in the "w/o DynKnow" variant, highlighting the critical importance of the dynamic context flow. Furthermore, we observe that the iterative analysis component is more crucial for the TWOSIDES dataset. We hypothesize this is because the DDI prediction task on TWOSIDES requires selecting multiple DDI types from a set of candidates, leading to a higher frequency of inter-expert disagreement and, consequently, making the iterative refinement process essential for consensus and accurate prediction.

Table 3: Comparison of different variants of DDIAgents

	DrugBank				TWOSIDES			
	S1		S2		S1		S2	
	Acc	F1	Acc	F1	Hit	NDCG	Hit	NDCG
DDIAgents	55.75	45.52	38.21	18.17	53.94	21.21	42.26	14.08
w/o DynExp	54.08	44.39	36.70	16.66	51.73	20.37	41.01	13.52
w/o DynKnow	52.23	40.19	35.57	16.01	50.34	19.52	38.71	12.07
w/o Iter	53.68	42.15	36.67	16.71	50.72	19.44	39.56	12.88

5.4 Case Study

Figure 4 shows an exemplar case of DDIAgents on a DDI prediction problem on DrugBank dataset using large-scale frontier model (GPT-4o). The case study details a two-iteration process to predict the DDI type between Pirfenidone and Teriflunomide, ultimately identifying the correct answer.

In the first iteration, the DDIAgents framework utilizes predefined pharmacist, pharmacokineticist, and pharmacologist as expert agents to analyze the question, but they failed to reach a consensus. In this case, the conclusion agent identifies conflicting evidence and generates targeted iteration guidance. Its feedback calls for instantiating toxicology and regulatory experts, identifies regulatory labeling as a key source for dynamic context flow, and prioritizes liver toxicity as the main focus for subsequent expert analysis.

Acting on this directive, expert agent instantiation in the second iteration gathers toxicology specialist, regulatory affairs expert, and liver injury specialist as updated experts better suited to the problem. The dynamic context flow then prioritizes official drug

labeling and mechanistic toxicology evidence to fill the identified gaps. With these refined inputs, expert analysis converges on a unified understanding of the mechanisms underlying hepatocellular injury. Finally, the decision-making stage confirms the prediction of "increased risk or severity of adverse effects", supported by an explanation of synergistic toxic burden and the clinical need to monitor liver function tests. To verify the reliability, we performed a secondary search of relevant literature, confirming that output analysis is well-supported by established clinical evidence [17]. More DDI prediction cases using Qwen 2.5 7B are shown in Appendix B.6.

5.5 Performance across DDI Type Frequencies

In this section, we evaluate the robustness of DDIAgents framework across various DDI types using the DrugBank dataset. Representative methods from different categories are selected for comparison, including MLP, TIGER, DDI-GPT, Reflexion, MDAgents and DDIAgents. Given the inherent class imbalance, we employ a frequency-based ranking to categorize DDI types into five quintile bins based on frequency. The "0%–20%" bin contains the most frequent DDI types, while the "80%–100%" bin represents the long-tail DDI types.

Figure 5 illustrates the average prediction accuracy across DDI type bins on S1 tasks. The results demonstrate that DDIAgents consistently outperforms all baseline methods across every frequency interval. Notably, the performance advantage of DDIAgents becomes increasingly significant on long-tail DDI types. This trend suggests that while traditional models depend on high-frequency patterns, DDIAgents better handles rare, data-sparse interactions through collaborative agents and dynamic context flow.

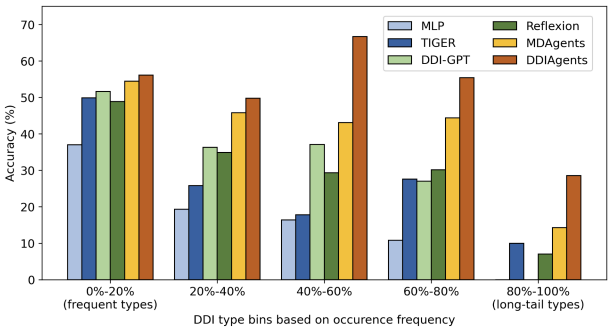
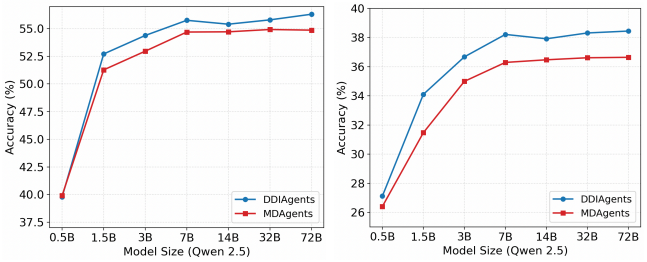


Figure 5: DDI prediction performance across DDI type frequency bins on DrugBank S1 tasks.

5.6 Performance across Different LLM Models

To rigorously validate the generalizability and inherent robustness, we conducted an extensive evaluation on DDIAgents using various LLMs as agents and compared them with the strongest baseline, MDAgents. Our initial set of experiments focused on assessing the influence of model scale within a unified family, utilizing the Qwen 2.5 series with sizes ranging from 0.5B to 72B parameters. The results in Figure 6 show that DDIAgents outperforms MDAgents at most tested model sizes, and that both frameworks improve steadily as model scale increases. Notably, the smaller models (0.5B and 1.5B) were unable to achieve satisfactory prediction performance for both methods, underscoring the requirement for a sufficiently capable LLM base to execute the complex tasks within the multi-agent system. Furthermore, DDIAgents achieves higher performance as model size increased, which suggests a greater capacity to stimulate the emergent capabilities of large-scale LLMs.



(a) S1 tasks **(b) S2 tasks**

Figure 6: The performance of DDIAgents and MDAgents with different LLM model sizes on the DrugBank dataset.

To evaluate the cross-architecture generalizability, we compared the performance of DDIAgents and MDAgents across LLMs of comparable scales, including Qwen 2.5 7B, Llama 3.1 8B [8], Gemma 2 9B [34] and ChatGLM 3 6B [7]. The results in Table 4 indicate that DDIAgents consistently outperforms MDAgents, demonstrating superior robustness and reasoning capabilities. Notably, while both frameworks perform well with Qwen 2.5, Llama 3.1 and Gemma 2, DDIAgents maintains a substantial lead in F1 scores, which indicates its enhanced capacity for analyzing long-tail DDI types. The relatively low performance of ChatGLM 3 may stem from its training that is less optimized for complex biomedical reasoning.

Table 4: The performance of DDIAgents and MDAgents using LLM models with similar sizes on DrugBank dataset.

LLM models	Method	DrugBank			
		S1		S2	
		Acc	F1	Acc	F1
Qwen 2.5 7B	DDIAgents	55.75	45.52	38.21	18.17
	MDAgents	54.69	39.52	36.29	15.97
Llama 3.1 8B	DDIAgents	55.63	46.56	37.21	17.77
	MDAgents	53.97	40.72	36.32	16.24
Gemma 2 9B	DDIAgents	55.30	44.19	38.34	18.39
	MDAgents	54.36	40.41	35.72	16.04
ChatGLM 3 6B	DDIAgents	53.71	42.60	35.98	16.44
	MDAgents	52.84	38.11	34.78	15.19

5.7 Comparison on Generated Contents

To evaluate the quality of the generated explanations, we conducted a comparative analysis between DDIAgents and MDAgents. We employed a state-of-the-art LLM (GPT-4o) as a judge to evaluate their responses on S2 tasks across four domain-specific dimensions [10]:

- Thoroughness: how comprehensively the system addresses the multi-faceted nature of DDIs;
- Diversity: the breadth of explored biological mechanisms;
- Rationality: whether the answer provides reasonable analysis;
- Overall: a holistic judgment of which system provides a more useful and trustworthy reference for drug interaction analysis.

