RAG-ENHANCED COLLABORATIVE LLM AGENTS FOR DRUG DISCOVERY

Namkyeong Lee[1], **Edward De Brouwer**[2*], **Ehsan Hajiramezanali**[2*], **Tommaso Biancalani** [2]
Chanyoung Park[1†], **Gabriele Scalia**[2†]
[1] KAIST [2] Genentech

ABSTRACT

Recent advances in large language models (LLMs) have shown great potential to accelerate drug discovery. However, the specialized nature of biochemical data often necessitates costly domain-specific fine-tuning, posing critical challenges. First, it hinders the application of more flexible general-purpose LLMs in cutting-edge drug discovery tasks. More importantly, it impedes the rapid integration of the vast amounts of scientific data continuously generated through experiments and research. To investigate these challenges, we propose CLADD, a retrieval-augmented generation (RAG)-empowered agentic system tailored to drug discovery tasks. Through the collaboration of multiple LLM agents, CLADD dynamically retrieves information from biomedical knowledge bases, contextualizes query molecules, and integrates relevant evidence to generate responses — all without the need for domain-specific fine-tuning. Crucially, we tackle key obstacles in applying RAG workflows to biochemical data, including data heterogeneity, ambiguity, and multi-source integration. We demonstrate the flexibility and effectiveness of this framework across a variety of drug discovery tasks, showing that it outperforms general-purpose and domain-specific LLMs as well as traditional deep learning approaches.

1 INTRODUCTION

Large language models (LLM) have revolutionized the landscape of natural language processing, emerging as general-purpose foundation models with remarkable abilities across multiple domains (Achiam et al., 2023; Touvron et al., 2023). However, given the inherent complexity and specialized nature of the biomolecular field, recent works emphasize the importance of domain-specific fine-tuning to boost tasks such as molecular captioning, property prediction, or binding affinity prediction (Fang et al., 2023; Chaves et al., 2024; Yu et al., 2024; Edwards et al., 2024). Consequently, rather than employing readily available general-purpose LLMs, most efforts in drug discovery have focused on fine-tuning LLMs using biochemical annotations or instruction-tuning datasets.

While promising, solely relying on these approaches poses significant challenges that can limit applications. On one hand, the rapid emergence of new LLM architectures and techniques (Minaee et al., 2024; Zhao et al., 2023b) complicates maintaining domain-specific models obtained through expensive fine-tuning. More importantly, drug discovery applications often require promptly incorporating new insights as they become available, for example, as a result of new experiments or through the scientific literature—a process exacerbated by the automation of experimental workflows (Tom et al., 2024). In addition to being impractical, regular rounds of fine-tuning to keep LLMs up-to-date with the latest scientific advances also introduce challenges such as catastrophic forgetting (Luo et al., 2023).

From this perspective, retrieval-augmented generation (RAG) methods offer a promising solution that enables dynamic adaptation of the model's knowledge without the need for expensive fine-tuning (Gao et al., 2023). However, applying this paradigm in the drug discovery domain presents important obstacles. First, retrieving relevant knowledge is difficult due to the limited domain expertise of general-purpose LLMs, combined with the vastness of the chemical space (Bohacek et al.,

[*]denotes Equal Contribution, and [†] denotes Corresponding Authors

1996) that renders exact retrieval suboptimal. Second, biochemical data is extremely heterogeneous, spanning diverse modalities such as molecules, proteins, diseases, and complex relationships between them (Wang et al., 2023). Such data can also exist across multiple sources, introducing challenges in factual integration (Harris, 2023). Finally, the available information is not necessarily relevant to the query, as it may be too general, ambiguous, or partial (Vamathevan et al., 2019).

In this study, we tackle these challenges by introducing a **C**ollaborative framework of **L**LM **A**gents for **D**rug **D**iscovery (CLADD). We assume a general setting where external knowledge is available as expert annotations associated with molecules or as knowledge graphs that flexibly represent diverse biochemical entities and their relationships. CLADD is powered by general-purpose LLMs, while also integrating domain-specific LMs, when necessary, to improve molecular understanding. Notably, external knowledge can be updated dynamically without LLM fine-tuning.

The multi-agent collaborative framework enables each agent to specialize in a specific data source and/or role based on their team, offering a modular solution that can improve overall information processing (Chan et al., 2024). In particular, CLADD includes a *Planning Team* to determine relevant data sources, a *Knowledge Graph Team* to retrieve external heterogeneous information in the knowledge graph and summarize it, also through a novel anchoring approach to retrieve related information when the query molecule is not present in the knowledge base, and a *Molecule Understanding Team*, which analyzes the query molecule based on its structure along with summaries of external data and tools. The flexibility of the framework enables CLADD to address a wide range of tasks for drug discovery, including zero-shot settings, while also improving interpretability through the transparent interaction of its agents.

Overall, we highlight the following contributions:

- We present CLADD, a multi-agent framework for RAG-based question-answering in drug discovery applications. The framework leverages generalist LLMs and dynamically integrates external biochemical data from multiple sources without requiring fine-tuning.
- We demonstrate the flexibility of the framework by tackling diverse applications, including property-specific molecular captioning, drug-target prediction, and molecular toxicity prediction.
- We provide comprehensive experimental results showcasing the effectiveness of CLADDcompared to general-purpose and domain-specific LLMs, as well as standard deep learning approaches. A further appeal of CLADD is its flexibility and explainability, improving the interaction between scientists and AI.

2 METHODOLOGY

2.1 PROBLEM SETUP

Given a query molecule g_q and a textual prompt describing a task of interest $\mathcal{I}$, we consider the general problem of generating a relevant response $\mathcal{A}_{g_q}$. For instance, given $g_q =$ 'C1=CC(=C(C=C1CCN)O)O ' and $\mathcal{I} =$ '*Predict liver toxicity*', our model should be able to generate an answer stating that $\mathcal{A}_{g_q} =$ '*this molecule does not have liver toxicity concerns*'.

We assume access to two types of external databases: (1) molecular annotation databases $\mathcal{C}$, which contain textual annotation about molecules (for example, detailing their functions and properties) and (2) knowledge graphs (KGs) connecting molecules to other biomedical entities. In particular, a KG $\mathcal{G}$ is composed of a set of entities $\mathcal{E}$ and a set of relations $\mathcal{R}$ connecting them. KG can include various types of entities, such as drugs, proteins, and diseases. In this paper, we only assume that molecule (or drug) entities are present in KG, while any other types of entities can exist.

Additionally, we assume access to pre-trained molecular captioning models that can be used as external tools to complement the external databases. In general, any predictive model on molecules can be considered a captioning model (Edwards et al., 2022; Pei et al., 2023), given that its output can be simply represented as text.

2.2 CLADD

Here, we introduce CLADD, a multi-agent framework for general molecular question-answering that supports multiple drug discovery tasks. Each agent is implemented by an off-the-shelf LLM

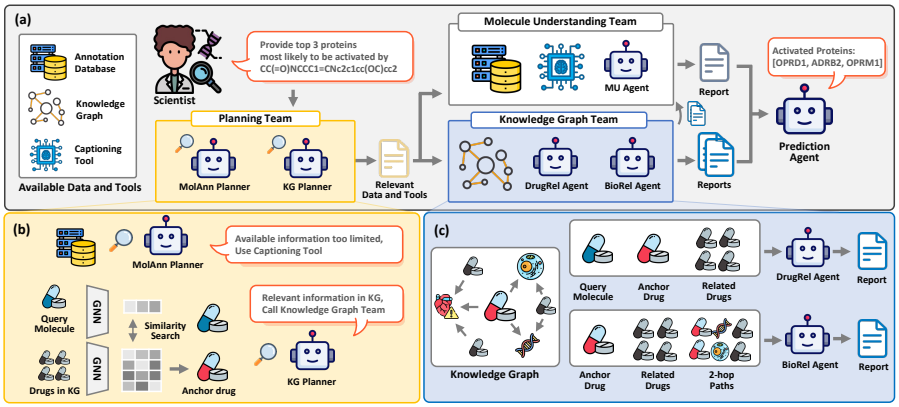

Figure 1: Overview of CLADD.

prompted to elicit a particular behavior. Our framework is composed of three teams, each composed of several agents: the **Planning Team**, which identifies the most appropriate data sources and overall strategy given the task and the query molecule (Section 2.2.1); the **Knowledge Graph (KG) Team**, which retrieves relevant contextual information about the molecule from available KG databases (Section 2.2.2); and the **Molecular Understanding (MU) Team**, which retrieves and integrates information from molecular annotation databases and external tools for molecule description (Section 2.2.3); Finally, the **Prediction Agent** integrates the findings from the MU and KG teams to generate the final answer. In the following sub-sections, we describe each team in detail. The overall framework is depicted in Figure 1.

