RAG-Enhanced Collaborative LLM Agents for Drug Discovery

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

[1] Korea Advanced Institute of Science and Technology
[2] Genentech
{namkyeong96, cy.park}@kaist.ac.kr, {debroue1, hajiramm, biancalt, scaliag}@gene.com

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 major challenges. First, it hinders the application of more flexible general-purpose LLMs for 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.

Code — https://github.com/Genentech/CLADD

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. 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 (Zhang et al. 2024; Pei et al. 2024). 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.

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 (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) complicates maintaining domain-specific models through expensive fine-tuning. More importantly, drug discovery applications often require promptly incorporating new insights as they become available, for example, through new experiments or the scientific literature. In addition to being impractical, regular rounds of fine-tuning also introduce challenges such as catastrophic forgetting, while not necessarily providing grounded answers. Above all, real-world drug discovery questions are inherently complex, open-ended, and context-dependent, spanning heterogeneous data types. 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 (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 biochemical space that renders exact retrieval ineffective. Second, biochemical data is extremely heterogeneous, spanning diverse modalities such as molecules, proteins, diseases, and complex relationships between them (Wang et al. 2023), which can also exist across multiple sources. 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) while re-

[*] Work done while the author was an intern at Genentech.
[†] Corresponing Authors

maining 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 heterogeneous biochemical entities and their relationships. CLADD is powered by general-purpose LLMs, while also integrating domain-specific LLMs, 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 (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 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 biochemical data without fine-tuning, while addressing 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.

2 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 (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. 2023) and Mol-Instructions (Fang et al. 2023) adopt instruction tuning 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 and Buehler 2024). 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.

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, Song, and Fu 2024) introduces a multi-agent framework integrating multiple external data sources, but is limited to predicting drug-target interaction. Another study with the same name employs an agentic framework for automating machine learning programming for drug discovery tasks (Liu et al. 2024). In contrast, our work seeks to tackle a diverse array of drug discovery tasks, grounding the agent capabilities in external knowledge.

3 Methodology

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 generate an answer such as $\mathcal{A}_{g_q}$ = *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, among others.

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 rela-

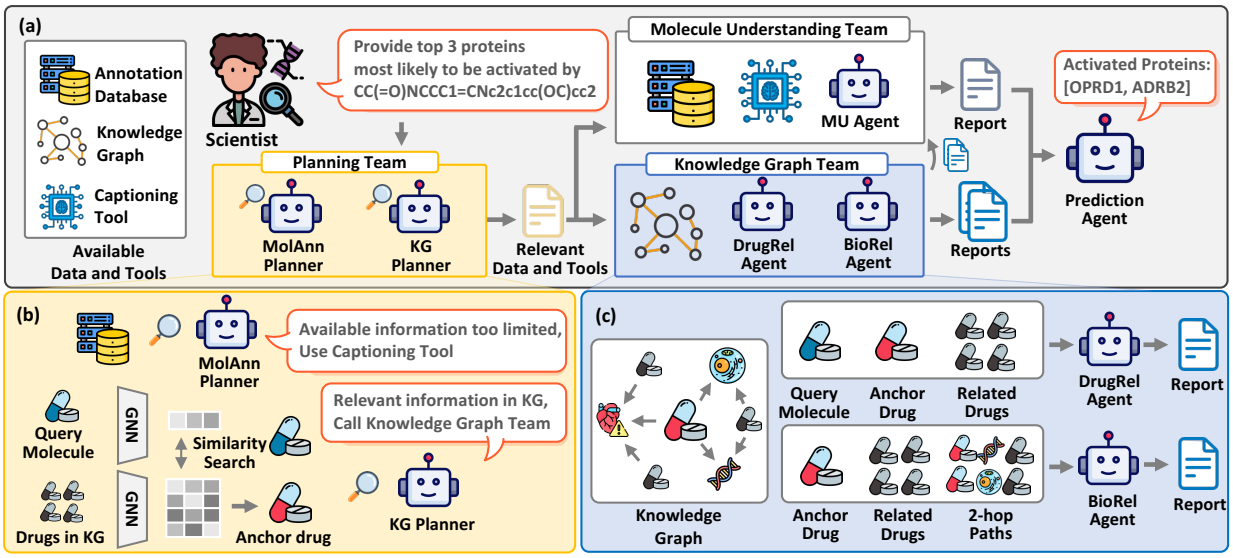

Figure 1: Overview of CLADD.

