RAG-Enhanced Collaborative LLM Agents for Drug Discovery

Anonymous Author(s)
Affiliation
Address
email

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 limits the rapid integration of the vast amounts of scientific data continuously generated through experiments and research. Compounding these challenges is the fact that real-world scientific questions are typically complex and open-ended, requiring reasoning beyond pattern matching or static knowledge retrieval. To address 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. Our code is publicly available at https://anonymous.4open.science/r/CLADD-EEDE.

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 [1, 60]. In particular, their application in biomolecular studies has recently gained significant interest, motivated by the potential to profoundly accelerate scientific innovation and drug discovery applications [75, 51, 13]. LLMs provide novel ways to understand and reason about molecular data, building on the wealth of available scientific literature. Additionally, their reasoning and zero-shot abilities help overcome the limitations of task-specific deep learning models, streamlining data needs and improving human-AI collaboration [22, 73].

However, given the inherent complexity and specialized nature of the field, recent works emphasize the importance of domain-specific fine-tuning to boost tasks such as molecular captioning, property prediction, or binding affinity prediction [22, 13, 73, 19]. 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 [47, 78]

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

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. 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 [43], while not necessarily providing grounded answers [25]. Above all, real-world drug discovery questions are inherently complex, open-ended, and context-dependent, spanning heterogeneous data types [53]. As a consequence, static LLMs—either general-purpose or fine-tuned—may struggle to generalize to novel tasks or adapt to new evidence.

From this perspective, retrieval-augmented generation (RAG) methods offer a promising direction that enables dynamic adaptation of the model's knowledge without the need for continuous, expensive fine-tuning [24, 21]. 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 biochemical space [8] that renders exact retrieval ineffective. Second, biochemical data is extremely heterogeneous, spanning diverse modalities such as molecules, proteins, diseases, and complex relationships between them [62], which can also exist across multiple sources, introducing challenges in factual integration [28]. Finally, many real-world tasks are open-ended and require the LLM to extrapolate beyond the available external knowledge (which may also be ambiguous or partial [61]) while remaining grounded in it.

In this study, we tackle these challenges by introducing a **C**ollaborative framework of **LLM** **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 (KGs) that flexibly represent diverse biochemical entities and their relationships. CLADD is powered by general-purpose LLMs, while also integrating domain-specific LLMs, 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, offering a modular solution that can improve overall information processing [11]. In particular, CLADD includes a *Planning Team* to determine relevant data sources, a *Knowledge Graph Team* to retrieve external heterogeneous information in the KG 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 and open-ended 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 heterogeneous biochemical data without requiring fine-tuning, while addressing zero-shot and open-ended settings.
- We demonstrate the flexibility of the framework by tackling diverse applications, including drug-target prediction, property-specific molecular captioning, and biological activity prediction tasks.
- We provide comprehensive experimental results showcasing the effectiveness of CLADD compared to both 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*'. Such a general QA setup can be flexibly adapted to multi-class classification, captioning, and set-based predictions.

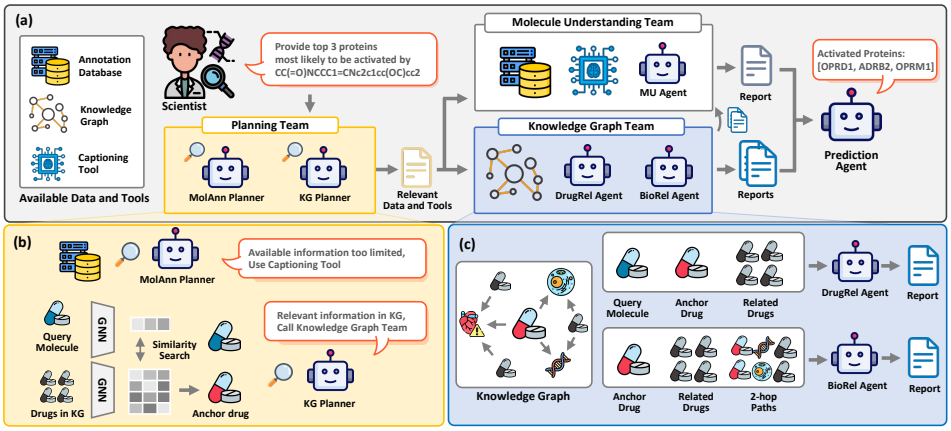

Figure 1: Overview of CLADD.

We assume access to two types of external databases: (1) molecular annotation databases $\mathcal{C}$, which include 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 heterogeneous entities $\mathcal{E}$ (such as drugs, proteins, and diseases) and a set of relations $\mathcal{R}$ connecting them. 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 [18, 50], 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 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 the relevance of external knowledge for a given query molecule. The team separately assesses the molecular annotations database and the knowledge graph through the MolAnn Planner and the KG Planner agents, respectively.

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 [73], they are often sparse, with many molecules entirely missing or lacking sufficient details due to the vastness of the chemical space [36].

To this end, the MolAnn Planner determines whether the retrieved annotations provide enough information for subsequent analyses. Specifically, given a query molecule g_q, retrieved annotations c_q, and the task instruction $\mathcal{I}$, the agent is invoked as follows:

$$o_{\text{MAP}} = \text{MolAnn Planner}(g_q, c_q, \mathcal{I}). \tag{1}$$

o_{MAP} indicates 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. While previous

works on general QA tasks focus on identifying entities in the knowledge graph that exactly match those in the query [3, 31], the vast chemical search space and the limited coverage of existing knowledge bases limit the effectiveness of such approaches in the field of drug discovery.

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 [44]. Specifically, we define the *anchor drug* g_a as the entity drug with the maximum cosine similarity between its embedding and that of the query molecule, among the set of all molecules in the KG ($g_{\mathcal{G}}$), $g_a = \underset{g \in g_{\mathcal{G}}}{\text{argmax}} \frac{emb(g_q) \cdot emb(g)}{\|emb(g_q)\| \|emb(g)\|}$, where *emb* is a representation produced by a graph neural network (GNN) pre-trained with 3D geometry [39], which 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:

$$o_{\text{KGP}} = \text{KG Planner}(g_q, g_a, s_{q,a}, \mathcal{I}), \tag{2}$$

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 contextual biological knowledge 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 [3, 66]. 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 [16]. 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 *semantically* 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 generates a report as follows:

$$o_{\text{DRA}} = \text{DrugRel Agent}\,(g_q, g_a, g_{r^1}, \ldots, g_{r^k}, \mathcal{T}, \mathcal{I}), \tag{3}$$

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. The agent leverages its internal knowledge about related drugs while effectively assessing the relatedness of the information to the target molecule based on the Tanimoto similarity.