As shown in Table 5, DDIAgents consistently outperforms MDAgents across all dimensions. The expert agent instantiation phase enhances pharmacological thoroughness and mechanistic diversity by integrating specialized domain expertise into the reasoning process. Simultaneously, the dynamic context flow and iterative feedback loops refine agent outputs, significantly boosting clinical rationality. By replacing static discussions with active collaboration, DDIAgents captures complex biochemical interactions more accurately, providing more reliable and comprehensive DDI explanations than generic multi-agent frameworks.

Table 5: Win rates (%) for generated contents of DDIAgents vs. MDAgents across four dimensions on S2 tasks.

	DrugBank		TWOSIDES	
	DDIAgents	MDAgents	DDIAgents	MDAgents
Thoroughness	66.39%	33.61%	72.27%	27.73%
Diversity	67.38%	32.62%	67.33%	32.67%
Rationality	75.02%	24.98%	70.87%	29.13%
Overall	69.08%	30.92%	71.34%	28.66%

6 Conclusion

In this work, we presented DDIAgents, a mechanism-conditioned multi-agent framework for DDI prediction. Rather than treating DDI analysis as static feature fusion or historical case reuse, DDIAgents frames the task as knowledge orchestration: for each drug pair, dynamic context flow routes heterogeneous biomedical evidence to specialized agents according to mechanism-specific reasoning needs. Extensive experiments demonstrate that DDIAgents consistently outperforms state-of-the-art baselines. Furthermore, DDIAgents provides interpretable agent-level mechanistic insights into predicted interactions, offering a useful direction for adaptive and evidence-grounded AI4Science reasoning.

Despite these advancements, the current implementation relies on a fixed agent collaboration scheme, which may restrict the system's adaptability. Future work will develop a more flexible collaboration scheme with multiple interaction patterns based on the specific context. Additionally, we plan to incorporate more critical knowledge sources to expand the dynamic context flow and further enhance predictive accuracy and robustness.

7 Limitations and Ethical Considerations

Limitations. The current DDIAgents adopts a fixed agent collaboration scheme with predefined interaction protocols. While this simplifies orchestration, it may reduce adaptability when different DDI queries would benefit from alternative coordination patterns. In addition, DDIAgents currently relies on a limited set of knowledge-source types in the dynamic context flow, which may omit clinically decisive evidence.

Ethical considerations. Our experiments use publicly available drug resources rather than patient-identifiable records. Nevertheless, any clinical deployment should ensure proper consent, governance and strong security for data access and storage. DDIAgents is intended for decision support but not final decisions. The outputs should be verified with reliable evidence to mitigate harms from hallucinations or misuse.

GenAI Disclosure

This work uses Generative AI. The LLM-based baselines and agent-based methods rely on large language models to generate predictions and explanations for DDI queries. The case studies include representative model-generated outputs to illustrate the agent workflows. All such content is produced at inference time for research evaluation and demonstration, and may contain inaccuracies as discussed in the Limitations and Ethical Considerations section.

Acknowledgment

Q. Yao is supported Beijing Natural Science Foundation (No.4242039) and Beijing Science and Technology Program (No.Z251100008125003).

References

[1] Tassallah Abdullahi et al. 2025. K-paths: reasoning over graph paths for drug repurposing and drug interaction prediction. arXiv preprint arXiv:2502.13344 (2025).
[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023).
[3] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. Nature 624, 7992 (2023), 570–578.
[4] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. 2023. Agentverse: facilitating multi-agent collaboration and exploring emergent behaviors. In International Conference on Learning Representations.
[5] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. 2016. Innovation in the pharmaceutical industry: new estimates of R&D costs. Journal of Health Economics 47 (2016), 20–33.
[6] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In International Conference on Machine Learning.
[7] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling. In Association for Computational Linguistics. 320–335.
[8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024).
[9] Alireza Ghafarollahi and Markus J Buehler. 2025. SciAgents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. Advanced Materials 37, 22 (2025), 2413523.
[10] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. LightRAG: simple and fast retrieval-augmented generation. arXiv preprint arXiv:2410.05779 (2024).
[11] Ke Han, Peigang Cao, Yu Wang, Fang Xie, Jiaqi Ma, Mengyao Yu, Jianchun Wang, Yaoqun Xu, Yu Zhang, and Jie Wan. 2022. A review of approaches for predicting drug–drug interactions based on machine learning. Frontiers in Pharmacology 12 (2022), 814858.
[12] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. 2023. MetaGPT: meta programming for a multi-agent collaborative framework. In International Conference on Learning Representations.
[13] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. MDagents: an adaptive collaboration of LLMs for medical decision-making. Advances in Neural Information Processing Systems 37 (2024), 79410–79452.
[14] Xuan Lin, Lichang Dai, Yafang Zhou, Zu-Guo Yu, Wen Zhang, Jian-Yu Shi, Dong-Sheng Cao, Li Zeng, Haowen Chen, Bosheng Song, et al. 2023. Comprehensive evaluation of deep and graph learning on drug–drug interactions prediction. Briefings in Bioinformatics 24, 4 (2023), bbad235.
[15] Guangyi Liu, Yongqi Zhang, Xunyuan Liu, and Quanming Yao. 2025. Case-based reasoning enhances the predictive power of LLMs in drug-drug interaction. arXiv preprint arXiv:2505.23034 (2025).
[16] Zun Liu et al. 2022. Predict multi-type drug–drug interactions in cold start scenario. BMC Bioinformatics (2022), 75.
[17] José Macías-Barragán, Ana Sandoval-Rodríguez, Jose Navarro-Partida, and Juan Armendáriz-Borunda. 2010. The multifaceted role of pirfenidone and its novel targets. Fibrogenesis & Tissue Repair 3, 1 (2010), 16.
[18] Jin Niu, Robert M Straubinger, and Donald E Mager. 2019. Pharmacodynamic drug–drug interactions. Clinical Pharmacology & Therapeutics 105, 6 (2019),

1395–1406.

[19] Arnold K Nyamabo, Hui Yu, and Jian-Yu Shi. 2021. SSI–DDI: substructure–substructure interactions for drug–drug interaction prediction. Briefings in Bioinformatics 22, 6 (2021), bbab133.

[20] Caterina Palleria, Antonello Di Paolo, Chiara Giofrè, Chiara Caglioti, Giacomo Leuzzi, Antonio Siniscalchi, Giovambattista De Sarro, and Luca Gallelli. 2013. Pharmacokinetic drug–drug interaction and their implication in clinical management. Journal of Research in Medical Sciences 18, 7 (2013), 601.

[21] Vinay Prasad, Kevin De Jesús, and Sham Mailankody. 2017. The high price of anticancer drugs: origins, implications, barriers, solutions. Nature Reviews Clinical Oncology 14, 6 (2017), 381–390.

[22] Yang Qiu, Yang Zhang, Yifan Deng, Shichao Liu, and Wen Zhang. 2021. A comprehensive review of computational methods for drug-drug interaction detection. IEEE/ACM Transactions on Computational Biology and Bioinformatics 19, 4 (2021), 1968–1985.

[23] Kithusshand Raveendran, Deshan Sumanathilaka, Kanagasabai Thiruthanigesan, Pavinithan Retnakumar, and Yathukulan Sivapiragasam. 2025. PharmaMap-LLM: Fine-Tuning Large Language Models for Drug-Drug Interaction (DDI) Analysis. In International Conference on Advancements in Computing (ICAC). 1–6.