2.2.1 Planning Team

The Planning Team assesses prior knowledge for a given query molecule. The team separately assesses the relevance of the molecular annotations and the knowledge graph using a MolAnn Planner agent and a KG Planner agent.

Molecule Annotation (MolAnn) Planner. This agent first retrieves annotations for the query molecule, c_q, from the annotation database $\mathcal{C}$. While these annotations can provide valuable biochemical knowledge (Yu et al., 2024), they are often sparse and may lack sufficient details due to the vastness of the chemical space (Lee et al., 2024). Given the vastness of the chemical space, it is also not uncommon for molecules to be completely absent from databases.

To this end, the MolAnn Planner determines whether the retrieved annotations provide enough information for subsequent analyses. Specifically, given a query molecule g_q and retrieved annotations c_q, the agent is invoked as follows: $o_{\text{MAP}} = \text{MolAnn Planner}(g_q, c_q)$. o_{MAP} is a Boolean indicating whether annotations should be complemented with additional information from tools.

Knowledge Graph (KG) Planner. In parallel to analyzing the available description for the query molecule, we analyze the relevance of the contextual information present in the KG. Unlike previous works in general QA tasks, which primarily rely on identifying exact entity matches within the KG (Baek et al., 2023), the vast chemical search space and the limited coverage of existing knowledge bases hinder such approaches.

To address this challenge, we propose leveraging the knowledge of drugs that are structurally similar to the query drug, building upon the well-established biochemical principle that structurally similar molecules often exhibit related biological activity (Martin et al., 2002). Specifically, we define the *anchor drug* g_a as the entity drug that maximizes the cosine similarity between its embedding and that of the query molecule, among the set of all molecules in the KG ($g_{\mathcal{G}}$):

$$g_a = \underset{g \in g_{\mathcal{G}}}{\text{argmax}} \frac{emb(g_q) \cdot emb(g)}{\|emb(g_q)\| \|emb(g)\|}, \tag{1}$$

where *emb* is a graph neural network (GNN) pre-trained with 3D geometry (Liu et al., 2021) that outputs structure-aware molecular embeddings.

Then, the KG Planner agent decides whether to use the KG based on the structural similarity between the query molecule and the retrieved anchor drug. To do so, we also provide the Tanimoto

similarity[1] to the KG Planner, as this domain-specific metric can be leveraged by the LLM's reasoning about chemical structural similarity as follows: $o_{\text{KGP}} = \text{KG Planner}(g_q, g_a, s_{q,a})$, where $s_{q,a}$ is the Tanimoto similarity between the query and anchor molecules. o_{KGP} is a Boolean indicating whether the KG should be used for the prediction.

2.2.2 KNOWLEDGE GRAPH TEAM

This team aims to provide relevant contextual information about the query molecule by leveraging the KG, and it is only called if $o_{\text{KGP}} = \text{TRUE}$. It consists of the Drug Relation (DrugRel) Agent and the Biological Relation (BioRel) Agent, both of which generate reports on the query molecule based on different aspects of the KG. Specifically, the DrugRel Agent focuses on related drug entities within the KG, primarily leveraging its internal knowledge, whereas the BioRel Agent focuses on summarizing and assessing biological relationships between drugs present in the KG.

Related Drugs Retrieval. The typical approach to leveraging a KG for QA tasks involves identifying multiple entities in the query and extracting the subgraph that encompasses those entities (Baek et al., 2023; Wen et al., 2023). However, in molecular understanding for applications related to drug discovery tasks, the question often involves only a single entity, i.e., the query molecule g_q, making it challenging to identify information in the KG relevant to the task.

Here, we introduce a novel approach for extracting relevant information for the query molecule g_q by utilizing the retrieved anchor drug g_a, which exhibits high structural similarity to the query molecule. In particular, while the drug entities in the KG $\mathcal{G}$ are mainly connected to other types of biological entities (e.g., proteins, diseases), we can infer relationships among drugs by considering the biological entities they share. For example, we can determine the relatedness of the drugs Trastuzumab and Lapatinib by observing their connectivity to the protein HER2 in the KG, as both drugs specifically target and inhibit HER2 to treat HER2-positive breast cancer (De Azambuja et al., 2014). Therefore, to identify relevant related drugs, we first compute the 2-hop paths connecting the anchor drug g_a to other drugs $g^i_{\mathcal{G}}$ in the KG $\mathcal{G}$, i.e., $(g_a, r_{a\to e}, e, r_{i\to e}, g^i_{\mathcal{G}})$, where $r \in \mathcal{R}$, $e \in \mathcal{E}$, and i denotes the index of the other drug. Then, we select the top-k *related drugs*, denoted as $g_{r^1}, \ldots, g_{r^k}$, corresponding to the molecules that have the greatest number of 2-hop paths to the anchor drug. Note that while the anchor drug g_a is selected based on its structural similarity to the query molecule g_q, these reference drugs are biologically related to g_a, reflecting the relationships captured within the KG.

Drug Relation (DrugRel) Agent. The DrugRel Agent generates a report on the query molecule, contextualizing it in relation to relevant drugs present in the knowledge base for the specific task instruction. Given a query molecule g_q, its anchor drug g_a, and the set of related drugs $g_{r^1}, \ldots, g_{r^k}$, the DrugRel Agent is defined as follows: $o_{\text{DRA}} = \text{DrugRel Agent}\,(g_q, g_a, g_{r^1}, \ldots, g_{r^k}, \mathcal{T}, \mathcal{I})$, where $\mathcal{T} = \{s_{q,a}, s_{q,r^1}, \ldots, s_{q,r^k}\}$ is the set of Tanimoto similarities between the query molecule and the retrieved drugs, and $\mathcal{I}$ is the task instruction. This allows the agent to leverage its internal knowledge about related drugs while effectively assessing the relatedness of the information to the target molecule based on the Tanimoto similarity.

Biological Relation (BioRel) Agent. On the other hand, drugs related to the query molecule in the KG may exhibit limited structural similarity, underscoring the importance of utilizing additional relevant information, such as shared toxicity profiles or interactions with the same target, rather than relying solely on structural resemblance. Therefore, the BioRel Agent summarizes how the anchor drug and the related drugs are biologically related, integrating additional biochemical information present in the KG. Specifically, given an anchor drug g_a, a set of reference drugs $g_{r^1}, \ldots, g_{r^k}$, the collection of all 2-hop paths $\mathcal{P}$ linking the anchor drug to the reference drugs, and the instruction $\mathcal{I}$, the agent generates the report as follows: $o_{\text{BRA}} = \text{BioRel Agent}(\mathcal{P}, \mathcal{I}, g_q, g_a, s_{q,a})$. This enables us to obtain a task-relevant summary of the subgraph connected to the anchor drug.

Importantly, while both the DrugRel Agent and BioRel Agent aim to reason about the query molecule in relation to other relevant drugs in the KG for the specific task, they leverage distinct knowledge sources and perform different roles. Specifically, the BioRel Agent primarily leverages the connectivity between drugs and other biological entities in the KG, focusing on interpreting and summarizing this network of relationships and aligning it with the specific task at hand. In contrast, the DrugRel Agent primarily draws on its internal knowledge, triggered by the names of the

[1] We provide details on the Tanimoto similarity in Appendix B.

related drug entities in the KG, and incorporates structural similarity between them. In Section 3, we demonstrate how these agents complement each other effectively, producing a synergistic effect when combined together.

2.2.3 MOLECULAR UNDERSTANDING TEAM

While the KG Team compiles the report by aggregating contextual knowledge, the Molecule Understanding (MU) Team focuses primarily on the query molecule itself. The MU Team is composed of a single Molecule Understanding (MU) Agent, which aims to write a report on the query molecule by leveraging its structural information, annotations from tools, and reports from other agents.

Molecule Understanding (MU) Agent. The MU Agent retrieves annotations for the query molecule, denoted as c_q. If the Planning Team decides to use external annotation tools (*i.e.,* $o_{\text{MAP}} = \text{TRUE}$), additional captions $\tilde{c}_q$ are generated with the external captioning tools as follows: $\tilde{c}_q = \text{Captioning Tools}(g_q)$, and concatenated to the annotations retrieved from the database: $c_q = c_q || \tilde{c}_q$. External captioning tools allow the system to easily harness recent advances in LLM-driven molecular understanding (Pei et al., 2023; Yu et al., 2024).