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

Biological Relation (BioRel) Agent. The BioRel Agent summarizes how the anchor drug and the related drugs are biologically related, integrating additional biochemical entities present in the KG, such as targets, indications, side effects, etc. 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:

$$o_{\text{BRA}} = \text{BioRel Agent}(\mathcal{P}, \mathcal{I}, g_q, g_a, s_{q,a}). \tag{4}$$

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 entities in the KG for the specific task, they leverage distinct knowledge sources and perform different roles. Specifically, the BioRel Agent focuses on summarizing the network of relationships between drugs and other biological entities in the KG, contextualizing it with respect to the specific task at hand. In contrast, the DrugRel Agent primarily draws on its internal knowledge, triggered by the names of the 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

The Molecular Understanding (MU) Team compiles a report on the query molecule by leveraging external annotations and integrating them with structural information and reports from other agents.

Molecule Annotations. Annotations from the external database are retrieved for the query molecule, denoted as c_q. If the Planning Team decided 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:

$$\tilde{c}_q = \text{Captioning Tools}(g_q), \tag{5}$$

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 [50, 73], and can potentially include any tools, given that the output can be transformed into text.

Molecule Understanding (MU) Agent. The MU agent then analyzes the structure of the molecule, combining it with annotations and reports generated by the KG Team and generating a comprehensive report as follows:

$$o_{\text{MUA}} = \text{MU Agent}(g_q, c_q, o_{\text{DRA}}, o_{\text{BRA}}, \mathcal{I}). \tag{6}$$

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:

$$\mathcal{A}_{g_q} = \text{Task Agent}(g_q, o_{\text{MUA}}, o_{\text{DRA}}, o_{\text{BRA}}, \mathcal{I}). \tag{7}$$

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 an open-ended answer. Such behavior leverages the zero-shot capabilities of LLMs [34] and does not require additional fine-tuning. Therefore, a key advantage of CLADD is its flexibility, which enhances scientist-AI interactions.

3 Experiments

We assess the effectiveness of CLADD by conducting a range of drug discovery applications spanning different predictive tasks, including drug-target prediction (Section 3.1), property-specific molecular captioning (Section 3.2), and drug biological activity prediction (Section 3.3).

Implementation Details. In all experiments, we utilize GPT-4o mini through the OpenAI API for each agent. We use PrimeKG [12] as the KG, PubChem [32] as an annotation database, and MolT5 [18] as an external captioning tool. Additional implementation details and agent templates can be found in Appendix F and H, respectively.

Table 1: Performance in drug-target prediction 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**

Table 2: 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)**

3.1 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 [56, 6]. Here, we evaluate the models' ability to *accurately identify which proteins a given molecule is most likely to activate or inhibit* in a set prediction setting.

Datasets. We use molecular targets present in the Drug Repurposing Hub [15], DrugBank [67], and STITCH v5.0 [57], as preprocessed in Zheng et al. [79], including 13,688 molecules in total (details are presented in Appendix D).

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 [58] (details are presented in Appendix E).

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, and we calculate the precision with respect to ground truth data. 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 1 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 CLADD is confirmed for molecules not present in the caption database or knowledge graph (Table 1 (b)), showcasing CLADD's ability to leverage external knowledge to generalize to novel molecules. **3)** We observe that domain-specific fine-tuned models, such as Galactica, GIMLET, and MolecularGPT, *could not perform this task in a zero-shot setting* when prompted to do so, likely because this task is not included in 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.

3.2 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 [27, 19]. 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 task of interest.

Table 3: Performance in biological activity prediction task including (a) toxicity and (b) antibacterial activity (Macro-F1). **Avg.** indicates the average performance over toxicity datasets. * indicates whether the model always outputs the same response, either "Yes" or "No".

	(a) Toxicity				(b) MLSMR
	hERG	**DILI**	**Skin**	**Avg.**	**Mtb**
General LLMs					
GPT-4o mini	28.42	33.47	41.84	34.58	33.33*
GPT-4o	40.45	25.76	**54.51**	40.24	36.68
Domain LLMs					
Galactica 125M	40.78*	33.56	42.43	38.92	33.33*
Galactica 1.3B	48.57	34.37	42.43	41.79	33.33*
Galactica 6.7B	23.75*	57.67	40.41*	40.61	33.33*
GIMLET	36.50	35.51	42.28	38.09	39.81
LlasMol	23.75*	**61.20**	31.92	38.95	33.33*
CLADD	**51.46**	41.10	50.43	**47.66**	**50.92**

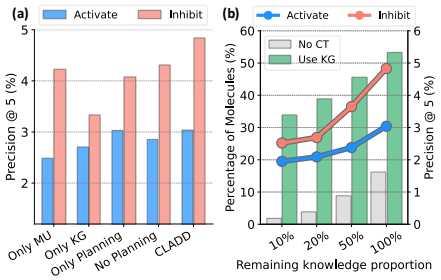

Figure 2: **Ablation studies. (a)** On model components. **(b)** On external knowledge.

Datasets. We leverage four widely recognized molecular property prediction datasets from the MoleculeNet benchmark [68]: **BBBP**, **Sider**, **ClinTox**, and **BACE** (further details in Appendix D).

Methods Compared. We consider different baseline approaches. First, we compare recent molecular captioning methods designed to generate general descriptions of molecules, including MolT5 [18], LlasMol [73], and BioT5 [50]. Furthermore, we assess general-purpose LLMs, namely GPT-4o mini and GPT-4o. Finally, we consider standard molecular property prediction baselines for references, including two GNNs pre-trained with different methodologies: GraphMVP [39] and MoleculeSTM [40]. We provide further details on the baseline models in Appendix E.

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 [69, 27, 19], we assess whether the generated captions can drive a classification model that categorizes molecules based on their properties. Specifically, we pose this evaluation as a molecular property prediction problem, and fine-tune a SciBERT model [7] on the generated caption concatenated to the SMILES representation to predict the property of interest. The "Only SMILES" model utilizes only the SMILES string as input for the SciBERT classifier. For baseline GNNs, each SMILES string is converted into a molecular graph. For all the experiments, we use a scaffold splitting strategy to simulate realistic distribution shifts, following previous work [40] (train/validation/test data split as 80/10/10%, with five independent runs). This evaluation protocol is further illustrated in Appendix D.2.