tions $\mathcal{R}$ connecting them. In this study, we only assume that molecule (or drug) entities are present in the KG, while any other types of entities can exist.

In addition to external databases, we assume access to pretrained *molecular captioning models* that can be used as external tools. 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.

3.1 CLADD

Here, we introduce CLADD, a multi-agent framework for general molecular question-answering that supports multiple drug discovery tasks and external databases/tools. 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 3.1); the **Knowledge Graph (KG) Team**, which retrieves relevant contextual information about the molecule from available KG databases (Section 3.1); and the **Molecular Understanding (MU) Team**, which retrieves and integrates information from molecular annotation databases and external tools for molecule description (Section 3.1). 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. CLADD provides a general framework, with a concrete instantiation described in Section 4. The overall framework is depicted in Figure 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 (Yu et al. 2024), they are often sparse, with many molecules entirely missing or lacking sufficient details due to the vastness of the chemical space (Lee et al. 2024).

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:

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

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 (Baek, Aji, and Saffari 2023; Jiang et al. 2023), 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 principle that structurally similar molecules often exhibit related biochemical activity (Martin, Kofron, and Traphagen 2002). 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 KG molecules ($g_{\mathcal{G}}$), $g_a = \underset{g \in g_{\mathcal{G}}}{\text{argmax}} \frac{emb(g_q) \cdot emb(g)}{\|emb(g_q)\| \|emb(g)\|}$, where *emb* is a representation produced by a graph neural network (GNN) pretrained with 3D geometry (Liu et al. 2021), which outputs structure-aware molecular embeddings.

Then, the KG Planner agent decides whether to use the KG based on query-anchor similarity, for the specific task. To do so, we also provide the KG Planner with the Tanimoto similarity, a domain-specific metric the LLM can use to reason about chemical similarity, as follows:

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

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.

Knowledge Graph Team This team provides 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, 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 (Wen, Wang, and Sun 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 g_q. 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 related 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:

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

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 the related drugs while effectively assessing the relatedness of such information, also based on the Tanimoto similarity.

Biological Relation (BioRel) Agent. The BioRel Agent summarizes how the anchor drug and the retrieved drugs are biologically related, integrating additional biochemical entities present in the KG, such as targets, indications, and side effects. Specifically, given an anchor drug g_a, a set of related drugs $g_{r^1}, \ldots, g_{r^k}$, the collection of all 2-hop paths $\mathcal{P}$ linking the anchor drug to the related drugs, and the instruction $\mathcal{I}$, the agent generates the report as follows:

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

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 reason about the query molecule in relation to other relevant entities in the KG for the specific task, they leverage distinct knowledge sources and play 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. 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 4, we demonstrate how they complement each other, with a synergistic effect when combined.

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 other reports.

Molecule Annotations. Annotations from the external databases 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 descriptions $\tilde{c}_q$ are generated with the external captioning tools as follows:

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

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 (Pei et al. 2023; Yu et al. 2024), and can potentially include any tools with molecules as input, given that the output can be transformed into text.