The agent then analyzes the structure of the molecule, contextualizing it with reports generated by the other agents. Using the SMILES representation, the caption of the query molecule, and the reports from the KG Team, it compiles a comprehensive molecular annotation report as follows: $o_{\text{MUA}} = \text{MU Agent}(g_q, c_q, o_{\text{DRA}}, o_{\text{BRA}}, \mathcal{I})$.

2.2.4 PREDICTION AGENT

Finally, the Prediction Agent performs the user-defined task by considering the reports from the various agents, including the MU and KG teams, as follows: $\mathcal{A}_{g_q} = \text{Task Agent}(g_q, o_{\text{MUA}}, o_{\text{DRA}}, o_{\text{BRA}}, \mathcal{I})$. By integrating this evidence, the Prediction Agent can perform a comprehensive analysis of the query molecule. Importantly, the output of the Prediction Agent can be flexibly adjusted based on the specific task requirements. For instance, it can be a descriptive caption, a simple yes/no response for binary classification, or a more complex answer listing the top-k targets associated with the query molecule. Such behavior leverages the zero-shot capabilities of LLMs (Kojima et al., 2022) and does not require additional fine-tuning. Therefore, a key advantage of CLADDis its flexibility, which enhances scientist-AI interactions.

3 EXPERIMENTS

Implementation Details. In all experiments, we utilize GPT-4o mini through the OpenAI API for each agent. We use PrimeKG (Chandak et al., 2023) as the KG, PubChem (Kim et al., 2021) as an annotation database, and MolT5 (Edwards et al., 2022) as an external captioning tool. Additional implementation details and agent templates can be found in Appendix E and G, respectively.

3.1 PROPERTY-SPECIFIC MOLECULAR CAPTIONING TASK

Earlier studies on molecular captioning tasks have primarily focused on generating general descriptions of molecules without targeting specific areas of interest, raising concerns about their practical applicability in real-world drug discovery tasks. Indeed, the usefulness of a molecular description is often task-dependent, and scientists may be interested in detailed explanations of specific characteristics of a molecule rather than a general description (Guo et al., 2024; Edwards et al., 2024). Hence, in this paper, we introduce *property-specific molecular captioning*, where the model is required to generate a description for a given molecule customized to a particular area of interest.

Datasets. We leverage four widely recognized molecular property prediction datasets from the MoleculeNet benchmark (Wu et al., 2018): **BBBP**, **Sider**, **ClinTox**, and **BACE**. We provide further details on the datasets in Appendix C.

Methods Compared. We consider different baseline approaches. First, we compare recent molecular captioning methods designed to generate general descriptions of molecules, including MolT5 (Edwards et al., 2022), LlasMol (Yu et al., 2024), and BioT5 (Pei et al., 2023). Furthermore, we assess general-purpose LLMs, namely GPT-4o mini and GPT-4o. Finally, we consider two

Table 1: Performance in molecular captioning tasks, mean AUROC with standard deviation (in parentheses). **Bold** and underline indicate the best and second-best language model-based methods.

	BBBP	Sider	ClinTox	BACE
GNNs				
GraphMVP	69.59 (1.29)	60.88 (0.41)	87.57 (3.26)	80.24 (2.92)
MoleculeSTM	70.14 (0.90)	58.69 (0.89)	92.19 (2.79)	79.24 (3.40)
Only SMILES	70.95 (1.14)	60.80 (1.18)	91.62 (2.18)	74.21 (1.32)
General LLMs				
GPT-4o mini	67.85 (1.50)	58.18 (1.55)	90.74 (1.91)	74.22 (1.95)
GPT-4o	66.43 (1.47)	60.41 (1.21)	88.13 (1.74)	67.82 (4.14)
Domain LMs				
MolT5	69.77 (1.89)	57.20 (0.98)	87.91 (1.25)	74.28 (4.00)
LlasMol	68.12 (1.48)	61.50 (1.66)	89.67 (0.57)	75.42 (2.98)
BioT5	69.68 (1.23)	64.65 (2.01)	92.80 (2.92)	77.23 (1.95)
CLADD	**72.28 (1.04)**	**66.42 (1.31)**	**93.80 (2.30)**	**77.74 (3.15)**

Table 2: Performance in drug target tasks (Precision @ 5). **Bold** and underline indicate best and second-best language model-based methods.

	(a) Overlap		(b) No overlap	
	Activate	Inhibit	Activate	Inhibit
GNNs (Fine-tune)				
GraphMVP	1.76	1.03	1.67	0.73
MoleculeSTM	1.66	0.89	1.48	0.65
General LLMs (Zero-shot)				
GPT-4o mini	1.15	1.02	1.13	0.87
GPT-4o	0.62	0.79	0.68	0.65
Domain LMs (Zero-shot)	N/A	N/A	N/A	N/A
Domain LMs (Fine-tune)				
Galactica 125M	1.36	1.03	0.86	0.69
Galactica 1.3B	1.65	1.09	1.37	0.80
Galactica 6.7B	1.52	0.97	1.22	0.71
CLADD(**Zero-Shot**)	**3.04**	**4.83**	**2.67**	**3.24**

GNNs pre-trained with different methodologies: GraphMVP (Liu et al., 2021) and MoleculeSTM (Liu et al., 2023). We provide further details on the baseline models in Appendix D.

Evaluation Protocol. Although property-specific captions are practical, no ground truth property-specific captions exist for individual molecules, rendering traditional text generation evaluation methods inapplicable. Thus, in line with recent works (Xu et al., 2024; Guo et al., 2024; Edwards et al., 2024), we assess whether the generated captions can drive a classification model that categorizes molecules based on their properties. Specifically, for a generated `caption` and the `SMILES` representation of the target molecule, we concatenate them using a `[CLS]` token, forming `SMILES[CLS]caption`, and fine-tune a SciBERT (Beltagy et al., 2019) model for property prediction. The "Only SMILES" model utilizes only the SMILES string as input for the SciBERT classifier. For baseline GNNs, we convert the SMILES into a molecular graph and provide it to the model. For all the experiments, we use a scaffold splitting strategy to simulate realistic distribution shifts (train/validation/test data split as 80/10/10%), following previous works (Liu et al., 2023). We perform five independent fine-tuning runs of SciBERT (or GNN baselines) and report the mean and standard deviation of the AUROC.

Experimental Results. Table 1 summarizes the results. We note the following observations: **1)** While domain-specific models outperform general-purpose LLMs, their performance remains sub-optimal, occasionally falling behind the "Only SMILES" approach. This means that the generated captions occasionally reduce model performance compared to using only the SMILES representation of the molecule. This aligns with previous work that found that general descriptors may lack property-specific relevance (Guo et al., 2024; Edwards et al., 2024). **2)** On the other hand, CLADD-generated captions consistently outperform all the baseline captioners and successfully improve over "Only SMILES" across all datasets. We attribute this improvement to the ability of CLADDto draw on external biochemical knowledge to ground its generation and its task-specificity. **3)** Moreover, CLADDconsistently outperforms pre-trained GNN baselines, except on the BACE dataset. Interestingly, this is also the only dataset for which the "Only SMILES" baseline falls short compared to GNN models, thus highlighting the critical role of 2D topological and 3D geometric information in this case. This paves the way for future research on injecting essential aspects of molecules, such as topological and geometric information, into LLM understanding.

3.2 Drug-Target Prediction Task

Accurately predicting a drug's protein target is essential for understanding its mechanism of action and optimizing its therapeutic efficacy while minimizing off-target effects (Santos et al., 2017; Batool et al., 2019). Here, we evaluate the models' ability to accurately identify which proteins a given molecule is most likely to activate or inhibit.

Datasets. We use molecular targets present in the Drug Repurposing Hub (Corsello et al., 2017), DrugBank (Wishart et al., 2018), and STITCH v5.0 (Szklarczyk et al., 2016), as preprocessed in Zheng et al. (2023), including 13,688 molecules in total (details are presented in Appendix C).

Methods Compared. We evaluate two pre-trained GNNs, GraphMVP and MoleculeSTM, along with two general-purpose LLMs—GPT-4o mini and GPT-4o, and the domain-specific language model Galactica (Taylor et al., 2022) (details are presented in Appendix D).