[24] Arthur G Roberts and Morgan E Gibbs. 2018. Mechanisms and the clinical relevance of complex drug–drug interactions. Clinical Pharmacology: Advances and Applications (2018), 123–134.

[25] David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. Journal of Chemical Information and Modeling 50 (2010), 742–754.

[26] Jae Yong Ryu et al. 2018. Deep learning improves prediction of drug–drug and drug–food interactions. Proceedings of the National Academy of Sciences 115 (2018), E4304–E4311.

[27] Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. 2024. MoME: mixture of multimodal experts for generalist multimodal large language models. Advances in Neural Information Processing Systems 37 (2024), 42048–42070.

[28] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In International Conference on Machine Learning. 31210–31227.

[29] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems 36 (2023), 8634–8652.

[30] Xiaorui Su et al. 2024. Dual-channel learning framework for drug-drug interaction prediction via relation-aware heterogeneous graph transformer. In AAAI Conference on Artificial Intelligence. 249–256.

[31] Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2025. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. Nature (2025), 1–3.

[32] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: large language models as collaborators for zero-shot medical reasoning. In Findings of the Association for Computational Linguistics. 599–621.

[33] Nicholas P Tatonetti et al. 2012. Data-driven prediction of drug effects and interactions. Science Translational Medicine (2012).

[34] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: improving open language models at a practical size. arXiv preprint arXiv:2408.00118 (2024).

[35] Qwen Team. 2024. Qwen2.5: a party of foundation models. https://qwenlm.github.io/blog/qwen2.5/

[36] Yaqing Wang, Zhenlin Luo, Peiyao Zhao, Yunfeng Cai, and Quanming Yao. 2026. Beyond scaling: A survey on data-efficient agentic learning. In International Joint Conference on Artificial Intelligence.

[37] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. Comput. Surveys 53, 3, Article 63 (2020), 34 pages. doi:10.1145/3386252

[38] Rick A Weideman, Ira H Bernstein, and W Paul McKinney. 1999. Pharmacist recognition of potential drug interactions. American Journal of Health-System Pharmacy 56, 15 (1999), 1524–1529.

[39] David S Wishart et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research 46 (2018), 1074–1082.

[40] Lingxuan Xie, Tengfei Ma, Yuqin He, Yiping Liu, and Xiangxiang Zeng. 2025. Predicting Drug–Drug Interaction via Dual-Drug Visual Representation. Journal of Chemical Information and Modeling 65, 19 (2025), 9999–10010.

[41] Chengqi Xu et al. 2024. DDI-GPT: explainable prediction of drug-drug interactions using large language models enhanced with knowledge graphs. bioRxiv (2024), 2024–12.

[42] 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 (2022), 8693–8703.

[43] Junfeng Yao et al. 2022. Effective knowledge graph embeddings based on multidirectional semantics relations for polypharmacy side effects prediction. Bioinformatics 38 (2022), 2315–2322.

[44] Jingjing Yu, Ichiko D Petrie, René H Levy, and Isabelle Ragueneau-Majlessi. 2019. Mechanisms and clinical significance of pharmacokinetic-based drug-drug interactions with drugs approved by the US Food and Drug Administration in 2017. Drug Metabolism and Disposition 47, 2 (2019), 135–144.

[45] Yue Yu et al. 2021. SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization. Bioinformatics 37 (2021), 2988–2995.

[46] Yongqi Zhang et al. 2023. Emerging drug interaction prediction enabled by a flow-based graph neural network with biomedical network. Nature Computational Science 3 (2023), 1023–1033.

[47] Fangqi Zhu et al. 2023. Learning to describe for predicting zero-shot drug-drug interactions. In Conference on Empirical Methods in Natural Language Processing.

[48] Marinka Zitnik et al. 2018. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 34 (2018), 457–466.

A Supplementary Information

A.1 Description of DDI Examples in Motivation Part

In Table 6, we provide textual descriptions of diverse DDI mechanisms illustrated in Figure 1. Here we provide more comprehensive explanations on DDI mechanisms and details of reference knowledge used in exemplar cases.

Table 6: Three DDI examples with different mechanisms.

DDI examples
Drug 1: Roxatidine acetate; **Drug 2:** Risedronic acid; **Interaction:** Increased serum concentration and adverse side effect. **Mechanism:** Roxatidine acetate raises the gastric pH, which unexpectedly increases the absorption of Risedronic acid, leading to a higher serum concentration and a greater risk of adverse effects. **Reference knowledge:** Drug usage instruction (the administration instructions for Risedronic acid)
Drug 1: Esomeprazole; **Drug 2:** Clopidogrel; **Interaction:** Decreased antiplatelet activity. **Mechanism:** Esomeprazole, by inhibiting the liver enzyme CYP2C19, prevents the necessary metabolic conversion of the inactive Clopidogrel, consequently decreasing its antiplatelet activity. **Reference knowledge:** Biomedical networks (the drug-target interaction between esomeprazole and enzyme CYP2C19)
Drug 1: Erythrityl tetranitrate; **Drug 2:** Sildenafil; **Interaction:** Increased vasodilatory activity. **Mechanism:** Erythrityl tetranitrate increases the concentration of the vasodilator cGMP by donating NO, while Sildenafil prevents its breakdown by inhibiting PDE5, leading to a dangerous and synergistic accumulation of cGMP that causes severe hypotension. **Reference knowledge:** Drug molecular structure (Nitrate functional group in Erythrityl tetranitrate)

A.2 The Classification of Knowledge Sources

As summarized in Table 7, the DDI prediction is enhanced by introducing several types of knowledge sources. For both DrugBank and TWOSIDES, this includes drug SMILES, drug-target interactions, and drug-side effect information. A critical component of the knowledge source is the pharmacological categorization of known DDIs. Specifically, DDIs in DrugBank are divided into six distinct classes based on their pharmacokinetic and pharmacodynamic mechanisms, while DDIs in TWOSIDES are categorized into nine medical system-based classes.

A.3 Implementation Details of Knowledge Retrieval by Expert Agents

Before expert agent analysis, the knowledge context is obtained through a retrieval process as follows:

- A textual encoder transforms the DDI question and all knowledge entries within the assigned sources into embeddings.
- Relevance scores are determined via cosine similarity between the query embedding and each entry embedding.
- We perform a top-k retrieval operation independently for each type of knowledge source assigned to the agent. This prevents a single dominant source from overshadowing other relevant but less frequent data types. The final context is the union of these retrieved subsets.

A.4 Algorithm of DDIAgents

The algorithm of DDIAgents is shown in Algorithm 1, which includes the iteration of three critical stages: Expert agent instantiation, Dynamic context flow, Analysis and Decision making.

Algorithm 1 DDIAgents for DDI Prediction through Multi-Agent System

Require: DDI Question, Knowledge sources and Knowledge textual encoder

$t \leftarrow 1$
while True **do**
  // Stage (a): Expert agent instantiation
  if $t = 1$ **then**
    Instantiate expert agents as predefined domain experts (Pharmacist, Pharmacokineticist and Pharmacologist)
  else
    Expert agent instantiation based on feedback of conclusion agent in $t - 1$ iteration
  end if
  // Stage (b): Dynamic context flow
  for each expert agent **do**
    Evaluate knowledge source assigned for the expert agent
  end for
  // Stage (c): Analysis and decision making
  Identify allocated data source types from dynamic context flow
  For each expert, retrieve top-k knowledge entries based on semantic similarity through knowledge textual encoder
  for each expert agent **do**
    Generate specific expert analysis
  end for
  Synthesize findings from all agents
  if O_C^t requires refinement **then**
    Generate iteration guidance
    $t \leftarrow t + 1$
  else
    break
  end if
end while
return final answer and explanation

A.5 Statistics of Datasets

We present the general statistics of datasets in experiments in Table 8. Here $\mathcal{D}$ represents the set of drugs, $\mathcal{R}$ denotes the set of interactions and $\mathcal{N}$ refers to the set of DDIs.