Experimental Results. Table 2 summarizes the results. **1)** While domain-specific models outperform general-purpose LLMs, their performance remains suboptimal, 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 [27, 19]. **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 CLADD to draw on external biochemical knowledge to ground its generation and its task-specificity. **3)** Moreover, CLADD consistently 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.

3.3 Biological Activity Prediction: Toxicity and Antibacterial Activity

Accurately predicting molecular bioactivity is a cornerstone of drug discovery, which is often hindered by the existence of countless biological contexts and sparse experimental data. We therefore explore the *zero-shot characterization of biological activity for unseen compounds*. To this goal, we focus on *drug toxicity* [5] and *antibacterial activity* [46] prediction.

Datasets. For drug toxicity prediction, we use three benchmark datasets: **hERG** [65], **DILI** [71], and **Skin** [2]. For antibacterial activity prediction, we use the dataset published in Eke et al. [20],

hereafter referred to as **MLSMR_Mtb**. In addition to its relevance, we selected MLSMR_Mtb for its recency, as it was *published after GPT-4o training and in parallel to the preparation of this study*, therefore avoiding the risk of pre-training data leakage. Dataset details are presented in Appendix D.

Methods Compared. We compare five domain-specific LLMs—Galactica 125M, Galactica 1.3B, Galactica 6.7B [58], LlasMol [73], and GIMLET [77], alongside two general-purpose LLMs, GPT-4o and GPT-4o mini (details in Appendix E).

Evaluation Protocol. Evaluation follows a zero-shot QA setting. The input includes a SMILES-based structural description of the molecule and the task description. Using the text-formatted output generated by each model, we compute the Macro-F1 score [49] as the evaluation metric.

Experimental Results. Table 3 summarizes the results. **1)** Both on toxicity datasets (average score) and the recently published antibacterial activity dataset, CLADD outperforms all the baselines. This highlights its ability to perform zero-shot predictions without domain-specific fine-tuning by effectively incorporating external knowledge into general-purpose LLMs at inference time. **2)** Notably, for three datasets (hERG, Skin and MLSMR_Mtb), several baseline models often output the same response, either “Yes" or “No", indicating their inability to perform the given task. In contrast, CLADD did not suffer from this limitation.

3.4 Ablation studies

Model Components Ablations. In Figure 2 (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 when only Molecular Understanding or Knowledge Graph team is available (“Only MU”, “Only KG”). **2)** *Dynamically selecting the relevant data sources with Planning Team improves 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*, as highlighted by the lower performance of “Only Planning” where all the relevant data sources are directly included in the prompt of a single Prediction Agent, bypassing intermediate reports. Additional ablation studies are presented in Appendix G.1. Furthermore, *we confirmed results across different LLMs, including open-source models*, showcasing the LLM-agnostic nature of CLADD in Appendix G.2.

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 2 (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*. A more detailed analysis of how external knowledge is utilized and its impact on model performance is provided in Appendix G.3.

4 Conclusion and Limitations

In this work, we introduced CLADD, a RAG-enhanced multi-agent framework for zero-shot molecular question-answering that can support various drug discovery tasks. We showcased its flexibility and effectiveness across multiple real-world tasks, outperforming both general-purpose and domain-specific fine-tuned LLMs. 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, fostering more interpretable scientist-AI interactions. While we focused on open-ended, set-based, and classification predictions, a limitation is the lack of focus on regression-based tasks, which would rely on the LLM’s ability to interpret assay details and numerical answers. Another limitation is the lack of uncertainty intervals, which could be tackled through recent orthogonal work. Beyond serving as a standalone tool, CLADD can also have a broader impact as a component of more complex agentic workflows, for example, combining computational and experimental systems [59], which will be the subject of future work.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Alves, V. M., Muratov, E., Fourches, D., Strickland, J., Kleinstreuer, N., Andrade, C. H., and Tropsha, A. 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.

[3] Baek, J., Aji, A. F., and Saffari, A. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *arXiv preprint arXiv:2306.04136*, 2023.

[4] Bajusz, D., Rácz, A., and Héberger, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7:1–13, 2015.

[5] Basile, A. O., Yahi, A., and Tatonetti, N. P. Artificial intelligence for drug toxicity and safety. *Trends in pharmacological sciences*, 40(9):624–635, 2019.

[6] Batool, M., Ahmad, B., and Choi, S. A structure-based drug discovery paradigm. *International journal of molecular sciences*, 20(11):2783, 2019.

[7] Beltagy, I., Lo, K., and Cohan, A. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

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

[9] Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

[10] Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

[11] Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. 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`.

[12] Chandak, P., Huang, K., and Zitnik, M. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023.

[13] Chaves, J. M. Z., Wang, E., Tu, T., Vaishnav, E. D., Lee, B., Mahdavi, S. S., Semturs, C., Fleet, D., Natarajan, V., and Azizi, S. Tx-llm: A large language model for therapeutics. *arXiv preprint arXiv:2406.06316*, 2024.

[14] Christofidellis, D., Giannone, G., Born, J., Winther, O., Laino, T., and Manica, M. Unifying molecular and textual representations via multi-task language modelling. *arXiv preprint arXiv:2301.12586*, 2023.

[15] Corsello, S. M., Bittker, J. A., Liu, Z., Gould, J., McCarren, P., Hirschman, J. E., Johnston, S. E., Vrcic, A., Wong, B., Khan, M., et al. The drug repurposing hub: a next-generation drug library and information resource. *Nature medicine*, 23(4):405–408, 2017.

[16] De Azambuja, E., Holmes, A. P., Piccart-Gebhart, M., Holmes, E., Di Cosimo, S., Swaby, R. F., Untch, M., Jackisch, C., Lang, I., Smith, I., 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.

[17] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] Edwards, C., Lai, T., Ros, K., Honke, G., Cho, K., and Ji, H. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

[19] Edwards, C., Lu, Z., Hajiramezanali, E., Biancalani, T., Ji, H., and Scalia, G. Molcap-arena: A comprehensive captioning benchmark on language-enhanced molecular property prediction. *arXiv preprint arXiv:2411.00737*, 2024.

[20] Eke, I. E., Williams, J. T., and Abramovitch, R. B. Genetic and cheminformatic characterization of mycobacterium tuberculosis inhibitors discovered in the molecular libraries small molecule repository. *ACS Infectious Diseases*, 11(4):882–893, 2025.

[21] Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., and Li, Q. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6491–6501, 2024.

[22] Fang, Y., Liang, X., Zhang, N., Liu, K., Huang, R., Chen, Z., Fan, X., and Chen, H. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*, 2023.

[23] Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J. R., Ektefaie, Y., Kondic, J., and Zitnik, M. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151, 2024.

[24] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[25] Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R., and Herzig, J. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7765–7784. Association for Computational Linguistics.

[26] Ghafarollahi, A. and Buehler, M. J. Protagents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 2024.

[27] Guo, H., Zhao, S., Wang, H., Du, Y., and Qin, B. 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.

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

[29] Inoue, Y., Song, T., and Fu, T. Drugagent: Explainable drug repurposing agent with large language model-based reasoning. *arXiv preprint arXiv:2408.13378*, 2024.

[30] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38, 2023.

[31] Jiang, J., Zhou, K., Dong, Z., Ye, K., Zhao, X., and Wen, J.-R. StructGPT: A general framework for large language model to reason over structured data. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9237–9251. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.574. URL `https://aclanthology.org/2023.emnlp-main.574/`.

[32] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., et al. Pubchem in 2021: new data content and improved web interfaces. *Nucleic acids research*, 49(D1):D1388–D1395, 2021.

[33] Klotsman, M., Weinberg, C., Davis, K., Binnie, C., and Hartmann, K. 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.

[34] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[35] Lála, J., O'Donoghue, O., Shtedritski, A., Cox, S., Rodriques, S. G., and White, A. D. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

[36] Lee, N., Laghuvarapu, S., Park, C., and Sun, J. 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.

[37] Li, D., Yang, S., Tan, Z., Baik, J. Y., Yun, S., Lee, J., Chacko, A., Hou, B., Duong-Tran, D., Ding, Y., et al. Dalk: Dynamic co-augmentation of llms and kg to answer alzheimer's disease questions with scientific literature. *arXiv preprint arXiv:2405.04819*, 2024.

[38] Li, J., Liu, Y., Fan, W., Wei, X.-Y., Liu, H., Tang, J., and Li, Q. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE transactions on knowledge and data engineering*, 2024.

[39] Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021.

[40] Liu, S., Nie, W., Wang, C., Lu, J., Qiao, Z., Liu, L., Tang, J., Xiao, C., and Anandkumar, A. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023.

[41] Liu, S., Lu, Y., Chen, S., Hu, X., Zhao, J., Fu, T., and Zhao, Y. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. *arXiv preprint arXiv:2411.15692*, 2024.

[42] Liu, Y., Ding, S., Zhou, S., Fan, W., and Tan, Q. Moleculargpt: Open large language model (llm) for few-shot molecular property prediction. *arXiv preprint arXiv:2406.12950*, 2024.

[43] Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

[44] Martin, Y. C., Kofron, J. L., and Traphagen, L. M. Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry*, 45(19):4350–4358, 2002.

[45] McNaughton, A. D., Sankar Ramalaxmi, G. K., Kruel, A., Knutson, C. R., Varikoti, R. A., and Kumar, N. Cactus: Chemistry agent connecting tool usage to science. *ACS omega*, 9(46): 46563–46573, 2024.

[46] Melo, M. C., Maasch, J. R., and de la Fuente-Nunez, C. Accelerating antibiotic discovery through artificial intelligence. *Communications biology*, 4(1):1050, 2021.

[47] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[48] ODonoghue, O., Shtedritski, A., Ginger, J., Abboud, R., Ghareeb, A. E., and Rodriques, S. G. 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`.

[49] Opitz, J. and Burst, S. Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*, 2019.

[50] Pei, Q., Zhang, W., Zhu, J., Wu, K., Gao, K., Wu, L., Xia, Y., and Yan, R. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *arXiv preprint arXiv:2310.07276*, 2023.

[51] Pei, Q., Wu, L., Gao, K., Zhu, J., Wang, Y., Wang, Z., Qin, T., and Yan, R. Leveraging biomolecule and natural language through multi-modal learning: A survey. *arXiv preprint arXiv:2403.01528*, 2024.

[52] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[53] Ramos, M. C., Collison, C. J., and White, A. D. A review of large language models and autonomous agents in chemistry. *Chemical Science*, 2025.

[54] Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[55] Roohani, Y., Lee, A., Huang, Q., Vora, J., Steinhart, Z., Huang, K., Marson, A., Liang, P., and Leskovec, J. Biodiscoveryagent: An ai agent for designing genetic perturbation experiments. *arXiv preprint arXiv:2405.17631*, 2024.

[56] Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T. I., et al. A comprehensive map of molecular drug targets. *Nature reviews Drug discovery*, 16(1):19–34, 2017.

[57] Szklarczyk, D., Santos, A., Von Mering, C., Jensen, L. J., Bork, P., and Kuhn, M. Stitch 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic acids research*, 44(D1):D380–D384, 2016.

[58] Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[59] Tom, G., Schmid, S. P., Baird, S. G., Cao, Y., Darvish, K., Hao, H., Lo, S., Pablo-García, S., Rajaonson, E. M., Skreta, M., et al. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732, 2024.

[60] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[61] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

[62] Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972): 47–60, 2023.

[63] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[64] Wang, Q., Downey, D., Ji, H., and Hope, T. SciMON: Scientific inspiration machines optimized for novelty. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 279–299, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.18. URL https://aclanthology.org/2024.acl-long.18/.

[65] Wang, S., Sun, H., Liu, H., Li, D., Li, Y., and Hou, T. Admet evaluation in drug discovery. 16. predicting herg blockers by combining multiple pharmacophores and machine learning approaches. *Molecular pharmaceutics*, 13(8):2855–2866, 2016.

[66] Wen, Y., Wang, Z., and Sun, J. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*, 2023.

[67] Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.

[68] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9 (2):513–530, 2018.

[69] Xu, J., Wu, Z., Lin, M., Zhang, X., and Wang, S. Llm and gnn are complementary: Distilling llm for multimodal graph learning. *arXiv preprint arXiv:2406.01032*, 2024.