Table 3: Performance in drug toxicity prediction task (Macro-F1). **Avg.** indicates the average performance over four datasets. **Bold** and underline indicate best and second-best methods.

	hERG	DILI	Skin	Carcinogens	Avg.
General LLMs					
GPT-4o mini	28.42	33.47	41.84	**69.51**	43.31
GPT-4o	40.45	25.76	**54.51**	38.91	39.90
Domain LMs					
Galactica 125M	40.78	33.56	42.43	17.64	33.60
Galactica 1.3B	48.57	34.37	42.43	44.00	42.34
Galactica 6.7B	23.75	57.67	40.41	65.20	46.76
GIMLET	36.50	35.51	42.28	48.47	40.69
LlasMol	23.75	**61.20**	31.92	51.68	42.14
CLADD	**51.46**	41.10	50.43	58.31	**50.33**

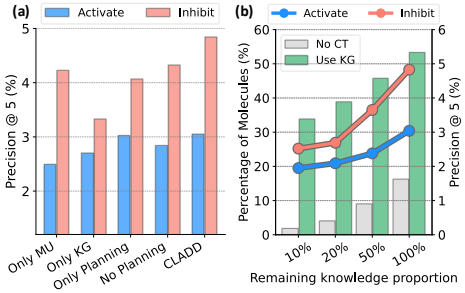

Table 4: **Ablation studies. (a)** On model components. **(b)** On external knowledge.

Evaluation Protocol. We assess the performance of LLMs in a zero-shot setting. Specifically, for a given target molecule, we prompt the LLMs to generate the top 5 proteins that the molecule is most likely to activate or inhibit. The precision is then calculated by determining whether the generated proteins match the correct answers. As baseline GNNs cannot perform this task without training in a zero-shot setting, we fine-tune them in a few-shot setting using 10% of the data. For domain-specific LMs, we also present fine-tuning results on the specific task. To better assess generalization power, we separately report the performance on the test set for molecules present/not present in the external databases ("Overlap"/"No Overlap").

Experimental Results. Table 2 summarizes the results. We observe the following: **1)** CLADD outperforms all the baselines, with a higher likelihood of correctly identifying proteins activated/inhibited by the input molecule. **2)** Importantly, the superiority of CLADDis confirmed for molecules not present in the caption database or knowledge graph (Table 2 (b)), showcasing CLADD's ability to leverage external knowledge to generalize to novel molecules. This outcome underscores the potential of CLADD for exploring entirely new molecular spaces. **3)** We observe that domain-specific fine-tuned models, such as Galactica, GIMLET, and MolecularGPT, could not generate five protein targets in a zero-shot setting when prompted to do so, as they are limited to providing answers based on their fine-tuning instruction dataset. By specifically fine-tuning Galactica on the task, we were able to answer the specific question, outperforming general-purpose LLMs in most experiments, but results were still inferior to CLADD. This further highlights the flexibility of CLADD, which leverages the zero-shot abilities of general-purpose LLMs in its architecture.

3.3 Drug Toxicity Prediction Task

Accurate predictions of drug toxicity are crucial to ensure patient safety and minimize the risk of adverse effects during drug development (Basile et al., 2019). Here, we evaluate the models' ability to predict the toxicity of a target molecule from its SMILES-based structural description.

Datasets. We use four datasets to comprehensively evaluate the performance of CLADD for drug toxicity prediction tasks: **hERG** (Wang et al., 2016), **DILI** (Xu et al., 2015), **Skin** (Alves et al., 2015), **Carcinogens** (Lagunin et al., 2009) datasets (details are presented in Appendix C).

Methods Compared. We compare five domain-specific LLMs—Galactica 125M, Galactica 1.3B, Galactica 6.7B (Taylor et al., 2022), LlasMol (Yu et al., 2024), and GIMLET (Zhao et al., 2023a) alongside two general-purpose LLMs, GPT-4o and GPT-4o mini (details in Appendix D).

Evaluation Protocol. As all the methods compared are foundation models, we evaluate their performance in a zero-shot setting. Specifically, given a SMILES-based structural description of the target molecule and a task description, the model outputs whether the molecule possesses the target property (binary classification). Using the text-formatted output generated by each model, we compute the Macro-F1 score (Opitz & Burst, 2019) as the evaluation metric.

Experimental Results. Table 3 summarizes the results. **1)** Comparing Galactica models, we observe that performance improves as the model size increases. However, this trend is not confirmed for GPT-4o and GPT-4o mini. This suggests that larger general-purpose models do not necessarily excel in domain-specific tasks, a trend also reported in prior research (Edwards et al., 2024). These findings emphasize the necessity of domain-specific training in combination with model scaling. **2)** Meanwhile, CLADD outperforms all the baselines (average score across datasets), without requiring domain-specific training by effectively incorporating external knowledge into general-purpose

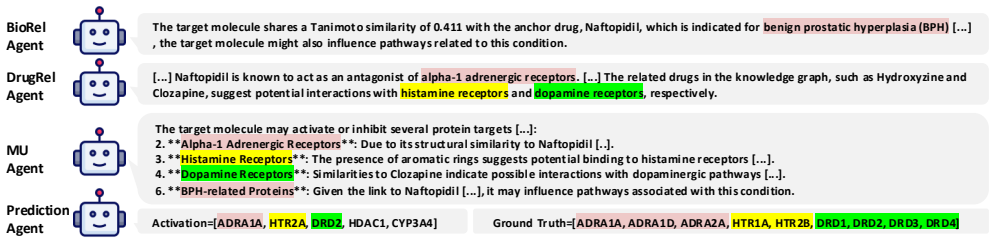

Figure 2: Example of collaboration between agents in CLADD on the drug-target prediction task.

LLMs. This highlights the potential of integrating external knowledge at inference time to improve molecular understanding, acting as an alternative (or complementary approach) to expensive domain-specific model fine-tuning.

3.4 Ablation studies

Model Components Ablations. In Figure 4 (a), we report the results of ablations on the components of CLADD. We observe: **1)** *The knowledge graph and the molecular annotations are important and complementary data sources*, as shown by the lower performance of the models with only Molecular Understanding or Only Knowledge Graph team available ("Only MU", "Only KG"). **2)** *Dynamically selecting the relevant data sources with our Planning Team leads to better performance*, leveraging their complementarity, as suggested by the lower performance of the "No Planning". **3)** *The distributed architecture of the multi-agent system is a more effective way of processing the retrieved information*. In "Only Planning", we concatenate all the relevant data sources directly into the prompt of the Prediction Agent, bypassing the preprocessing and report generation of the KG and MU teams. Our results show the limitations of a single model in handling heterogeneous biological data sources.

External Knowledge Ablations. To further assess the impact of external knowledge on model performance, we evaluate the model after progressively pruning the available databases and present our results in Figure 4 (b). We observe the following: **1)** *Model performance depends on external knowledge size*, validating the key role of the external knowledge to the framework. **2)** Interestingly, *we do not observe any performance plateau*, indicating that further expanding the external knowledge could provide additional performance improvements. **3)** From the bar plots, *i.e.,* "No CT (No Captioning Tool)" and "Use KG (Call Knowledge Graph Team)," we observe that as the amount of external knowledge grows, the planning team increasingly depends on it. This indicates that CLADD actively leverages external knowledge more effectively during the decision-making process when such knowledge is more abundant.

3.5 Case Studies

In Figure 2, we provide an example of agent collaboration in CLADD to identify the top 5 proteins a query molecule is most likely to activate. First, the BioRel Agent extracts from the knowledge graph that the anchor drug, Naftopidil, is indicated for benign prostatic hyperplasia (BPH), pointing to likely activation of pathways related to that condition. The DrugRel Agent then complements these findings by **1)** connecting BPH with alpha-1 adrenergic receptors using its internal knowledge (which is confirmed in the literature (Klotsman et al., 2004)), and **2)** analyzing the related drugs in the knowledge graph (*e.g.,* Hydroxyzine and Clopamine), inferring potential interaction with histamine and dopamine receptors. Finally, the MU agent integrates these findings with its own evidence to provide a summarized report of the likely activated proteins. This example highlights the complementary nature of the agents in CLADD, where the integration of diverse information from each agent is key to its success. The distributed architecture of the model, grounded in real data sources, also leads to increased interpretability and reliability of the generated answers.

4 Conclusion

In this work, we introduced CLADD, a multi-agent framework for molecular question-answering that dynamically retrieves and integrates external knowledge to support various drug discovery tasks. We showcased its flexibility and effectiveness across multiple tasks, outperforming both general-purpose and domain-specific LLMs as well as standard deep learning methods, without requiring

expensive domain-specific fine-tuning. Our analyses highlighted the complementarity of external knowledge sources, internal LLM reasoning, and multi-agent orchestration. CLADD's chain of messages also provides insight into its decision-making process and the role of different data sources, fostering more interpretable scientist-AI interactions.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Vinicius M Alves, Eugene Muratov, Denis Fourches, Judy Strickland, Nicole Kleinstreuer, Carolina H Andrade, and Alexander Tropsha. Predicting chemically-induced skin reactions. part i: Qsar models of skin sensitization and their application to identify potentially hazardous compounds. *Toxicology and applied pharmacology*, 284(2):262–272, 2015.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *arXiv preprint arXiv:2306.04136*, 2023.

Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7:1–13, 2015.

Anna O Basile, Alexandre Yahi, and Nicholas P Tatonetti. Artificial intelligence for drug toxicity and safety. *Trends in pharmacological sciences*, 40(9):624–635, 2019.

Maria Batool, Bilal Ahmad, and Sangdun Choi. A structure-based drug discovery paradigm. *International journal of molecular sciences*, 20(11):2783, 2019.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

Regine S Bohacek, Colin McMartin, and Wayne C Guida. The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal research reviews*, 16(1):3–50, 1996.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FQepisCUWu.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023.

Juan Manuel Zambrano Chaves, Eric Wang, Tao Tu, Eeshit Dhaval Vaishnav, Byron Lee, S Sara Mahdavi, Christopher Semturs, David Fleet, Vivek Natarajan, and Shekoofeh Azizi. Tx-llm: A large language model for therapeutics. *arXiv preprint arXiv:2406.06316*, 2024.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. Unifying molecular and textual representations via multi-task language modelling. *arXiv preprint arXiv:2301.12586*, 2023.

Steven M Corsello, Joshua A Bittker, Zihan Liu, Joshua Gould, Patrick McCarren, Jodi E Hirschman, Stephen E Johnston, Anita Vrcic, Bang Wong, Mariya Khan, et al. The drug repurposing hub: a next-generation drug library and information resource. *Nature medicine*, 23(4): 405–408, 2017.

Evandro De Azambuja, Andrew P Holmes, Martine Piccart-Gebhart, Eileen Holmes, Serena Di Cosimo, Ramona F Swaby, Michael Untch, Christian Jackisch, Istvan Lang, Ian Smith, et al. Lapatinib with trastuzumab for her2-positive early breast cancer (neoaltto): survival outcomes of a randomised, open-label, multicentre, phase 3 trial and their association with pathological complete response. *The lancet oncology*, 15(10):1137–1146, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

Carl Edwards, Ziqing Lu, Ehsan Hajiramezanali, Tommaso Biancalani, Heng Ji, and Gabriele Scalia. Molcap-arena: A comprehensive captioning benchmark on language-enhanced molecular property prediction. *arXiv preprint arXiv:2411.00737*, 2024.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*, 2023.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151, 2024.

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

Alireza Ghafarollahi and Markus J Buehler. Protagents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 2024.

Haoqiang Guo, Sendong Zhao, Haochun Wang, Yanrui Du, and Bing Qin. Moltailor: Tailoring chemical molecular representation to specific tasks via text prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18144–18152, 2024.

Emily Harris. Large language models answer medical questions accurately, but can't match clinicians' knowledge. *JAMA*, 2023.

Yoshitaka Inoue, Tianci Song, and Tianfan Fu. Drugagent: Explainable drug repurposing agent with large language model-based reasoning. *arXiv preprint arXiv:2408.13378*, 2024.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem in 2021: new data content and improved web interfaces. *Nucleic acids research*, 49(D1):D1388–D1395, 2021.

M Klotsman, CR Weinberg, K Davis, CG Binnie, and KE Hartmann. A case-based evaluation of srd5a1, srd5a2, ar, and adra1a as candidate genes for severity of bph. *The Pharmacogenomics Journal*, 4(4):251–259, 2004.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Alexey Lagunin, Dmitrii Filimonov, Alexey Zakharov, Wei Xie, Ying Huang, Fucheng Zhu, Tianxiang Shen, Jianhua Yao, and Vladimir Poroikov. Computer-aided prediction of rodent carcinogenicity by pass and cisoc-psct. *QSAR & Combinatorial Science*, 28(8):806–810, 2009.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

Namkyeong Lee, Siddhartha Laghuvarapu, Chanyoung Park, and Jimeng Sun. Vision language model is not all you need: Augmentation strategies for molecule language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1153–1162, 2024.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023.

Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Tianfan Fu, and Yue Zhao. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. *arXiv preprint arXiv:2411.15692*, 2024a.

Yuyan Liu, Sirui Ding, Sheng Zhou, Wenqi Fan, and Qiaoyu Tan. Moleculargpt: Open large language model (llm) for few-shot molecular property prediction. *arXiv preprint arXiv:2406.12950*, 2024b.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

Yvonne C Martin, James L Kofron, and Linda M Traphagen. Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry*, 45(19):4350–4358, 2002.

Andrew D McNaughton, Gautham Krishna Sankar Ramalaxmi, Agustin Kruel, Carter R Knutson, Rohith A Varikoti, and Neeraj Kumar. Cactus: Chemistry agent connecting tool usage to science. *ACS omega*, 9(46):46563–46573, 2024.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Odhran ODonoghue, Aleksandar Shtedritski, John Ginger, Ralph Abboud, Ali Essam Ghareeb, and Samuel G Rodriques. Bioplanner: Automatic evaluation of LLMs on protocol planning in biology. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL `https://openreview.net/forum?id=pMCRGmB7Rv`.

Juri Opitz and Sebastian Burst. Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*, 2019.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *arXiv preprint arXiv:2310.07276*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. Biodiscoveryagent: An ai agent for designing genetic perturbation experiments. *arXiv preprint arXiv:2405.17631*, 2024.

Rita Santos, Oleg Ursu, Anna Gaulton, A Patrícia Bento, Ramesh S Donadi, Cristian G Bologa, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I Oprea, et al. A comprehensive map of molecular drug targets. *Nature reviews Drug discovery*, 16(1):19–34, 2017.

Damian Szklarczyk, Alberto Santos, Christian Von Mering, Lars Juhl Jensen, Peer Bork, and Michael Kuhn. Stitch 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic acids research*, 44(D1):D380–D384, 2016.

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

Gary Tom, Stefan P Schmid, Sterling G Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M Rajaonson, Marta Skreta, et al. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024a.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 279–299, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.18. URL `https://aclanthology.org/2024.acl-long.18/`.

Shuangquan Wang, Huiyong Sun, Hui Liu, Dan Li, Youyong Li, and Tingjun Hou. Admet evaluation in drug discovery. 16. predicting herg blockers by combining multiple pharmacophores and machine learning approaches. *Molecular pharmaceutics*, 13(8):2855–2866, 2016.

Yilin Wen, Zifeng Wang, and Jimeng Sun. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*, 2023.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Junjie Xu, Zongyu Wu, Minhua Lin, Xiang Zhang, and Suhang Wang. Llm and gnn are complementary: Distilling llm for multimodal graph learning. *arXiv preprint arXiv:2406.01032*, 2024.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Youjun Xu, Ziwei Dai, Fangjin Chen, Shuaishi Gao, Jianfeng Pei, and Luhua Lai. Deep learning for drug-induced liver injury. *Journal of chemical information and modeling*, 55(10):2085–2093, 2015.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*, 2024.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862, 2022.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

Haiteng Zhao, Shengchao Liu, Ma Chang, Hannan Xu, Jie Fu, Zhihong Deng, Lingpeng Kong, and Qi Liu. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *Advances in Neural Information Processing Systems*, 36:5850–5887, 2023a.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023b.

Menglin Zheng, Satoshi Okawa, Miren Bravo, Fei Chen, María-Luz Martínez-Chantar, and Antonio Del Sol. Chempert: mapping between chemical perturbation and transcriptional response for non-cancer cells. *Nucleic Acids Research*, 51(D1):D877–D889, 2023.

Supplementary Material

- RAG-Enhanced Collaborative LLM Agents for Drug Discovery -

A Related Works **15**

B Preliminaries **15**

C Datasets **15**

C.1 Drug Toxicity Prediction Task 15

C.2 Property-Specific Molecular Captioning Task 16

C.3 Drug-Target Prediction Task Task 16

D Baselines Setup **16**

E Implementation Details **17**

E.1 Software Configuration 17

E.2 External Databases 17

E.3 KG Planner 18

F Additional Experimental Results **18**

F.1 Additional Ablation Studies 18

F.2 Additional External Knowledge Analysis 18

F.3 Additional Case Studies 21

G Agent Templates **22**

A RELATED WORKS

LLMs for Molecules. Leveraging the extensive body of literature and string-based molecular representations such as SMILES, language models (LMs) have been successfully applied to molecular sciences. Inspired by the masked language modeling approach used in BERT training (Devlin et al., 2018), KV-PLM (Zeng et al., 2022) introduces a method to train LMs by reconstructing masked SMILES and textual data. Similarly, MolT5 (Edwards et al., 2022) adopts the "replace corrupted spans" objective (Raffel et al., 2020) for pre-training on both SMILES strings and textual data, followed by fine-tuning for downstream tasks such as molecule captioning and generation. Building on this foundation, Pei et al. (2023) and Christofidellis et al. (2023) extend MolT5 with additional pre-training tasks, including protein FASTA reconstruction and chemical reaction prediction. Furthermore, GIMLET (Zhao et al., 2023a), Mol-Instructions (Fang et al., 2023), and MolecularGPT (Liu et al., 2024b) adopt instruction tuning (Zhang et al., 2023) to improve generalization across a wide range of molecular tasks. While these approaches demonstrate enhanced versatility, they still rely on expensive fine-tuning processes to enable molecule-specific tasks or to incorporate new data.