Table 7: Knowledge source categorization for different datasets.

Dataset	Knowledge sources	Explanation
DrugBank	Drug SMILES	String representations of drug chemical structures.
	Drug target interactions	Binding relationships between drug molecules and their specific biological targets.
	Drug side effects	Adverse or unintended physiological responses observed after the administration of drugs.
	Drug description	Concise textual summary providing essential details about a medication's purpose, chemical composition, and safety profile.
	DDI description	A brief introduction of drug-drug interaction types provided as choices in DDI prediction question.
	Absorption DDIs	Interactions that alter the rate or extent of drug entry into the bloodstream.
	Distribution DDIs	Interactions affecting the movement of drugs between the blood and various body tissues.
	Metabolic DDIs	Interactions influencing the biochemical transformation of drugs.
	Excretion DDIs	Interactions impacting the processes through which drugs and metabolites are removed from the body.
	Synergistic DDIs	Interactions where the combined drug effect is greater than the sum of individual effects.
	Antagonistic DDIs	Interactions where one drug reduces or counteracts the pharmacological effect of another.
TWOSIDES	Drug SMILES	String representations of drug chemical structures.
	Drug description	Concise textual summary providing essential details about a medication's purpose, chemical composition, and safety profile.
	DDI description	A brief introduction of drug-drug interaction types provided as choices in DDI prediction question.
	Drug target interactions	Binding relationships between drug molecules and their specific biological targets.
	Drug side effects	Adverse or unintended physiological responses observed after the administration of drugs.
	Nervous System and Psychiatric Disorders	Side effects affecting brain function, neurological processes, or mental health.
	Cardiovascular and Circulatory Disorders	Adverse interaction effects impacting the heart, blood vessels, and systemic circulation.
	Respiratory Disorders	Side effects affecting the lungs and breathing processes caused by interacting drug pairs.
	Gastrointestinal and Hepatic Disorders	Drug interaction effects related to the digestive system and liver function.
	Renal and Metabolic Disorders	Adverse effects involving kidney function or systemic biochemical processes and regulation.
	Immune, Hematologic, and Endocrine Disorders	Interactions affecting the immune system, blood components, or hormone-producing glands.
	Musculoskeletal and Connective Tissue Disorders	Side effects involving muscles, bones, joints, and associated supportive tissues.
	Dermatologic and Sensory Disorders	Adverse reactions impacting the skin, hair, nails, or sensory organs.
	Neoplastic and Systemic Conditions	Interactions related to tumor growth or generalized, whole-body physiological complications.

Table 8: Statistics of DDI prediction datasets in experiments.

Dataset	DrugBank	TWOSIDES
$\lvert\mathcal{D}_k\rvert$ (Known drug set)	1020	292
$\lvert\mathcal{D}_n\rvert$ (New drug set)	129	38
$\lvert\mathcal{R}\rvert$ (Relation types)	86	209
$\lvert\mathcal{N}_{\text{train}}\rvert$ (Train DDIs)	83272	11300
$\lvert\mathcal{N}_{\text{test S1}}\rvert$ (Test DDIs for S1 tasks)	20031	3203
$\lvert\mathcal{N}_{\text{test S2}}\rvert$ (Test DDIs for S2 tasks)	1298	194

A.6 Evaluation Metrics

According to the evaluation metrics mentioned in Section 5.1, the evaluation metrics include F1-Score, accuracy for DrugBank:

- F1-Score (Macro) = $\frac{1}{||\mathcal{P}_D||}\sum_{p\in\mathcal{P}_D}\frac{2P_p\cdot R_p}{P_p+R_p}$, where P_p and R_p are the precision and recall for the interaction type p, respectively.
- Accuracy: the proportion of correctly predicted interaction types relative to the ground-truth interaction types.

For TWOSIDES dataset, we use Hit@5 and NDCG@5 as evaluation metrics. Denote $\mathcal{I}_{1:5}$ as the recommended interaction for the given query drug pairs, $\mathcal{T}$ as the ground-truth interaction list and $\mathbb{I}$ as the indicator function. These metrics can be represented as follows:

- Hit@5 = $\mathbb{I}(|\mathcal{I}_{1:5}\cap\mathcal{T}|)$.
- NDCG@5 = $\frac{\sum_{i=1}^{5}\mathbb{I}(\mathcal{I}_i\in T)/\log_2(i+1)}{\sum_{i=1}^{\min(|T|,5)}1/\log_2(i+1)}$.

A.7 Baseline Methods

Here we present a more detailed introduction of baseline methods compared in experiments:

- MLP [25] is a classic supervised learning approach that utilizes a multi-layer perceptron to map drug chemical fingerprints to specific interaction types between drug pairs.
- MSTE [43] specially designs a knowledge graph embedding scoring function to model complex relations in DDI prediction problems.
- Decagon [48] is a representative GNN-based approach that performs multi-relational link prediction by extracting and aggregating information from biomedical networks.
- EmerGNN [46] is a specialized GNN that utilizes a flow-based architecture to identify critical paths within biomedical networks, specifically optimized for predicting interactions involving emerging drugs.

Table 9: System prompt for planner in expert agent instantiation.

```
You are a medical expert who specializes in categorizing a specific drug-drug interaction prediction problem into specific areas of medicine.
You need to complete the following steps:
1. Carefully read the drug-drug interaction prediction problem with candidate choices and feedbacks from last iteration round.
2. Based on the problem, provide three experts that are suitable to answer the question from different perspectives.
3. Provide three expert names and use one sentence to describe their roles, respectively.
You should output in exactly the same format as: (1) [expert1 name]: [one sentence describes the role of expert1]. (2) [expert2 name]: [one sentence describes the role of expert2]. (3) [expert3 name]: [one sentence describes the role of expert3].
Question: {{Question}}
Candidate interaction types: {{Candidate DDI types}}
Analysis in last round: {{Feedback from conclusion agent}}
Please provide your split of three experts:
```

Table 10: System prompt for planner in dynamic context flow (for DrugBank dataset).

```
You are a medical librarian who specializes in managing medical knowledge sources for medical experts.
Given the medical expert name, his expertise, drug-drug interaction prediction problem and knowledge sources, please select the relevant information needed.
Expert name: {{Expert name}}. Expertise: {{Expertise}}.
Question: {{Question}}
Candidate interaction types: {{Candidate DDI types}}
Knowledge sources:
A. Absorption DDIs: an absorption DDI occurs when one drug alters the rate or extent of absorption and bioavailability of a second drug, leading to a decrease in its concentration and potential loss of efficacy or an increase in its concentration and potential worsening of adverse effects.
B. Distribution DDIs: a distribution DDI occurs when one drug alters the distribution of a second drug throughout the body, typically by affecting plasma protein binding or transport, leading to changes in the free serum concentration of Drug2, which may result in altered efficacy or toxicity.
C. Metabolic DDIs: a Metabolic DDI occurs when one drug alters the rate of metabolism of a second drug, typically via enzyme inhibition or induction, leading to either a change in Drug2's serum concentration or an alteration in the concentration of its active or inactive metabolites.
D. Excretion DDIs: An Excretion DDI occurs when one drug alters the rate of renal or biliary excretion of a second drug, leading to either an increase in the elimination rate and potential loss of Drug2 efficacy or a decrease in the elimination rate and potential Drug2 toxicity due to elevated serum levels.
E. Synergistic DDIs: a synergistic DDI is an interaction where two drugs combined produce an effect (therapeutic or adverse) that is greater than the sum of the effects of each drug given alone, typically resulting in an increase in the intensity of specific pharmacodynamic activities, such as toxicity or efficacy.
F. Antagonistic DDIs: an antagonistic DDI is an interaction where one drug opposes the effect of a second drug by competing for the same receptor or having counteracting physiological effects, resulting in a decrease in the therapeutic efficacy or an attenuation of the adverse effects of Drug2.
G. SMILES of two drugs.
```

Table 11: System prompt for expert agent analysis.

```
You are a {{Expert name}} who specializes in {{Agent expertise}}.
You are provided with:
- Two drugs and their descriptions.
- Candidate interaction types predicted by a deep learning model, each with a short description.
- Relevant knowledge sources as a reference.
- Analysis summary of different agents in the last round.
Note:
- You MUST give your answer in English, MUST not contain characters in other language.
- The interaction could be from Drug 1 to Drug 2, or from Drug 2 to Drug 1.
Please choose the most plausible interaction type based on the provided information and your medical knowledge.
Input format: Question: What is the drug-drug interaction type between [Drug 1] and [Drug 2]?
[Drug 1]: [Description of Drug 1]; [Drug 2]: [Description of Drug 2].
Candidate interaction types:
- A. [Interaction Type A]: [Description of Interaction Type A]
- B. [Interaction Type B]: [Description of Interaction Type B]
...
Reference knowledge sources: ...
The answer of other agents in the last round: ...
Output Format:
Interaction type A/B/C/D/E.
[Your analysis to explain your choice. ]
Now answer this question: ...
```

- Single LLM is a straightforward method that directly uses LLMs to answer textual DDI prediction questions without adding any supplementary information. In our experiments, we implement this baseline using Qwen2.5-7B as the backbone model.
- TIGER [30] is a hybrid architecture featuring a dual-channel graph transformer, which simultaneously captures the internal structural information of drug molecular graphs and the global topological features of biomedical networks.
- TextDDI [47] is the first framework to leverage the zero-shot capabilities of LLMs for DDI prediction, utilizing textual drug descriptions and interaction histories.

Table 12: System prompt for decision making.

```
You are a medical expert specializing in pharmacology and drug safety.
Your task is to predict the most likely drug-drug interaction (DDI) between a given pair of drugs.
You are provided with the prediction and information from multiple experts.
Please take their answer into comprehensive consideration and judge whether the evidence provided is enough to obtain a conclusion.
You are provided with:
- Two drugs and their descriptions.
- Candidate interaction types predicted by a deep learning model, each with a short description.
- The answer of other agents for reference.
Note:
- You MUST give your answer in English, MUST not contain characters in other language.
- If there are contradictions among the answer of different agents, it is encouraged to require a new round of analysis for different experts.
Input format: Question: What is the drug-drug interaction type between [Drug 1] and [Drug 2]?
[Drug 1]: [Description of Drug 1]; [Drug 2]: [Description of Drug 2].
Candidate interaction types:
- A. [Interaction Type A]: [Description of Interaction Type A]
- B. [Interaction Type B]: [Description of Interaction Type B]
...
The answer of other agents: ...
Output must strictly follow the format below.
If you think the evidence provided is enough, the output format MUST be:
Yes. Interaction type A/B/C/D/E.
[Your analysis to explain your choice. ]
If you think the evidence provided is not enough, the output format MUST be:
No. [One sentence to summarize the answer of each expert agent, respectively.]
[Suggestion for further analysis on the question. ]
Now answer this question: ...
```

- DDI-GPT [41] captures relevant information of query drugs from biomedical networks and uses biomedical large language model to enhance DDI prediction performance.
- CBR-DDI [15] adapts the "Case-Based Reasoning" paradigm to the LLM setting, enabling the LLM model to retrieve and learn from past DDI instances to solve new DDI queries.
- K-path [1] extracts critical paths from biomedical networks as important background to enhance the LLM prediction performance on DDI prediction problem.
- Reflexion [29] is a reinforcement learning-inspired architecture that enables single agent to improve its performance through self-reflection, where it critiques past mistakes and stores them in a verbal memory to inform more accurate future attempts.
- Debate [6] is a multi-agent collaboration protocol where specialized agents iteratively critique and refine each other's responses until a consensus is reached.
- AgentVerse [4] is a dynamic framework that adjusts the composition of an agent group based on the specific progress and complexity of the problem-solving task.
- MDAgents [13] is a multi-agent system specifically tailored for medical decision-making, employing adaptive protocols that scale in complexity based on the perceived difficulty of the clinical case.

A.8 Prompt used in Experiments

In this section, we provide the prompt templates of different agents in DDIAgents in Table 9-Table 12.

B More Experimental Results

B.1 Experimental Results on S0 Setting

Table 16 details the performance of various methods within the S0 tasks, which focuses on predicting interactions between known drugs. While the proposed DDIAgents maintain strong performance, the performance gap between models is less pronounced in S0 than in S1 and S2 tasks. This suggests that traditional methods can effectively leverage the rich interaction history of known drugs, making the prediction task less challenging than in zero-shot or cold-start scenarios.

B.2 Experimental Results on TWOSIDES Dataset with Traditional Evaluation Setting

To provide a comprehensive comparison with existing literature, we report the performance of DDIAgents on the TWOSIDES dataset using the traditional binary classification setting. While we maintain that the ranking-based metrics (Hit@5 and NDCG@5) introduced in Section 5.1 better reflect clinical realities by addressing the multi-label nature of drug-drug interactions, these supplementary results in Table 13 serve as a benchmark against standard methodologies. Here we use the following two evaluation metrics:

- ROC-AUC = $\sum_{k=1}^{n} TP_k \Delta FP_k$ measures the area curve of receiver operating characteristics.
- Accuracy: the proportion of correctly predicted DDIs for each DDI type.

As shown, DDIAgents consistently outperforms all baseline methods across both S1 and S2 tasks, demonstrating its robustness when evaluated under traditional binary classification setting.

Table 13: Performance of representative methods on TWOSIDES with traditional setting. The best and second-best results are highlighted in bold and underline, respectively.

Method	TWOSIDES			
	S1		S2	
	ROC-AUC	Acc	ROC-AUC	Acc
MLP	74.72±1.39	66.54±0.93	54.82±1.72	51.64±2.29
TIGER	81.59±0.82	72.48±1.36	61.48±0.91	52.77±0.55
DDI-GPT	<u>86.50±0.89</u>	78.25±0.40	<u>64.57±1.53</u>	<u>61.26±1.27</u>
Reflexion	83.07±0.61	75.41±0.49	60.71±0.88	56.44±1.21
MDAgents	85.16±0.69	<u>78.39±0.71</u>	63.08±1.98	60.74±1.49
DDIAgents	**87.36±0.41**	**80.09±0.59**	**66.41±1.02**	**62.77±0.78**

B.3 Candidate Retrieval Performance of the Traditional Model

To enable efficient agent-based DDI prediction, we first train a traditional model as a candidate retriever that filters a small set of highly probable interaction types for each drug pair. Specifically, we keep the top-5 candidates for DrugBank and the top-20 candidates for TWOSIDES, aiming to maintain a high upper bound on recall; otherwise, the downstream agents would be fundamentally limited by missed candidates.

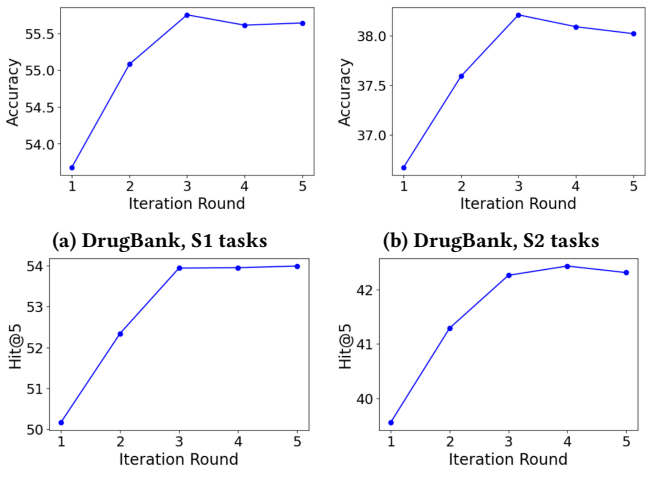
Table 14 reports the retrieval statistics of this traditional model. Overall, the model can successfully narrow the DDI candidate choice while retaining most true interaction types among the retrieved candidates. However, its prediction quality remains limited, as reflected by substantially lower accuracy/Hit@5 rates. For instance, on DrugBank, the retriever achieves high Recall@5 (93.92% in S1 and 89.46% in S2), yet its accuracy is only 50.68% (S1) and 33.47% (S2), indicating that high-recall filtering alone is insufficient for accurate DDI prediction. These results support our design choice: the traditional model serves to maximize recall and reduce the candidate space, while the performance gains of multi-agent systems mainly come from the subsequent agent-based analysis that improves precision through mechanism-aware reasoning.

Table 14: Candidate retrieval performance of the traditional model used for agent-based DDI prediction. Here "(agt)" refers to the result of DDIAgents.

DrugBank					
S1			S2		
Acc	Acc(agt)	Recall@5	Acc	Acc(agt)	Recall@5
50.68	55.75	93.92	33.47	38.21	89.46
TWOSIDES					
S1			S2		
Hit@5	Hit@5(agt)	Hit@20	Hit@5	Hit@5(agt)	Hit@20
43.90	53.94	78.23	33.51	42.26	68.56

B.4 Ablation Study on Iteration Round of DDIAgents

Here we present the results varying the iterative analysis round t in Figure 7 to assess its impact on model efficacy. For initial analysis rounds ($t \leq 3$), the performance of DDIAgents consistently improves as t increases, a gradual enhancement that strongly underscores the necessity of the iterative analysis process for effective DDI prediction. For analysis rounds extending beyond $t = 3$, performance across both datasets stabilizes, indicating that the model achieves peak efficacy within a few iterations.



(a) DrugBank, S1 tasks (b) DrugBank, S2 tasks

(c) TWOSIDES, S1 tasks (d) TWOSIDES, S2 tasks

Figure 7: Varying the iteration round of DDIAgents.

Table 15: The statistics of newly instantiated agents.

Expert name	Total number	Accuracy
Cardiologist	142	40.14
Neurologist	71	36.62
Toxicologist	55	41.82
Pulmonologist	53	33.96
Endocrinologist	52	44.23
Hepatologist	44	38.64
Oncologist	41	39.02
Gastroenterologist	39	43.59
Nephrologist	36	36.11
Pharmacogeneticist	31	32.26

B.5 Prediction Performance Using Large-scale Frontier Model

Table 17 reports the S2 results of DDIAgents when replacing the backbone LLM with a large-scale frontier model (GPT-4o). On DrugBank dataset, GPT-4o yields a clear gain in F1 score (18.17→20.75), suggesting that a stronger model particularly improves performance on the harder and less frequent DDI types captured by F1 score. On TWOSIDES, GPT-4o brings consistent improvements on both ranking metrics, which further highlights the prediction potential of DDIAgents under S2 tasks.

B.6 More DDI prediction cases

In this section, we provide additional DDI prediction case studies for DDIAgents using Qwen 2.5 7B as the backbone LLM. Table 18 reports the same drug pair as the case illustrated in Figure 4 (Section 5.4). While DDIAgents with GPT-4o correctly predicts the ground-truth interaction in Section 5.4, the Qwen 2.5 7B backbone produces an incorrect prediction after a single iteration. This comparison suggests that stronger backbone reasoning can translate into improved overall performance for DDIAgents. In addition, Table 19 presents a case where DDIAgents with Qwen 2.5 7B arrives at the correct prediction after two iterations, highlighting the benefit of iterative refinement even with a smaller backbone model.

B.7 The Statistics of Newly Instantiated Agents

To characterize the dynamics of expert agent instantiation, we summarize the most frequently created new expert agents and their predictive utility on DrugBank S2 tasks in Table 15. The total number reflects how often DDIAgents identifies a mechanism-specific reasoning gap that is not well covered by the default experts and therefore requests a specialized perspective.

Overall, "Cardiologist", "Neurologist", and "Toxicologist" are instantiated most often, suggesting that a substantial portion of hard cases in DrugBank S2 require specialized clinical safety assessment and adverse-event reasoning beyond general pharmacology. Notably, instantiation frequency does not always align with per-agent accuracy: "Endocrinologist" and "Gastroenterologist" achieve higher accuracy despite being created less often. This pattern indicates that certain expert perspectives are high-yield when triggered (i.e., they provide decisive evidence for a narrower subset of queries), whereas other experts are broad-coverage but face more ambiguous cases.

These statistics provide empirical evidence that DDIAgents benefits from dynamic expert instantiation: rather than relying on a fixed agent pool, it adaptively composes the experts needed to match the interaction-dependent evidence required by each query.

B.8 Case Study: the Correlation between Expert and Knowledge Source in Dynamic Context Flow

We conduct a case study to analyze the context flow of DDIAgents on the DrugBank dataset under S2 tasks (as shown in Figure 8). For clarity, we illustrate three predefined expert agents, noting that the instantiated expert profiles can vary across iterations. The retrieval rates of each expert over different knowledge sources are illustrated.

Overall, Metabolic DDIs, Synergistic DDIs, and DDI descriptions are retrieved most frequently across experts, suggesting that these sources provide particularly informative evidence for accurate DDI prediction. Moreover, the experts exhibit distinct retrieval preferences consistent with their functional roles. The Pharmacist emphasizes clinically oriented evidence (e.g., drug side effects, drug targets, and textual DDI descriptions), the Pharmacokineticist focuses on ADME-related sources (Absorption, Distribution, Metabolic and Excretion DDIs), and the Pharmacologist places relatively more weight on molecular structure signals (e.g., Drug SMILES) for chemical mechanism analysis.

Figure 8: Knowledge retrieval rates across expert agents and knowledge sources on DrugBank S2 tasks.

B.9 An Illustrative Example for Comparison on Generation Contents

In order to provide a clearer comparison between the generated contents of MDAgents and DDIAgents in Section 5.7, we present an illustrative example in Table 21. The contents include the DDI question, the answer by MDAgents and DDIAgents, and the LLM judgment. We can see that DDIAgents outperforms in all of the evaluation dimensions including thoroughness, diversity, rationality and overall quality.

Table 16: Performance comparison between different DDI prediction methods on S0 tasks. The best and second-best results are highlighted in bold and underline, respectively.