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

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

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:

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

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

	(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	<u>0.87</u>
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	<u>1.65</u>	<u>1.09</u>	<u>1.37</u>	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 1: Performance in drug-target prediction tasks (Precision @ 5). **Bold** and <u>underline</u> indicate 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	<u>70.95</u> (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 LLMs				
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)	<u>64.65</u> (2.01)	<u>92.80</u> (2.92)	<u>77.23</u> (1.95)
CLADD	**72.28 (1.04)**	**66.42 (1.31)**	**93.80 (2.30)**	**77.74 (3.15)**

Table 2: Performance in molecular captioning tasks, mean AUROC with standard deviation (in parentheses). **Bold** and <u>underline</u> indicate the best and second-best language model-based methods.

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 (Kojima et al. 2022) and does not require additional fine-tuning. Therefore, a key advantage of CLADD is its flexibility, which enhances scientist-AI interactions.

4 Experiments

Implementation Details. In all experiments, we utilize GPT-4o mini through the OpenAI API for each agent. In our experiments, we use PrimeKG (Chandak, Huang, and Zitnik 2023) as the KG, PubChem (Kim et al. 2021) as an annotation database, and MolT5 (Edwards et al. 2022) as an external captioning tool.

4.1 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 (Santos et al. 2017; Batool, Ahmad, and Choi 2019). 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 (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.

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).

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 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.

4.2 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 (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 task 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**.

Methods Compared. We consider different baseline ap-

	(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**

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. **Bold** and underline indicate best and second-best methods. * indicates whether the model always outputs the same response, either "Yes" or "No".

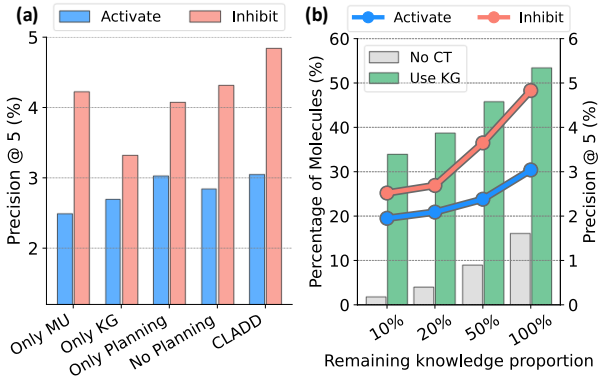

Figure 2: Ablation studies. (a) On model components. (b) On external knowledge.

proaches. 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, for reference, we consider standard molecular property prediction baselines, including two GNNs pre-trained with different methodologies: GraphMVP (Liu et al. 2021) and MoleculeSTM (Liu et al. 2023).

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), 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 (Beltagy, Lo, and Cohan 2019) 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 (Liu et al. 2023) to simulate realistic distribution shifts.

Experimental Results. Table 2 summarizes the results. **1)** While domain-specific LLMs 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 (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 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 geometric information, into LLM understanding.

4.3 Biological Activity Prediction

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* (Basile, Yahi, and Tatonetti 2019) and *antibacterial activity* (Melo, Maasch, and de la Fuente-Nunez 2021) prediction.

Datasets. For drug toxicity prediction, we use three benchmark datasets: **hERG** (Wang et al. 2016), **DILI** (Xu et al. 2015), and **Skin** (Alves et al. 2015). For antibacterial activity prediction, we use a dataset based on Eke, Williams, and Abramovitch (2025), 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.

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. 2023), alongside two general-purpose LLMs, GPT-4o and GPT-4o mini.

Evaluation Protocol. Evaluation follows a zero-shot QA setting. The input includes a SMILES representation of the molecule and the task description. Using the text-formatted output generated by each model, we compute the Macro-F1 score 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 includes GPT-4o mini, which is used as building block of CLADD. This highlights its ability to perform zero-shot predictions without domain-specific