[70] Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[71] Xu, Y., Dai, Z., Chen, F., Gao, S., Pei, J., and Lai, L. Deep learning for drug-induced liver injury. *Journal of chemical information and modeling*, 55(10):2085–2093, 2015.

[72] Yang, R., Liu, H., Marrese-Taylor, E., Zeng, Q., Ke, Y. H., Li, W., Cheng, L., Chen, Q., Caverlee, J., Matsuo, Y., et al. Kg-rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques. *arXiv preprint arXiv:2403.05881*, 2024.

[73] Yu, B., Baker, F. N., Chen, Z., Ning, X., and Sun, H. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*, 2024.

[74] Zeng, Z., Yao, Y., Liu, Z., and Sun, M. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862, 2022.

[75] Zhang, Q., Ding, K., Lyv, T., Wang, X., Yin, Q., Zhang, Y., Yu, J., Wang, Y., Li, X., Xiang, Z., et al. Scientific large language models: A survey on biological & chemical domains. *arXiv preprint arXiv:2401.14656*, 2024.

[76] Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

[77] Zhao, H., Liu, S., Chang, M., Xu, H., Fu, J., Deng, Z., Kong, L., and Liu, Q. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *Advances in Neural Information Processing Systems*, 36:5850–5887, 2023.

[78] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[79] Zheng, M., Okawa, S., Bravo, M., Chen, F., Martínez-Chantar, M.-L., and Del Sol, A. Chempert: mapping between chemical perturbation and transcriptional response for non-cancer cells. *Nucleic Acids Research*, 51(D1):D877–D889, 2023.

Supplementary Material for RAG-Enhanced Collaborative LLM Agents for Drug Discovery

This is an appendix for the paper **RAG-Enhanced Collaborative LLM Agents for Drug Discovery**.

A Related Work

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 [17], KV-PLM [74] introduces a method to train LMs by reconstructing masked SMILES and textual data. Similarly, MolT5 [18] adopts the "replace corrupted spans" objective [52] 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. [50] and Christofidellis et al. [14] extend MolT5 with additional pre-training tasks, including protein FASTA reconstruction and chemical reaction prediction. Furthermore, GIMLET [77], Mol-Instructions [22], and MolecularGPT [42] adopt instruction tuning [76] 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 [63]. Recently, LLM agents have attracted significant interest in scientific applications and biomedical discovery [23], with applications including literature search [35], experiment design [55], and hypothesis generation [64], among others. In particular, agents focusing on drug discovery applications have emerged. Systems like ChemCrow [10], CACTUS [45], and Coscientist [9] 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 [48] or design [26]. In contrast to existing works, we investigate an agent-based framework that can effectively incorporate external knowledge to improve open-ended and zero-shot molecular QA. This could be used either independently or as part of a larger system for automated drug discovery [59].

Multi-Agent Collaborations for Drug Discovery. Only a limited number of studies have explored multi-agent frameworks in the context of drug discovery. DrugAgent [29] 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 [41]. In contrast, our work seeks to tackle a diverse array of drug discovery tasks, grounding the agent capabilities in external knowledge.

LLMs with Knowledge Graphs. While large language models (LLMs) have been successfully adapted to numerous domains, they have faced criticism for their lack of factual accuracy. Specifically, LLMs often struggle to recall reliable facts and are prone to hallucinations [30], which can be a bottleneck for scientific applications, and are still persistent after fine-tuning [25]. A promising approach to mitigate these issues is the integration of external knowledge sources, such as knowledge graphs (KGs), into LLMs during the generation process. For instance, Baek et al. [3] proposes a method where relevant triplets are retrieved from KGs based on the input query. These triplets are then verbalized and provided as additional input to the LLM, enhancing its factual grounding and accuracy. KG-Rank [72] focuses on medical question-answering, leveraging a medical knowledge graph to match terms in the question and expand them. DALK [37] leverages an LLM to construct an Alzheimer's disease-specific KG, which is then used to enhance the accuracy and relevance of LLM-generated responses. Although these methods retrieve entities from KGs that are related to those in the query, the virtually infinite number of potential molecules of interest in drug discovery, combined with the limited domain expertise of general-purpose LLMs, makes it challenging to directly apply existing techniques to molecular question-answering.

B Additional Related Works

LLMs with Knowledge Graphs. While large language models (LLMs) have been successfully adapted to numerous domains, they have faced criticism for their lack of factual accuracy. Specifically, LLMs often struggle to recall reliable facts and are prone to hallucinations [30], which can be a bottleneck for scientific applications, and are still persistent after fine-tuning [25]. A promising

approach to mitigate these issues is the integration of external knowledge sources, such as knowledge graphs (KGs), into LLMs during the generation process. For instance, Baek et al. [3] proposes a method where relevant triplets are retrieved from KGs based on the input query. These triplets are then verbalized and provided as additional input to the LLM, enhancing its factual grounding and accuracy. KG-Rank [72] focuses on medical question-answering, leveraging a medical knowledge graph to match terms in the question and expand them. DALK [37] leverages an LLM to construct an Alzheimer's disease-specific KG, which is then used to enhance the accuracy and relevance of LLM-generated responses. Although these methods retrieve entities from KGs that are related to those in the query, the virtually infinite number of potential molecules of interest in drug discovery, combined with the limited domain expertise of general-purpose LLMs, makes it challenging to directly apply existing techniques to molecular question-answering.

C Preliminaries

Tanimoto Similarity. The Tanimoto similarity is a widely accepted criterion for calculating the similarity between two molecules based on their molecular fingerprint [4], which are the binary sequences that denote the presence or absence of specific substructures [54]. 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:

$$s_{i,j} = \frac{|\texttt{fp}_i \cap \texttt{fp}_j|}{|\texttt{fp}_i| + |\texttt{fp}_j| - |\texttt{fp}_i \cap \texttt{fp}_j|}. \tag{8}$$

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

D 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 4.

Table 4: Data statistics.

	hERG	**DILI**	**Skin**	**MLSMR_Mtb**	**BBBP**	**Sider**	**ClinTox**	**BACE**	**ChemPert**	
									Overlap	No Overlap
# Molecules	648	475	404	200	2039	1427	1477	1513	7917	5771
# Tasks	1	1	1	1	1	27	2	1	2	2

D.1 Drug Biological Activity Prediction Task

For the drug biological activity prediction task, we use four datasets: **hERG**, **DILI**, **Skin**, and **MLSMR_Mtb**.

- The Human ether-a-go-go related gene (**hERG**) [65] 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**) [71] 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** [2] sensitization. In this task, we evaluate the model's ability to predict whether the drug induces a skin reaction.
- The Molecular Libraries Small Molecule Repository - *Mycobacterium tuberculosis* dataset (**MLSMR_Mtb**) has been released as part of Eke et al. [20]. Antimycobacterial activity against *M. tuberculosis* was measured in a dose-response assay and quantified as AUC. Following the original study, we used an AUC cutoff of 25 for classification. Out of the 935 molecules tested, we randomly selected 200 compounds with a balanced positive/negative ratio. For this task, we evaluate the model's ability to predict antimycobacterial activity. In addition to its relevance, we selected this dataset for its recency, as **it was published after GPT-4o and in parallel to the preparation of this study**, ensuring no overlap with pre-training data and thus allowing benchmarking against leakage risks. To the best of our knowledge, our work is the first study leveraging this dataset.

D.2 Property-Specific Molecular Captioning Task

For the property-specific molecular captioning task, we use four datasets in MoleculeNet [68]: **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.

Evaluation Protocol. While previous works on molecular captioning generate general molecule descriptions and evaluate them with standard NLP metrics like BLEU. However, because a molecule can be described in multiple ways (some more relevant to certain tasks [27, 19]), we focus on property-specific captioning. Here, the main challenge is the lack of ground-truth captions for each property. Therefore, similar to previous work [19], we use an evaluation protocol that checks how well the generated captions aid in property prediction by fine-tuning a language model (SciBERT) on them. 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 [7] model for property prediction. Importantly, **fine-tuning SciBERT is only part of the evaluation protocol, as CLADD itself does not involve any fine-tuning**. This process is illustrated in Figure 3.

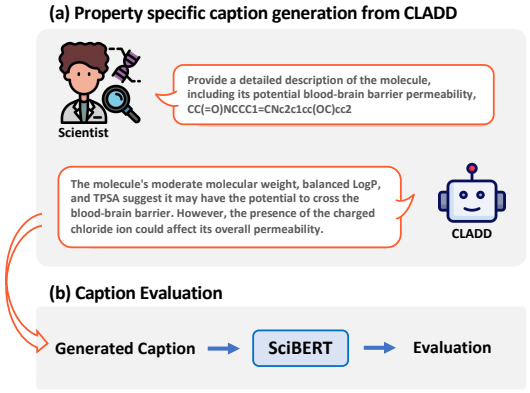

Figure 3: (a) After CLADD (or baseline models) generates a property-specific caption, (b) SciBERT is used for evaluation. In other words, **fine-tuning SciBERT is not part of CLADD**; it is only used for evaluation purposes.

D.3 Drug-Target Prediction Task

We rely on annotated molecular targets present in the Drug Repurposing Hub [15], DrugBank [67], and STITCH v5.0 [57], as combined and preprocessed in 79. 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 F. 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 the "No Overlap" scenario.

E Baselines Setup

This section provides 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 5: 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** [58] 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 and associated activated/inhibited proteins. For fine-tuning, we searched for the optimal hyperparameters (learning rate of $\{1e-3, 1e-4, 1e-5, 1e-6\}$ and epoch number of $\{50, 100, 150, 200\}$), reporting the best performance achieved.
- **GIMLET** [77] 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** [73] 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.

F Implementation Details

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

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.

Computational Resources. For LLMs, we utilize the OpenAI API, thereby leveraging OpenAI's computational resources. All other computations, such as GNN retrievers, are performed on a 24GB NVIDIA GeForce RTX 3090 GPU.

External Databases. In all experiments, we employ the PubChem database [32] as the annotation database $\mathcal{C}$ and PrimeKG [12] 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 4 (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 4 (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,`

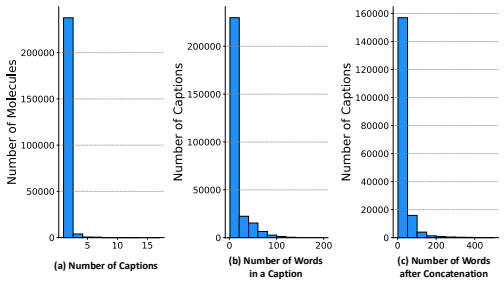

Figure 4: Data analysis on PubChem database.

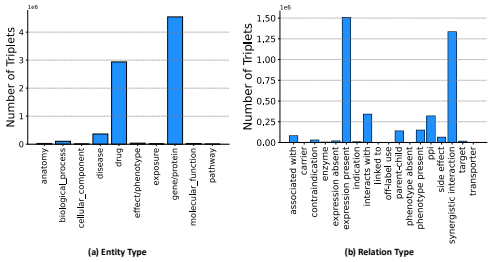

Figure 5: Data analysis on PrimeKG knowledge graph.

enzyme, expression 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 5 (a) and (b), respectively.

F.1 KG Planner

As explained in section 2.2.1, we utilize a pre-trained GNN (with 3D information) to retrieve molecules highly related to the query molecule. In particular, the model has a GIN architecture [70], which is pre-trained with the GraphMVP [39] approach. The checkpoint of the model is available at `https://huggingface.co/chao1224/MoleculeSTM/tree/main/pretrained_GraphMVP`.

G Additional Experimental Results

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

G.1 Additional Ablation Studies

In Table 6, we conduct a model analysis by removing one component of the model at a time for the drug-target prediction task. 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. Still, their combination leads to the best overall performance. **2)** Among the three agents—DrugRel Agent, BioRel Agent, and MU Agent—we could not determine a clear superiority in their relative importance, as it was task-dependent (Activation or Inhibition). **3)** Overall, we observe a decline in performance when any single component of CLADD is removed, emphasizing the significance of each module.

We perform additional ablation studies in the property-specific molecule captioning task in Figure 6. Similarly, we observe that including all components (i.e., CLADD) leads to the best performance except for the BACE dataset. Our analysis showed that this is because, as illustrated in Figure 10, 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 to boost performance.

G.2 LLM-Agnostic Nature of CLADD

Due to the expensive API costs, we mainly report the results using GPT-4o mini in the main manuscript to validate the proposed framework. In this section, we performed additional experiments replacing it with different LLMs, including Llama3.3-70b and DeepSeek-V3. As shown in Table 7, the proposed framework (+ CLADD) consistently improves each individual LLM, showcasing its LLM-agnostic advantage.

Table 6: Additional ablation studies in drug-target prediction task (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**

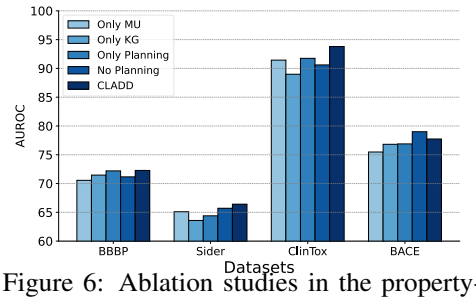

Figure 6: Ablation studies in the property-specific molecular captioning task.

Table 7: Performance of CLADD with each agent replaced by a different closed-source and open-source LLM (drug-target prediction task).

	Overlap		No overlap	
	Activate	**Inhibit**	**Activate**	**Inhibit**
GPT-4o mini	1.15	1.02	1.13	0.87
+ CLADD	**3.04**	**4.83**	**2.67**	**3.24**
Llama-3.3-70B	0.84	0.94	0.88	0.81
+ CLADD	**3.13**	**6.40**	**2.73**	**4.14**
DeepSeek-V3	1.91	1.46	1.90	1.11
+ CLADD	**3.60**	**7.75**	**3.15**	**5.01**

G.3 Additional External Knowledge Analysis

In Table 7, we analyze how the retrieval accuracy affects the model performance. To do so, we investigated two settings: one where the anchor drug selection in the knowledge graph is done randomly, and another where annotations are randomly sampled from the annotation database. As expected, we observe that the performance of both these models is significantly lower compared to the original model. We also observe that there is still a significant performance gap when compared to GPT-4o mini. This is expected, as our model still includes a planning team that ensures that the anchor drug and annotations are only used when they are relevant to the query molecule and task.

Moreover, we further investigate how the quality of retrieved knowledge affects the model performance. Firstly, we analyzed how performance changes as a function of the length of the annotation retrieved from the annotation database. In Figure 8(a), “Zero" indicates that no annotation is available in the annotation database, while Q1, Q2, Q3, and Q4 represent the quartiles of the retrieved annotation length. We highlight two interesting trends: (1) in general, performance increases with the annotation length, which is in line with the intuition that longer annotations include more relevant information, and (2) on average, “no annotation” leads to better results than the shortest annotations, which could indicate that the shortest annotations are often not informative enough to boost performance. However, for all groups except the shortest annotations, the additional information provides a proportional improvement.

Secondly, we analyzed how performance changes as a function of the similarity between the query molecule and the anchor molecule in the knowledge graph. In Figure 8 (b), Molecules with a Tanimoto similarity of 1 are excluded from the evaluation. “High": Tanimoto similarity between 0.7$\sim$1.0, “Middle": Tanimoto similarity between 0.3$\sim$0.7, “Low": Tanimoto similarity between 0.0$\sim$0.3. Here, we found a very positive correlation, which is in line with the intuition that a higher similarity provides more relevant contextual information.

In Figure 9, we analyze how external knowledge is used during the decision-making process for the drug-target prediction task. We have the following observations: **1)** As shown in Figures 9(a) and 9(b), the average length of human descriptions is considerably longer in the “Correct" case, and 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 Figures 9(c)

Figure 7: Performance analysis on retrieval errors.

	Overlap		No overlap	
	Activate	**Inhibit**	**Activate**	**Inhibit**
GPT-4o mini	1.15	1.02	1.13	0.87
Random Anchor Drug in KG	2.49	4.46	1.86	2.31
Random Annotations in DB	2.62	4.08	2.51	2.85
CLADD	**3.04**	**4.83**	**2.67**	**3.24**

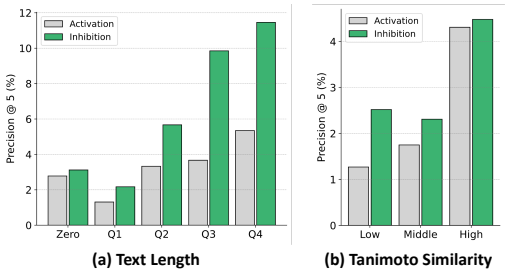

Figure 8: Performance as a function of (a) the length of the text retrieved from the annotation database and (b) the Tanimoto similarity between the anchor molecule and the knowledge graph.

and 9(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, integrating different sources of evidence and internal knowledge.

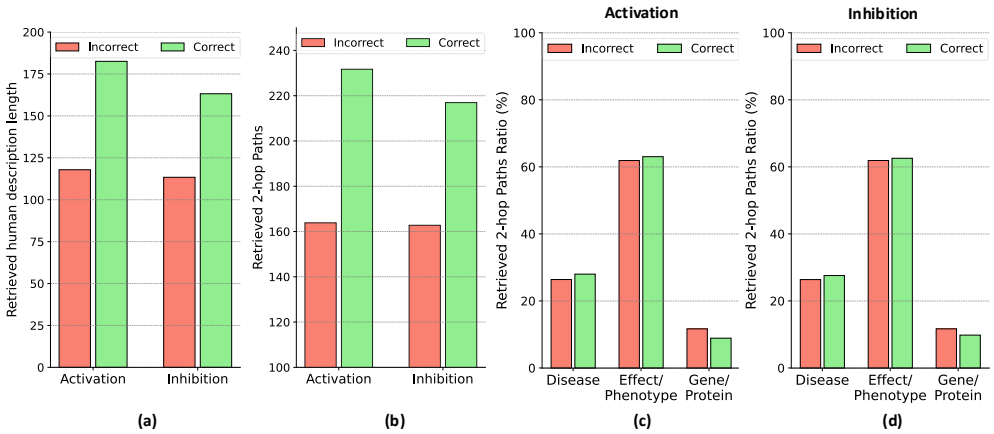

Figure 9: External knowledge analysis results. (a) The average length of retrieved human descriptions, (b) the average number of retrieved 2-hop paths in the knowledge graph, and (c-d) the proportion of entity types in 2-hop paths for correct and incorrect cases.

In Figure 10, 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.

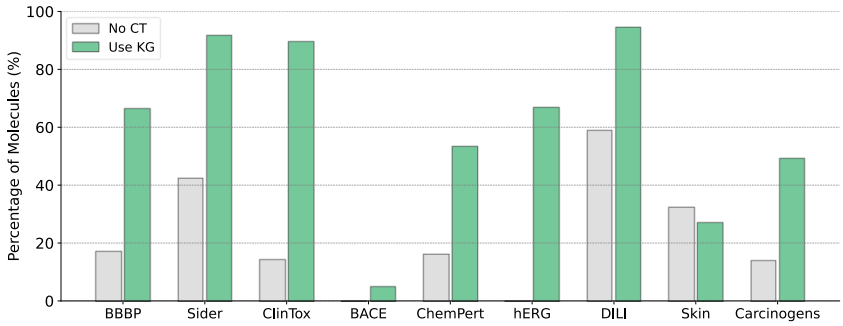

Figure 10: Planning team decision analysis based on different datasets. “No CT" signifies that the planning team has decided not to utilize the captioning tool, while “Use KG" indicates that the planning team intends to involve the Knowledge Graph Team.

G.4 Case Studies

Figure 11 showcases how the agents in CLADD collaborate to identify "the top-5 protein targets 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), implying the activation of related pathways. The DrugRel Agent complements these findings by **1)** linking BPH to alpha-1 adrenergic receptors using its internal knowledge (which is confirmed in the literature [33]), and **2)** analyzing related drugs in the knowledge graph (*e.g.,* Hydroxyzine, Clopamine), to infer interaction with histamine and dopamine receptors. Finally, the MU agent integrates these findings with the analysis of the molecular structure to provide a summarized report of the activated protein targets. This example highlights the agents' complementary strengths, which lead to interpretable and reliable predictions.

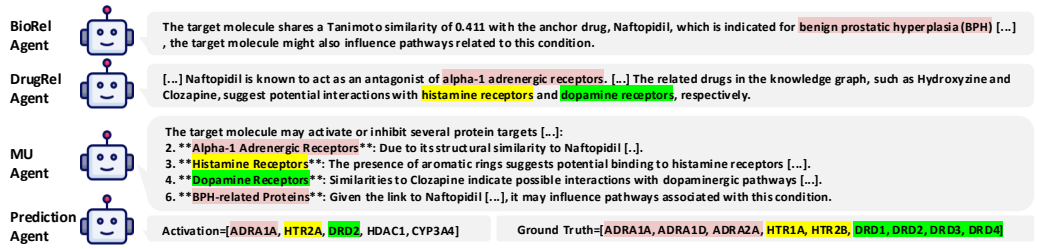

Figure 11: **Example of collaboration between agents in CLADD** (on the drug-target prediction task). Red represents adrenergic receptors, yellow represents histamine receptors, and green represents dopamine receptors. The full version is available in Appendix G.

Moreover, in Figure 12, 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 the prediction. Thus, we argue that our system can efficiently prioritize relevant information based on consensus, functioning similarly to a majority voting system.

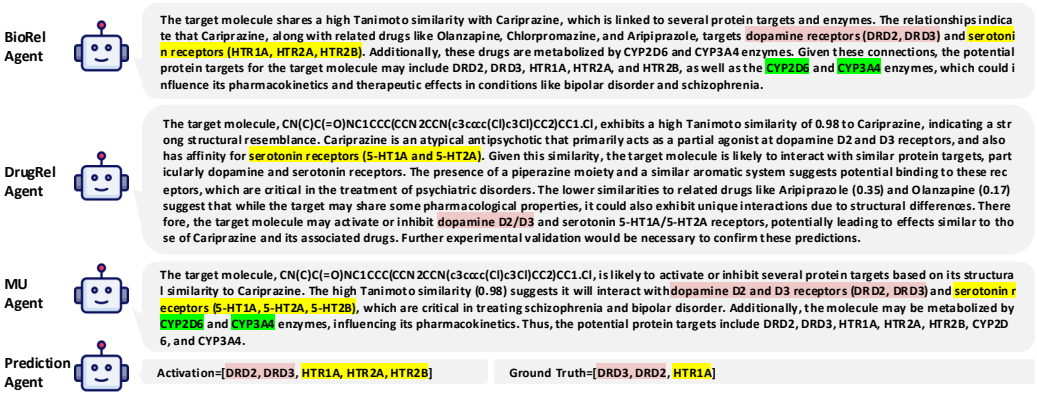

Figure 12: Additional case studies. Red represents dopamine-related proteins, yellow represents serotonin-related proteins, and green represents Cytochrome P450-related enzymes.

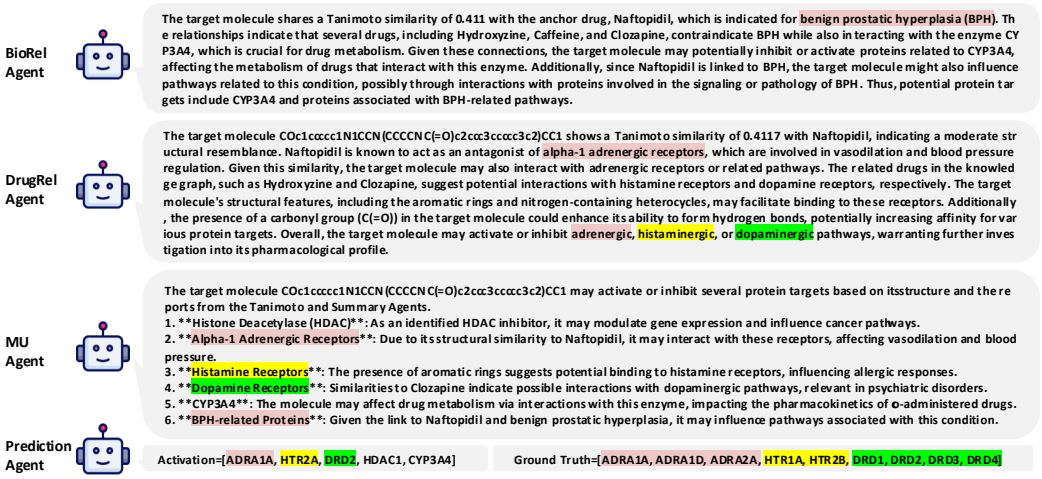

Figure 13: Full version of Figure 11.

H Agent Templates

In this section, we provide the templates for each agent used in Section 2. We follow the previous work for designing the system prompt [38].

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.