LLM Agents for Science. An LLM agent is a system that leverages LLMs to interact with users or other systems, perform tasks, and make decisions autonomously (Wang et al., 2024a). Recently, LLM agents have attracted significant interest in scientific applications and biomedical discovery (Gao et al., 2024), with applications including literature search (Lála et al., 2023), experiment design (Roohani et al., 2024), and hypothesis generation (Wang et al., 2024b), among others. In particular, agents focusing on drug discovery applications have emerged. Systems like ChemCrow (Bran et al., 2023), CACTUS (McNaughton et al., 2024), and Coscientist (Boiko et al., 2023) focus on automating cheminformatics tasks and experiments, streamlining computational and experimental pipelines. Other works leverage agent-based orchestration of tools and data to accelerate specific aspects of scientific workflows, such as search (ODonoghue et al., 2023) or design (Ghafarollahi & Buehler, 2024). In contrast to existing works, we investigate an agent-based framework that can effectively incorporate external knowledge to improve general molecular QA. This could be used either independently or as part of a larger system for automated drug discovery (Tom et al., 2024).

Multi-Agent Collaborations for Drug Discovery. Only a limited number of studies have explored multi-agent frameworks in the context of drug discovery. DrugAgent (Inoue et al., 2024) introduces a multi-agent framework integrating multiple external data sources but is limited to predicting drug-target interaction scores. Another study with the same name employs an agentic framework for automating machine learning programming for drug discovery tasks (Liu et al., 2024a). In contrast, our work seeks to tackle a diverse array of drug discovery tasks, enabling applicability across a wide variety of use cases.

B PRELIMINARIES

Tanimoto Similarity. The Tanimoto similarity is a widely accepted criterion for calculating the similarity between two molecules based on their molecular fingerprint Bajusz et al. (2015), which are the binary sequences that denote the presence or absence of specific substructures Rogers & Hahn (2010). Given two molecules g_i and g_j with fingerprints $\texttt{fp}_i$ and $\texttt{fp}_j$, the Tanimoto similarity $s_{i,j}$ is computed as follows:

$$s_{i,j} = \frac{|\texttt{fp}_i \cap \texttt{fp}_j|}{|\texttt{fp}_i| + |\texttt{fp}_j| - |\texttt{fp}_i \cap \texttt{fp}_j|}. \tag{2}$$

Intuitively, the Tanimoto similarity is the intersection-over-union of the sets of molecular substructures of both molecules.

C DATASETS

In this section, we provide further details on the datasets we used in Section 3. We provide a summary of data statistics in Table 5.

C.1 DRUG TOXICITY PREDICTION TASK

For the drug toxicity prediction task, we use four datasets: **hERG**, **DILI**, **Skin**, and **Carcinogens**.

Table 5: Data statistics.

	hERG	DILI	Skin	Carcinogens	BBBP	Sider	ClinTox	BACE	ChemPert	
									Overlap	No Overlap
# Molecules	648	475	404	278	2039	1427	1477	1513	7917	5771
# Tasks	1	1	1	1	1	27	2	1	2	2

- The Human ether-a-go-go related gene (**hERG**) Wang et al. (2016) plays a critical role in regulating the heart's rhythm. Thus, accurately predicting hERG liability is essential in drug discovery. In this task, we assess the model's ability to predict whether a drug blocks hERG.
- Drug-induced liver injury (**DILI**) Xu et al. (2015) is a severe liver condition caused by medications. In this task, we evaluate the model's capability to predict whether a drug is likely to cause liver injury.
- Repeated exposure to a chemical agent can trigger an immune response in inherently susceptible individuals, resulting in **Skin** Alves et al. (2015) sensitization. In this task, we evaluate the model's capability to predict whether the drug induces a skin reaction.
- A **Carcinogen** Lagunin et al. (2009) refers to any substance, radionuclide, or type of radiation that contributes to carcinogenesis, the development of cancer. In this task, we assess the model's ability to predict whether a drug has carcinogenic properties.

C.2 Property-Specific Molecular Captioning Task

For the property-specific molecular captioning task, we use four datasets in MoleculeNet Wu et al. (2018): **BBBP**, **Sider**, **Clintox**, **BACE**

- The blood-brain barrier penetration **(BBBP)** dataset consists of compounds categorized by their ability to penetrate the barrier, addressing a significant challenge in developing drugs targeting the central nervous system.
- The side effect resource **(Sider)** dataset organizes the side effects of approved drugs into 27 distinct organ system categories.
- The **Clintox** dataset includes two classification tasks: 1) predicting toxicity observed during clinical trials, and 2) determining FDA approval status.
- The **BACE** dataset provides qualitative binding results for a set of inhibitors aimed at human β-secretase 1.

C.3 Drug-Target Prediction Task Task

We rely on annotated molecular targets present in the Drug Repurposing Hub Corsello et al. (2017), DrugBank Wishart et al. (2018) and STITCH v5.0 Szklarczyk et al. (2016), as combined and preprocessed in Zheng et al., 2023. As we explained in Section 3, we separately report the performance on the test set for molecules based on their information availability in the external databases ("Overlap"/"No Overlap"). More specifically, for "No Overlap" cases, we exclude the molecules in the following criteria:

- We exclude the molecules if they exist in the knowledge graph.
- However, we noticed that many molecules have uninformative annotations, as also discussed in Section E. Consequently, we decided to exclude molecules from the test set only if they have sufficient annotations relevant to the task, as determined by GPT-4o mini.

After this process, 5771 molecules remained in the test set for "No Overlap" scenario.

D Baselines Setup

In this section, we provide further details on the baselines we used in Section 3. For all baseline models, we utilize the pre-trained checkpoints provided by the authors of the original papers.

Table 6: Links to baseline model checkpoints.

Model	URL
Galactica 125M	https://huggingface.co/facebook/galactica-125m
Galactica 1.3B	https://huggingface.co/facebook/galactica-1.3b
Galactica 6.7B	https://huggingface.co/facebook/galactica-6.7b
GIMLET	https://huggingface.co/haitengzhao/gimlet
LlasMol	https://huggingface.co/osunlp/LlaSMol-Mistral-7B
MolecularGPT	https://huggingface.co/YuyanLiu/MolecularGPT

- **Galactica** Taylor et al. (2022) is a large language model designed to store, integrate, and reason over scientific knowledge. The authors demonstrate Galactica's capabilities in simple molecule understanding tasks, such as predicting IUPAC names and performing binary classification for molecular property prediction. We also fine-tune Galactica for the Drug-Target Prediction task described in Section 3, using molecules linked to more than five proteins for activation and inhibition. For fine-tuning, we search for the optimal hyperparameters by experimenting with learning rates of $\{1e-3, 1e-4, 1e-5, 1e-6\}$ and epochs of $\{50, 100, 150, 200\}$, reporting the best performance achieved.
- **GIMLET** Zhao et al. (2023a) introduces a unified approach to leveraging language models for both graph and text data. The authors aim to enhance the generalization ability of language models for molecular property prediction through instruction tuning.
- **LlaSMol** Yu et al. (2024) presents a large-scale, comprehensive, and high-quality dataset designed for instruction tuning of large language models. This dataset includes tasks such as name conversion, molecule description, property prediction, and chemical reaction prediction, and it is used to fine-tune different open-source LLMs.

E Implementation Details

In this section, we provide further details on the implementation of CLADD.

E.1 Software Configuration

Our model is implemented using Python 3.11, PyTorch 2.5.1, Torch-Geometric 2.6.1, RDKit 2023.9.6, and LangGraph 0.2.59. The source code for CLADD is available at https://anonymous.4open.science/r/CLADD-0281.

E.2 External Databases

In all experiments, we employ the PubChem database Kim et al. (2021) as the annotation database $\mathcal{C}$ and PrimeKG Chandak et al. (2023) as the biological knowledge graph $\mathcal{G}$.