Type	Method	DrugBank S0		TWOSIDES S0	
		Acc	F1	Hit@5	NDCG@5
Feature-based	MLP	84.30±1.13	19.10±2.65	66.54±1.39	30.75±1.21
Graph-based	Decagon	90.61±0.71	85.65±1.10	54.07±0.47	21.10±1.06
	EmerGNN	96.71±0.11	94.12±0.39	68.71±0.29	33.09±0.31
	MSTE	92.17±0.27	71.30±2.74	30.91±0.07	12.91±0.04
	TIGER	95.85±0.83	92.31±0.31	62.84±0.77	29.51±0.41
LLM-based	Single LLM	9.17±0.26	4.79±0.19	20.44±0.19	9.82±0.12
	K-Path	92.49±2.27	89.86±2.71	55.39±1.81	24.14±1.36
	CBR-DDI	<u>96.84±0.62</u>	<u>94.92±0.76</u>	69.17±0.52	<u>35.40±0.16</u>
	TextDDI	96.19±0.48	91.54±0.33	57.09±0.77	28.51±0.39
	DDI-GPT	96.63±1.31	93.32±0.88	68.16±1.48	34.51±0.92
Agent-based	Reflexion	92.04±1.82	88.39±2.04	58.45±1.29	26.86±0.77
	Debate	94.41±0.92	93.12±1.36	63.11±0.68	32.10±0.99
	AgentVerse	95.77±1.41	94.04±1.27	67.08±0.82	33.09±0.55
	MDAgents	96.33±0.97	94.19±0.72	<u>69.63±1.22</u>	33.67±1.01
	DDIAgents	**96.95±0.84**	**95.02±1.09**	**71.57±0.72**	**35.82±0.44**

Table 17: Performance of DDIAgents using GPT-4o as the backbone LLM.

Method	DrugBank S2		TWOSIDES S2	
	Acc	F1	Hit@5	NDCG@5
DDIAgents (Qwen 2.5 7B)	38.21±1.03	18.17±0.51	42.26±0.92	14.08±0.15
DDIAgents (GPT-4o)	38.34±0.62	20.75±0.69	44.08±0.79	14.90±0.24

Table 18: The DDI prediction case by DDIAgents using Qwen 2.5 7B as the backbone LLM (The same DDI prediction question as illustrated in Section 5.4).

DDI prediction question	What is the drug-drug interaction type between Pirfenidone and Teriflunomide? A. Decreased metabolism B. Increased risk or severity of adverse effects C. Increased serum concentration D. Decreased serum concentration E. Decreased therapeutic efficacy
Pharmacist (Iteration 1)	**Assigned knowledge sources:** Metabolic DDIs, Excretion DDIs, Synergistic DDIs and Drug side effects. **Answer: Interaction type A.** Pirfenidone is a small molecule drug with multiple anti-fibrotic properties, while Teriflunomide is a disease-modifying treatment for multiple sclerosis. The metabolism of Teriflunomide can be decreased when combined with Pirfenidone, this suggests that Pirfenidone may inhibit enzymes that break down Teriflunomide, leading to elevated plasma levels and a higher risk of toxicity. This is a common scenario in drug interactions, especially in patients with multiple chronic conditions where polypharmacy is prevalent. Monitoring for adverse effects and potential dose adjustments would be necessary to manage this interaction effectively.
Pharmacokineticist (Iteration 1)	**Assigned knowledge sources:** Metabolic DDIs, Excretion DDIs, Synergistic DDIs and Antagonistic DDIs. **Answer: Interaction type A.** Pirfenidone is a small molecule with multiple anti-fibrotic properties, and Teriflunomide is a novel disease-modifying treatment for multiple sclerosis. Based on the candidate interaction types, the most plausible interaction is that the metabolism of Teriflunomide can be decreased when combined with Pirfenidone. This interaction indicates that Pirfenidone inhibits enzymes that break down Teriflunomide, leading to elevated plasma levels and a higher risk of toxicity. This aligns with the deep learning model's prediction, which has a high probability score, suggesting a strong likelihood of this interaction occurring. Therefore, monitoring for increased toxicity and potential dose adjustments would be necessary if these drugs are co-administered.
Pharmacologist (Iteration 1)	**Assigned knowledge sources:** Metabolic DDIs, Excretion DDIs, Synergistic DDIs, Antagonistic DDIs, Drug targets and Drug side effects. **Answer: Interaction type C.** Pirfenidone is a drug with multiple anti-fibrotic properties, while Teriflunomide is a novel disease-modifying treatment for multiple sclerosis (MS) that is in phase III development. Pirfenidone may inhibit enzymes that break down Teriflunomide, leading to elevated plasma levels and a higher risk of toxicity. This is a plausible interaction given that both drugs are in different therapeutic areas and their mechanisms of action do not directly overlap in terms of receptors or pathways. Therefore, the most plausible interaction type is A, where the metabolism of Teriflunomide is decreased when combined with Pirfenidone.
Conclusion agent (Iteration 1)	**Answer: Yes. Interaction type C.** The agents all agreed that the metabolism of Teriflunomide could be inhibited by Pirfenidone, leading to increased plasma levels and a higher risk of toxicity. Therefore, this interaction type is the most likely.

Table 19: A DDI prediction case by DDIAgents that different agents achieve consensus in two iterations.

DDI prediction question	What is the drug-drug interaction type between Riociguat and Macitentan? A. Decreased serum concentration B. Increased anticoagulant activities C. Increased risk or severity of adverse effects D. Increased serum concentration E. Increased hypotensive activities
Pharmacist (Iteration 1)	**Assigned knowledge sources:** Absorption DDIs, Metabolic DDIs and SMILES of two drugs. **Answer: Interaction type E.** Riociguat, a PDE5 inhibitor that increases cGMP levels, and Macitentan, a VEGFR-1 and VEGFR-2 inhibitor, both have vasodilatory effects. The combination of these drugs could lead to an increased risk of hypotension due to additive vasodilation. Reference examples such as bosentan increasing the hypotensive activities of penbutolol support this interaction. Therefore, the most plausible interaction type is that riociguat may increase the hypotensive activities of macitentan, leading to a higher risk of fainting, falls, or shock. Close clinical monitoring and adjustment of dosages may be necessary to manage this risk.
Pharmacokineticist (Iteration 1)	**Assigned knowledge sources:** Absorption DDIs, Metabolic DDIs and Synergistic DDIs. **Answer: Interaction type D.** Riociguat is known to increase cGMP levels by inhibiting PDE5, which can lead to vasodilation. Macitentan, on the other hand, is an endothelin receptor antagonist that also has vasodilatory properties. The combination of these two drugs, both of which have vasodilatory effects, could potentially lead to a significant increase in blood flow, resulting in elevated blood levels of Macitentan. This could increase both the efficacy and the toxicity risk of Macitentan, given its established side effects including hypotension and cardiovascular events. Therefore, the most plausible interaction is that Riociguat increases the serum concentration of Macitentan, potentially leading to higher efficacy and toxicity risk.
Pharmacologist (Iteration 1)	**Assigned knowledge sources:** Metabolic DDIs, Antagonistic DDIs, Drug targets and Drug side effects. **Answer: Interaction type E.** Riociguat increases cyclic guanosine monophosphate (cGMP) levels through the inhibition of phosphodiesterase type 5 (PDE5), leading to vasodilation. Macitentan, on the other hand, is a VEGFR-1 and VEGFR-2 inhibitor, which is not directly involved in the PDE5 pathway. However, the mechanism of action of both drugs involves cardiovascular effects. Riociguat is primarily used to treat pulmonary arterial hypertension, while Macitentan is used to treat pulmonary arterial hypertension and systemic sclerosis. The combination of these two drugs could potentially lead to an increased hypotensive effect due to the synergistic vasodilation, which aligns with the description of interaction type E. This interaction type suggests that riociguat would enhance the hypotensive activities of macitentan, increasing the risk of excessive blood pressure reduction and related adverse effects such as dizziness, fainting, or shock. Therefore, this is the most plausible interaction type based on the provided information and medical knowledge.
Conclusion agent (Iteration 1)	**No. Pharmacist:** Interaction type E is suggested by multiple experts, indicating an increased hypotensive effect. **Pharmacokineticist:** Interaction type D is proposed, indicating an increased serum concentration of Macitentan. **Pharmacologist:** Interaction type E is favored, suggesting an increased risk of adverse effects due to synergistic vasodilation. Further analysis is needed to clarify the most likely interaction type. A more detailed review of clinical studies and case reports involving these drugs in combination is recommended. Additionally, pharmacokinetic studies to assess the impact of riociguat on the metabolism or absorption of macitentan could provide more concrete evidence.
Planner agent for expert agent instantiation (Iteration 2)	**Cardiologist:** provide insights into the cardiovascular effects and the potential for increased hypotension. **Pulmonologist:** offer expertise on the respiratory effects and the interaction between these drugs, given their role in treating pulmonary arterial hypertension. **Nephrologist:** assess the potential renal implications of drug combination.
Cardiologist (Iteration 2)	**Assigned knowledge sources:** Absorption DDIs, Metabolic DDIs, Antagonistic DDIs and Drug targets. **Answer: Interaction type E.** Riociguat and Macitentan both have vasodilatory effects. Riociguat increases cyclic guanosine monophosphate (cGMP) levels, leading to vasodilation, while Macitentan inhibits VEGFR-1 and VEGFR-2, which also have vasodilatory effects. Given the expert opinions and the deep learning model predictions, the most plausible interaction is that Riociguat may increase the hypotensive activities of Macitentan. This interaction would result in a higher risk of excessive hypotension, dizziness, fainting, or falls. The clinical monitoring and titration recommended by experts align with this hypothesis. Therefore, interaction type E is the most plausible based on the provided information and medical knowledge.

Table 20: A DDI prediction case by DDIAgents that different agents achieve concensus in two iterations (continued).

Pulmonologist (Iteration 2)	**Assigned knowledge sources:** Absorption DDIs, Metabolic DDIs, Drug targets and Drug side effects. **Answer: Interaction type D.** Based on the provided descriptions and the deep learning model's predictions, there are indications that riociguat may enhance the hypotensive effects of macitentan, which is a significant concern given the clinical context of pulmonary arterial hypertension (PAH) where maintaining adequate blood pressure is critical. The pharmacologist's suggestion of interaction type E, which indicates an increased risk of adverse effects due to synergistic vasodilation, aligns with the known mechanisms of action for both drugs. Riociguat works by increasing nitric oxide bioavailability and cGMP levels, leading to vasodilation, while macitentan inhibits VEGFR-1 and VEGFR-2, also promoting vasodilation. The combination could result in excessive vasodilation, leading to hypotension. This is further supported by the pharmacist's and pharmacologist's consensus favoring interaction type E. The deep learning model's prediction of an increased hypotensive activity is the most plausible based on the available information and expert opinions.
Nephrologist (Iteration 2)	**Assigned knowledge sources:** Absorption DDIs, Metabolic DDIs, Excretion DDIs and Synergistic DDIs. **Answer: Interaction type E.** Riociguat is known to cause vasodilation, and Macitentan also has vasodilatory effects by inhibiting VEGFR-1 and VEGFR-2. Given that both drugs can lead to hypotension, the combination is likely to enhance this effect. The deep learning model suggests an increased hypotensive activity, which aligns with the pharmacological actions of both drugs. The pharmacist and pharmacologist have also favored this interaction type, indicating a consensus on the potential for increased hypotension. Therefore, the most plausible interaction type is that riociguat may increase the hypotensive activities of Macitentan, leading to a heightened risk of fainting, falls, or shock.
Conclusion agent (Iteration 2)	**Yes. Interaction type E.** Based on the comprehensive analysis of the provided information and expert opinions, the most plausible interaction between Riociguat and Macitentan is that Riociguat may increase the hypotensive activities of Macitentan. Both drugs have vasodilatory effects, with Riociguat increasing cyclic guanosine monophosphate (cGMP) levels and Macitentan inhibiting VEGFR-1 and VEGFR-2. The deep learning model also indicates an increased hypotensive activity, which aligns with the known mechanisms of action. Furthermore, the consensus among the pharmacologist, pharmacist, and other experts supports this interaction type, emphasizing the potential risk to explain.

Table 21: Illustrative case: generated content comparison between MDAgents and DDIAgents.

DDI prediction question	What is the drug-drug interaction type between Aclidinium and Fesoterodine? A. Increased risk or severity of adverse effects B. Increased anticholinergic activities C. Increased tachycardic activities D. Increased the QTc-prolonging activities E. Increased serum concentration
MDAgents	Interaction type A. The evidence provided from the deep learning model and the analysis of the agents strongly supports the interaction type A, where Aclidinium may increase the anticholinergic activities of Fesoterodine. The detailed explanations from the agents highlighting the anticholinergic effects of both drugs, provides a robust basis for this conclusion.
DDIAgents	Interaction type B. Both aclidinium and fesoterodine are anticholinergic agents, and their concurrent use can potentiate anticholinergic side effects, making interaction type B (increased anticholinergic activities) the most precise and clinically relevant interaction type. This is supported by the analysis provided by the pharmacists and pharmacologists, who emphasize the pharmacodynamic overlap and the potential for enhanced side effects when these drugs are used together.
LLM Decision	Thoroughness: "Winner": Answer 2 (DDIAgents). Explanation: Answer 2 is more comprehensive because it accurately aligns the drug mechanism with the specific multiple-choice option provided. It correctly identifies both drugs as anticholinergic agents and links this to Interaction type B. Answer 1, conversely, suffers from a mapping error; it selects "Interaction type A" but then proceeds to describe the definition of "Interaction type B" in its explanation, creating a contradiction that fails to fully address the question's structure. Diversity: "Winner": Answer 2 (DDIAgents). Explanation: Answer 2 offers a richer clinical perspective. It mentions "pharmacodynamic overlap" and the perspectives of "pharmacists and pharmacologists," which provides a broader professional context for why this interaction is significant. Answer 1 mentions a "deep learning model," which is a technical insight, but it lacks the clinical depth found in Answer 2. Rationality: "Winner": Answer 2 (DDIAgents). Explanation: Answer 2 is logically sound. It identifies the shared drug class and correctly matches it to the most specific answer choice (B). Answer 1 is irrational in its presentation: it claims the answer is "A" (the general risk of adverse effects) but then justifies it using the exact phrasing of option "B" (anticholinergic activities). This internal inconsistency makes Answer 1 unreliable. Overall: "Winner": Answer 2 (DDIAgents). Explanation: Answer 2 is the superior response. It is factually accurate, logically consistent, and maps the pharmacological mechanism to the correct multiple-choice option. Answer 1 contains a critical labeling error—matching the explanation for B to the label for A—which would lead a user to the wrong choice despite having the correct underlying reasoning. Answer 2 provides the "most precise and clinically relevant" answer, as it correctly identifies the specific nature of the interaction.