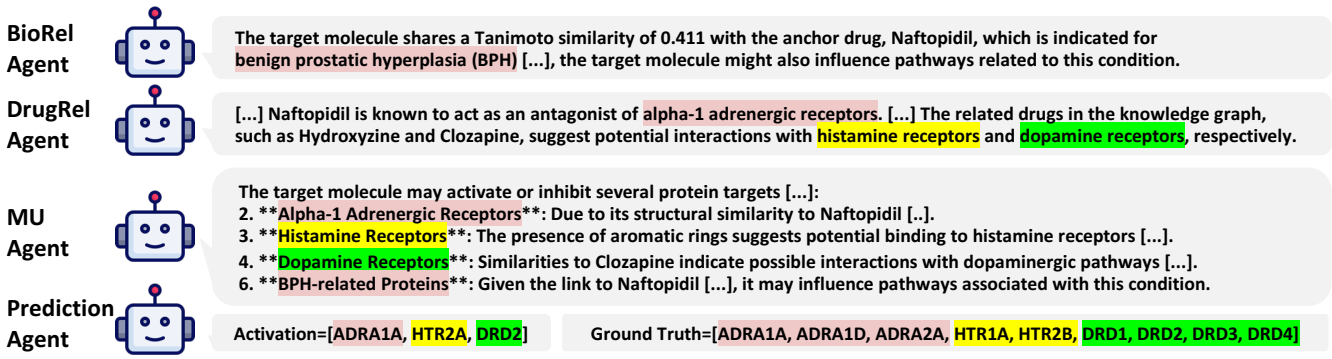

Figure 3: Example of collaboration between agents in CLADD (on the drug-target prediction task). Red represents adrenergic receptors, yellow represents histamine receptors, and green represents dopamine receptors.

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)** Moreover, most baselines performed especially poorly on the recently released dataset (MLSMR_Mtb). However, CLADD shows a significant improvement over the baselines, demonstrating its reasoning ability on unseen tasks.

4.4 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 multi-agent architecture improves information processing*, 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.

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*.

4.5 Case Studies

Figure 3 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 (Klotsman et al. 2004)), and **2)** analyzing related drugs in the knowledge graph (*e.g.,* Hydroxyzine, Clopamine), to infer interactions 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 reliable predictions.

5 Conclusion

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. Moreover, as shown in our case studies, CLADD's chain of messages provides insight into its decision-making process, fostering more interpretable scientist-AI interactions.

Acknowledgements

N.L. and C.P. were supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2025-02304967, AI Star Fellowship (KAIST)). Additionally, N.L. and C.P. received funding from the National Research Foundation of Korea (NRF) through two separate grants: RS-2024-00335098 (funded by the Korea government (MSIT)) and RS-2022-NR068758 (funded by the Ministry of Science and ICT). E.D.B., E.H., T.B., G.S. are employees of Genentech and shareholders of Roche.

References

Alves, V. M.; Muratov, E.; Fourches, D.; Strickland, J.; Kleinstreuer, N.; Andrade, C. H.; and Tropsha, A. 2015. Predicting chemically-induced skin reactions. *Toxicology and applied pharmacology*, 284(2): 262–272.

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

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

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

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

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

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

Chan, C.-M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; and Liu, Z. 2024. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. In *ICLR*.

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

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. 2024. Tx-LLM: A Large Language Model for Therapeutics. *arXiv preprint arXiv:2406.06316*.

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

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. 2017. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nature medicine*, 23(4): 405–408.

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. 2014. 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.

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

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

Edwards, C.; Lu, Z.; Hajiramezanali, E.; Biancalani, T.; Ji, H.; and Scalia, G. 2024. MolCap-Arena: A Comprehensive Captioning Benchmark on Language-Enhanced Molecular Property Prediction. *arXiv preprint arXiv:2411.00737*.

Eke, I. E.; Williams, J. T.; and Abramovitch, R. B. 2025. Genetic and Cheminformatic Characterization of Mycobacterium tuberculosis Inhibitors Discovered in the Molecular Libraries Small Molecule Repository. *ACS Infectious Diseases*, 11(4): 882–893.

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

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

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

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

Guo, H.; Zhao, S.; Wang, H.; Du, Y.; and Qin, B. 2024. Moltailor: Tailoring chemical molecular representation to specific tasks via text prompts. In *AAAI*, volume 38.

Inoue, Y.; Song, T.; and Fu, T. 2024. DrugAgent: Explainable Drug Repurposing Agent with Large Language Model-based Reasoning. *arXiv preprint arXiv:2408.13378*.

Jiang, J.; Zhou, K.; Dong, Z.; Ye, K.; Zhao, X.; and Wen, J.-R. 2023. StructGPT: A General Framework for Large Language Model to Reason over Structured Data. In Bouamor, H.; Pino, J.; and Bali, K., eds., *EMNLP*, 9237–9251. Association for Computational Linguistics.

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

Klotsman, M.; Weinberg, C.; Davis, K.; Binnie, C.; and Hartmann, K. 2004. A case-based evaluation of SRD5A1, SRD5A2, AR, and ADRA1A as candidate genes for severity of BPH. *The Pharmacogenomics Journal*, 4(4): 251–259.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *NeurIPS*, 35: 22199–22213.

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

Lee, N.; Laghuvarapu, S.; Park, C.; and Sun, J. 2024. Vision Language Model is NOT All You Need: Augmentation Strategies for Molecule Language Models. In *CIKM*.

Liu, S.; Lu, Y.; Chen, S.; Hu, X.; Zhao, J.; Fu, T.; and Zhao, Y. 2024. DrugAgent: Automating AI-aided Drug Discovery Programming through LLM Multi-Agent Collaboration. *arXiv preprint arXiv:2411.15692*.

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

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

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

McNaughton, A. D.; Sankar Ramalaxmi, G. K.; Kruel, A.; Knutson, C. R.; Varikoti, R. A.; and Kumar, N. 2024. CACTUS: Chemistry Agent Connecting Tool Usage to Science. *ACS omega*, 9(46): 46563–46573.

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

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

ODonoghue, O.; Shtedritski, A.; Ginger, J.; Abboud, R.; Ghareeb, A. E.; and Rodriques, S. G. 2023. BioPlanner: Automatic Evaluation of LLMs on Protocol Planning in Biology. In *EMNLP*.

Pei, Q.; Wu, L.; Gao, K.; Zhu, J.; Wang, Y.; Wang, Z.; Qin, T.; and Yan, R. 2024. Leveraging Biomolecule and Natural Language through Multi-Modal Learning: A Survey. *arXiv preprint arXiv:2403.01528*.

Pei, Q.; Zhang, W.; Zhu, J.; Wu, K.; Gao, K.; Wu, L.; Xia, Y.; and Yan, R. 2023. BioT5: Enriching Cross-modal Integration in Biology with Chemical Knowledge and Natural Language Associations. *arXiv preprint arXiv:2310.07276*.

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

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

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. 2017. A comprehensive map of molecular drug targets. *Nature reviews Drug discovery*, 16(1).

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

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

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

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

Wang, Q.; Downey, D.; Ji, H.; and Hope, T. 2024b. SciMON: Scientific Inspiration Machines Optimized for Novelty. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *ACL*, 279–299. Bangkok, Thailand: Association for Computational Linguistics.

Wang, S.; Sun, H.; Liu, H.; Li, D.; Li, Y.; and Hou, T. 2016. Predicting hERG blockers by combining multiple pharmacophores and machine learning approaches. *Molecular pharmaceutics*, 13(8).

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

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. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46(D1).

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

Xu, J.; Wu, Z.; Lin, M.; Zhang, X.; and Wang, S. 2024. LLM and GNN are Complementary: Distilling LLM for Multimodal Graph Learning. *arXiv preprint arXiv:2406.01032*.

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

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

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

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

Zhao, H.; Liu, S.; Chang, M.; Xu, H.; Fu, J.; Deng, Z.; Kong, L.; and Liu, Q. 2023. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *NeurIPS*.

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