The **PubChem** database is one of the most extensive public molecular databases available. Pubchem database consists of multiple data sources, including DrugBank, CTD, PharmGKB, and more (https://pubchem.ncbi.nlm.nih.gov/sources/). The PubChem database used in this study includes 299K unique molecules and 336K textual descriptions associated with them (that is, a single molecule can have multiple captions sourced from different datasets associated with it). On average, each molecule has 1.115 descriptions, ranging from a minimum of one to a maximum of 17, as shown in Figure 3 (a). In this study, if a molecule had multiple captions, they were concatenated to form a single caption. On the other hand, as shown in Figure 3 (b), most captions consist of fewer than 20 words, underscoring the limited informativeness of human-generated captions. Even after concatenating multiple captions for each molecule, the majority still contain fewer than 50 words.

PrimeKG is a widely used knowledge graph for biochemical research. The knowledge graph contains 4,037,851 triplets and encompasses 10 entity types, including {anatomy, biological processes, cellular components, diseases, drugs, effects/phenotypes, exposures, genes/proteins, molecular functions, and pathways}. Additionally, it includes 18 relationship types: {associated with, carrier, contraindication, enzyme, expression

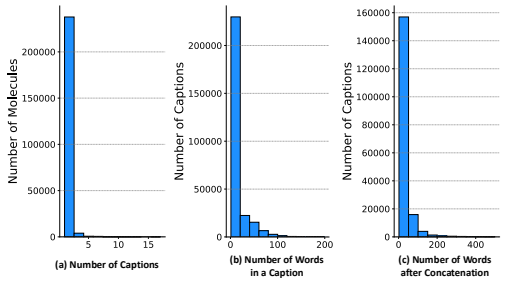

Figure 3: Data analysis on PubChem database.

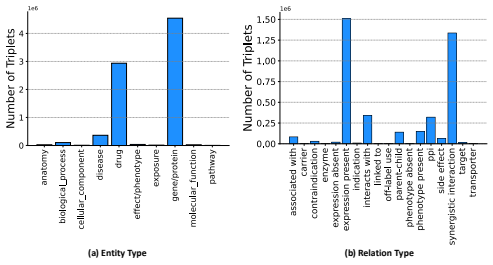

Figure 4: Data analysis on PrimeKG knowledge graph.

`absent, expression present, indication, interacts with, linked to, off-label use, parent-child, phenotype absent, phenotype present, ppi, side effect, synergistic interaction, target, and transporter`}. The number of triplets associated with each entity and relation type is shown in Figure 4 (a) and (b), respectively.

E.3 KG Planner

In section 2.2.1, we propose to utilize 3D geometrically pre-trained GNNs to retrieve molecules highly structurally similar to the query molecule. We use GIN architecture Xu et al. (2018), which is pre-trained with GraphMVP Liu et al. (2021) approach. The checkpoint of the model is available at [2].

F Additional Experimental Results

In this section, we provide additional experimental results that can supplement our experimental results in Section 3.

F.1 Additional Ablation Studies

In Table 7, we conduct a model analysis by removing one component of the model at a time. We have the following observations: **1)** By comparing "Only Expert Annotation" and "Only Generated Caption", we observe that relying solely on expert annotations yields significantly better performance. This highlights the critical importance of human-generated annotations over machine-generated captions. **2)** Among the three agents—DrugRel Agent, BioRel Agent, and MU Agent—we could not determine a clear superiority in their relative importance, as the agent leading to the best performance by itself varied by task (Activation or Inhibition). **3)** Overall, we observe a decline in performance when any single component of CLADD is removed, emphasizing the significance of each component.

Moreover, we perform additional ablation studies in the property-specific molecule captioning tasks in Figure 5. We observe that including all components (i.e., CLADD) leads to best performance except for the BACE dataset. This is because, as illustrated in Figure 7, the BACE dataset contains minimal relevant information in both the annotation database and the knowledge graph. Consequently, the model derives minimal benefit from external knowledge, highlighting the critical role of having relevant external information.

F.2 Additional External Knowledge Analysis

In Figure 6, we analyze how external knowledge is used during the decision-making process for drug-target prediction tasks. We have the following observations: **1)** As shown in Figure 6 (a) and (b), the average length of human descriptions is considerably longer in the "Correct" case, and

[2] `https://huggingface.co/chao1224/MoleculeSTM/tree/main/pretrained_GraphMVP`

Table 7: Additional ablation studies in protein target tasks (Precision @ 5). **Bold** and underline indicate best and second-best methods.

	(a) Overlap		(b) No overlap	
	Activate	Inhibit	Activate	Inhibit
No MolAnn Planner				
- Only Expert Annotation	2.99	4.80	2.63	3.20
- Only Generated Caption	2.72	3.96	2.61	2.80
No KG Planner	2.84	4.49	2.64	2.97
No DrugRel Agent	2.90	4.79	2.48	2.99
No BioRel Agent	2.96	4.50	2.63	3.00
No MU Agent	**3.04**	4.17	2.66	2.59
CLADD	**3.04**	**4.83**	**2.67**	**3.24**

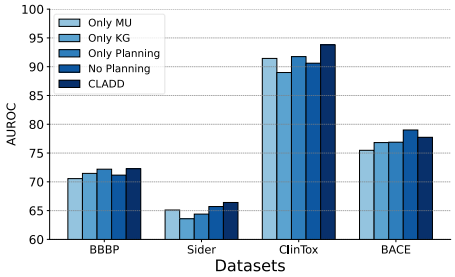

Figure 5: Ablation studies in the property-specific molecular captioning task.

the number of retrieved 2-hop paths is notably higher in the "Correct" case. This highlights the importance of having external information that is both high quality and abundant. **2)** On the other hand, although we anticipated a higher proportion of 2-hop paths containing Gene/Protein entities in the "Correct" case, no significant difference was observed between the "Correct" and "Incorrect" cases in Figure 6 (c) and (d). From these results, we argue that CLADD's performance is not solely reliant on retrieving external information that is directly linked to the correct answer, given that external information can be further processed and contextualized by the agents, also integrating different sources of evidence.

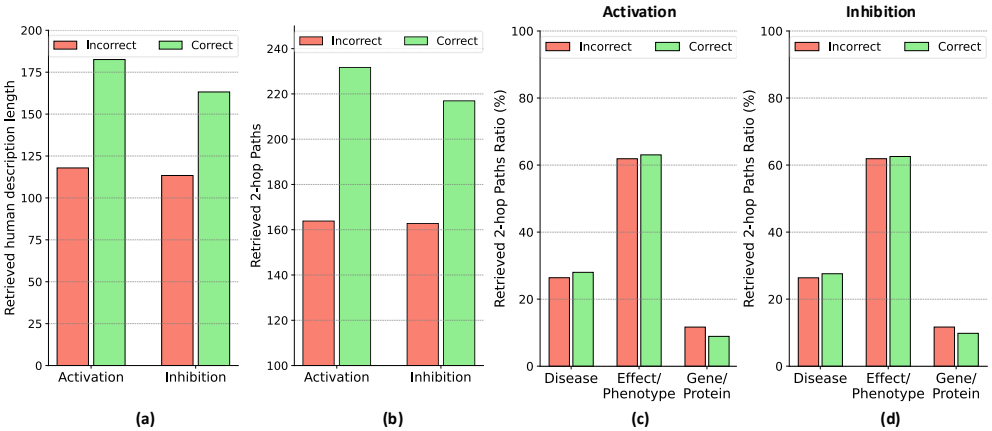

Figure 6: External knowledge analysis results. (a) The average length of retrieved human descriptions, (b) the average number of retrieved 2-hop paths in the knowledge graph, and (c) the proportion of entity types in 2-hop paths for correct and incorrect cases.

In Figure 7, we examine how the Planning Team determines the use of the captioning tool and collaborates with the Knowledge Graph Team based on the datasets. We observed that, in most cases, the KG was used for more than 50% of the query molecules, with the BACE and Skin Reaction datasets as significant exceptions. Furthermore, we observed that the BACE and hERG datasets lacked corresponding annotations for all query molecules.

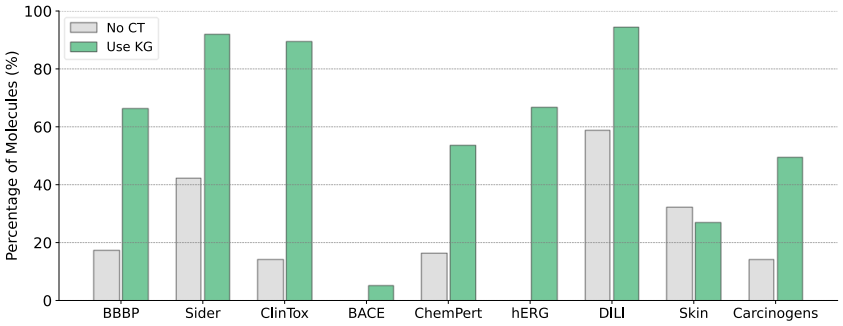

Figure 7: Planning team decision analysis based on different datasets. "No CT" signifies that the planning team has decided not to utilize the captioning tool, while "Use KG" indicates that the planning team intends to involve the Knowledge Graph Team.

F.3 ADDITIONAL CASE STUDIES

In this section, we provide additional case studies to analyze the behavior CLADD. In Figure 8, we observe that all three agents consistently predict dopamine-related and serotonin-related proteins as targets. Based on the reports, Prediction Agent prioritizes these proteins over Cytochrome P450-related enzymes in their predictions. Thus, we argue that our system efficiently prioritizes relevant information based on consensus, functioning similarly to a majority voting system.

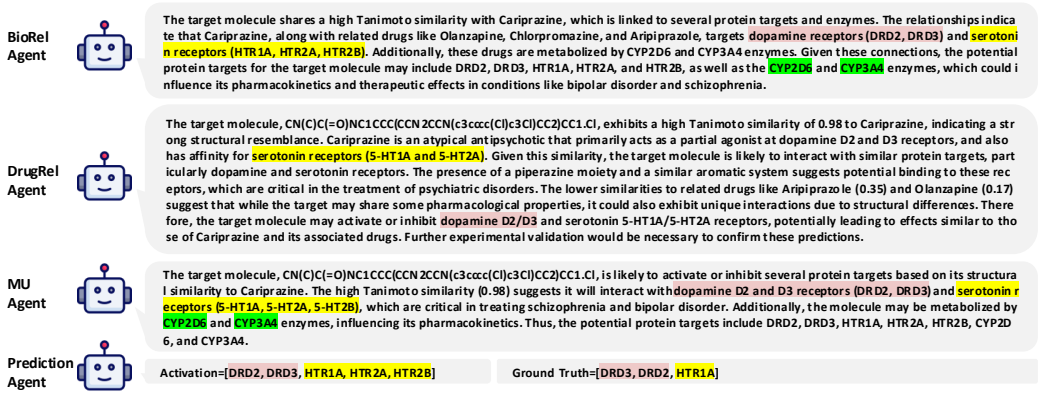

Figure 8: Additional case studies. Red represents dopamine-related proteins, yellow represents serotonin-related proteins, and green represents Cytochrome P450-related enzymes.

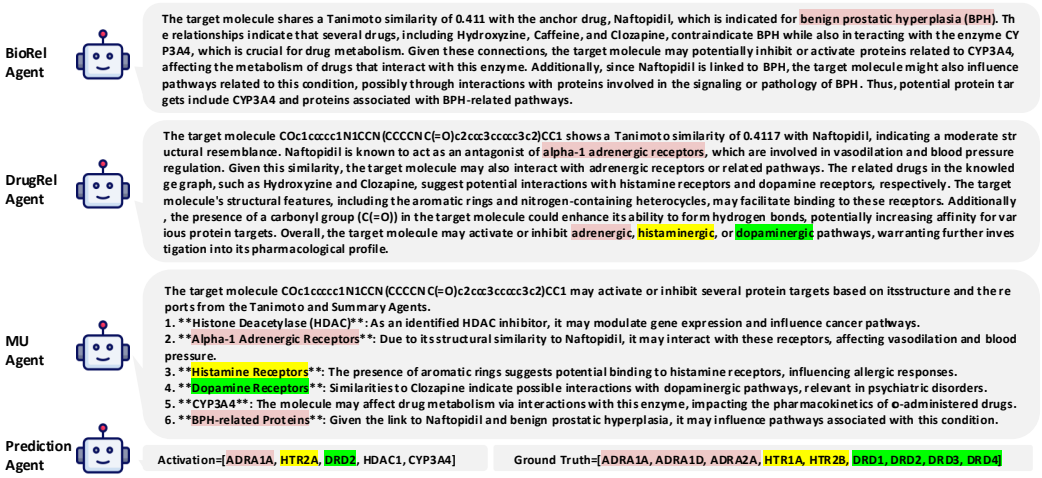

Figure 9: Full version of Figure 2.

G AGENT TEMPLATES

In this section, we provide the templates for each agent used in Section 2.

Table 8: Prompts for Molecule Annotation Planner (Section 2.2.1).

Prompt: You are now working as an excellent expert in chemistry and drug discovery.
Your task is to determine whether the provided description is enough for analyzing
the structure of the molecule.

Are you ready?

Description: {Retrieved Human Description}

You should answer in the following format:

Answer = YES or NO
REASON = YOUR REASON HERE

THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE.

Table 9: Prompts for Knowledge Graph Planner (Section 2.2.1).

Prompt: You are now working as an excellent expert in chemistry and drug discovery.
Your task is to decide whether to utilize the knowledge graph structure by evaluating the structural
similarity between the target molecule and the anchor drug within the knowledge graph.
If the target molecule and the anchor drug show high similarity, the knowledge graph should be
leveraged to extract relevant information.

The Tanimoto similarity between the target molecule {SMILES} and the anchor drug
{SMILES} ({Drug Name}) is {Tanimoto Similarity}.

You should answer in the following format:

Answer = YES or NO
REASON = YOUR REASON HERE

THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE.

Table 10: Prompts for Biology Relation Agent (Section 2.2.2).

Prompt: You are now working as an excellent expert in chemistry and drug discovery.
Your task is to predict {Task Description} by analyzing the relationships between the anchor drug,
which shares tanimoto similarity of {Tanimoto Similarity} with the target molecule,
and the most closely related drugs in the knowledge graph.

You should explain the reasoning based on the intermediate nodes between the
related drugs and the anchor drug, as well as the types of relationships they have.

The two-hop relationships between the drugs will be provided in the following format:
(Drug A, relation, Entity, relation, Drug B), where the entity can be one of the following
three types of entities: (gene/protein, effect/phenotype, disease)

Are you ready?

Target molecule: {SMILES}

Here are the two-hop relationships:
{Two-hop Paths}

DO NOT ANSWER IN THE PROVIDED FORMAT.
DO NOT WRITE MORE THAN 300 TOKENS.
THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE.

Table 11: Prompts for Drug Relation Agent (Section 2.2.2).

Prompt: You are now working as an excellent expert in chemistry and drug discovery.

Your task is to {Task Description} by analyzing its structural similarity to anchor drugs and related drugs, and provide an explanation grounded in its resemblance to these other drugs.

Are you ready?

The Tanimoto similarity between the target molecule {SMILES} and the anchor drug {SMILES} ({Drug Name} is {Tanimoto Similarity}.

The anchor drug {Drug Name} is highly associated with the following molecules in the knowledge graph: {Reference Drugs}.

The Tanimoto similarities between the target molecule {SMILES} and the related drugs in the knowledge graph are {Tanimoto Similarity}.

DO NOT WRITE MORE THAN 300 TOKENS.
THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE.

Table 12: Prompts for Molecule Understanding Agent (Section 2.2.3).

Prompt: You are now working as an excellent expert in chemistry and drug discovery.

Your task is to predict {Task Description} by using the SMILES representation and description of a molecule, and explain the reasoning based on its description.

You can also consider the report from other agents involved in drug discovery:
- Drug Relation Agent: Evaluates the structural similarity between the target molecule and related molecules.
- Biology Relation Agent: Examines the biological relationships among the related molecules.

Are you ready?
SMILES: {SMILES}
Description: {Caption}

Below is the report from other agents.
Drug Relation Agent:
{Report from Drug Relation Agent}

Biology Relation Agent:
{Report from Biology Relation Agent}

DO NOT WRITE MORE THAN 300 TOKENS.
THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE.

Table 13: Prompts for Prediction Agent (Section 2.2.4).

Prompt: You are now working as an excellent expert in chemistry and drug discovery.

Your task is to predict {Task Description} {SMILES}.

Your reasoning should be based on reports from various agents involved in drug discovery:
- Molecule Understanding Agent: Focuses on analyzing the structure of the target molecule.
- Drug Relation Agent: Evaluates the structural similarity between the target molecule and related molecules.
- Biology Relation Agent: Examines the biological relationships among the related molecules.

Below is the report from each agent.

Molecule Understanding Agent:
{Report from Molecule Understanding Agent}

Drug Relation Agent:
{Report from Drug Relation Agent}

Biology Relation Agent:
{Report from Biology Relation Agent}

Based on the reports, {Task Description and Answering Format}

THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE.