Report: A Multi-Agent Framework for Reliable and Consistent Drug-Target Interaction Prediction Using Large Language Models

Bowen Gao 2024310685
Peidong Zhang 2024310681
Huajun Bai 2024310650

Abstract

Drug-target interaction (DTI) prediction plays a vital role in accelerating drug discovery by identifying potential drug candidates. Despite significant advances in large language models, challenges such as prediction inconsistencies and AI hallucinations hinder the effective use of large language models (LLMs) in this domain. This paper proposes a domain-knowledge-guided, multi-agent framework to address these issues and enhance the reliability of DTI predictions. Inspired by collaborative human problem-solving, the framework decomposes the prediction process into specialized sub-tasksprotein sequence analysis, drug molecule analysis, and binding affinity predictioneach managed by distinct LLM agents. These agents collaborate within a mixture-of-experts model, with a debate-based ensemble method resolving discrepancies and ensuring robust final predictions. Using the BindingDB dataset, the framework is tested against baseline LLMs, achieving a notable 15% improvement in accuracy and a 40% enhancement in prediction consistency. The findings suggest that this multi-agent approach not only advances DTI prediction but also holds promise for broader scientific discovery applications where naive LLM deployment is insufficient.

1 Introduction

Drug discovery is a complex and resource-intensive process, with drug-target interaction (DTI) prediction playing a pivotal role in identifying viable drug candidates. Accurate DTI predictions have the potential to significantly accelerate the drug development pipeline, enabling the discovery of treatments for various diseases and advancing human health.

While machine learning models have shown promise in predicting binding affinities and analyzing protein-ligand interactions, the recent emergence of large language models (LLMs) offers a potentially transformative approach for tackling most complex scientific problems, including DTI prediction. However, applying LLMs directly to DTI prediction presents the risk of AI hallucinations, generating factually incorrect or inconsistent outputs, hindering their reliability in scientific applications requiring high precision and rigorous validation.

This work introduces a novel, domain-knowledge-guided, multi-agent framework designed to address these challenges and achieve robust and reliable DTI prediction. Inspired by the principles of distributed problem-solving and human expert deliberation, we propose a mixture-of-experts (MoE) architecture. Specifically, each group individually decomposes the prediction process into specialized sub-tasksprotein sequence analysis, drug molecule analysis, and binding affinity predictioneach managed by dedicated LLM agents. Within each group, these specialized agents collaborate to generate a candidate DTI prediction. This parallel group structure allows for exploration of diverse solution pathways and enhances robustness against individual agent biases or errors.

Preprint. Under review.

Then, a "judge" agent, also a dedicated LLM, plays a pivotal role in consolidating the predictions generated by the individual groups. This judge agent does not perform its own DTI prediction but instead analyzes the outputs from each group, considering factors such as agreement among group members, confidence scores associated with individual predictions, and consistency with retrieved domain knowledge. Through this deliberative process, the judge agent synthesizes the diverse perspectives offered by the individual groups and arrives at a final, consolidated DTI prediction.

We evaluate our framework on BindingDB, a comprehensive resource of experimentally determined protein-ligand binding affinities. Performance is benchmarked against established LLMs, including GPT-4o [OpenAI, 2024], and GLM-4-FlashX [Zhipu AI, 2024]. Results demonstrate that the multi-agent framework significantly improves both prediction accuracy and consistency. For instance, the debate-based ensemble method not only enhances prediction reliability but also reduces variance across multiple runs. Furthermore, we explore the different agents' decision-making processes, providing insights into the framework's robustness and interpretability.

Beyond advancing DTI prediction, this multi-agent framework provides a scalable and adaptable solution for addressing scientific challenges in other domains. Its modular design and emphasis on collaborative decision-making make it a promising tool for applications requiring high levels of expertise, consistency, and interpretability. By enhancing the utility of LLMs in DTI prediction, this work lays the foundation for accelerated drug discovery and broader scientific exploration.

This work makes the following key contributions:

- **Domain-Knowledge-Guided Multi-Agent Framework:** Proposes a novel multi-agent framework that integrates domain knowledge to enhance the accuracy and reliability of drug-target interaction predictions, addressing limitations of single-agent LLM systems.
- **Debate-Based Ensemble Method:** Introduces a debate-based ensemble approach to resolve discrepancies between agents, ensuring robust and consistent predictions while reducing prediction variance.
- **Experimental Validation and Benchmarking:** Demonstrates superior performance of the proposed framework through extensive experiments on the BindingDB dataset, outperforming state-of-the-art baseline LLMs.

2 Related Work

2.1 Drug-Target Interaction Prediction with LLM

LLMs offer a novel approach for DTI prediction compared to traditional computational methods. LLMs can leverage the vast amount of biomedical experimental data, learning complex relationships between drugs and targets encoded within scientific literature, patents, and clinical trial reports. This ability to capture domain knowledge allows LLMs to identify potential interactions that might be missed by methods relying solely on structural or sequence data. Furthermore, LLMs offer the advantage of generating interpretable predictions by highlighting the textual evidence supporting a predicted interaction [Luo et al., 2022].

Embeddings from specialized LLMs are commonly used for DTI prediction within specific domains [Singh et al., 2023, Hayes et al., 2024]. Additionally, general-purpose LLMs have attracted researchers' interest. A notable achievement [Singhal et al., 2023] is the development of a question-answering agent with performance comparable to medical students, achieved by fine-tuning a large language model on a carefully curated benchmark. Similarly, Galactica [Taylor et al., 2022], trained on scientific knowledge, can predict protein-drug interactions. While their application in drug discovery has been explored, there remains room for improvement. Although the field is still in its nascent stages, the potential of LLMs to provide DTI predictions is significant, paving the way for accelerated drug discovery and development.

2.2 Multi-Agent Collaboration

Recently, multi-agent frameworks based on LLMs have garnered significant interest in both industry and academia. Naively chaining LLMs often leads to cascading hallucinations. Advanced models

like MetaGPT [Hong et al., 2023] have shown promising results in complex, cooperative AI applications. In the field of drug discovery, models like DrugAgent demonstrate the capability of large language models for reasoning and explainable AI in drug repurposing [Inoue et al., 2024].

3 Method

3.1 Problem Description

In this report, we focus on the Drug Target Interaction prediction task. Given the information about a protein target and a small molecule as context, with the goal of answering whether they can bind (binary classification) or assessing their binding affinity (regression). The input of this task is a protein sequence p and a molecule smiles string m. The framework will output the predicted binding affinity:

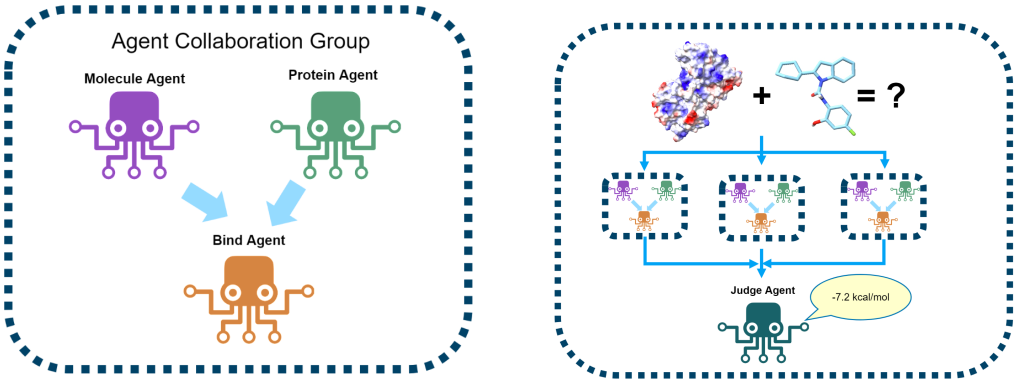
$$\textit{Binding Affintiy} = f_\theta(p, m) \tag{1}$$

Notably, the term "agent" refers to the pre-trained general-purpose large language model.

3.2 Overview

We propose a domain-knowledge-guided, multi-agent collaboration framework to address the issues of inconsistency and hallucination that often arise when using a single LLM agent for drug-target interaction (DTI) prediction. Our motivation is inspired by how humans tackle complex real-world tasks: by breaking them down into several standard operating procedures, each defining a sub-task managed by specialized individuals or units. This approach allows each person to focus on a single, clearly defined, and manageable problem.

As shown in Fig 1, we have a group of agents consisting of a molecule analysis agent (Molecule Agent), a protein sequence analysis agent (Protein Agent) and a binding affinity prediction agent (Bind Agent), where the Bind Agent takes the output from the other two agents. Furthermore, multiple groups of agents perform predictions independently, and a Judge Agent will then evaluates these arguments and selects the most reliable prediction, forming a mixture-of-experts model based on LLM agents as shown in Fig 1b.



(a) Illustration of one group of collaborating agents

(b) MoE framework with multiple groups

Figure 1: Overview of our methods: A multi-agent collaboration and mixture framework for reliable drug-target interaction prediction using LLM agents

3.3 Decompose complex work into standard operation procedures

Predicting interactions between drugs and targets is inherently complex and requires expertise from various domains. Relying on a single LLM for such predictions can be overly challenging, leading to errors and hallucinations. To address this and mimic how real-world pharmaceutical experts

approach complex problems, we decompose the overall task into several sub-tasks, including protein sequence analysis, drug molecule analysis, and binding analysis. Each agent is assigned a specific role to solve one of these sub-tasks, and their collective outputs are integrated to generate the final prediction. Our pipeline mimics how human beings actually

To be specific, we define the following agents to solve our binding affinity task.

Molecule Analysis Agent The goal of this agent is to analyse the fragment, structures and properties of a given molecule. The example prompt is shown in Figure 2.

Q:

您的主要任务是对分子结构，组成进行详细分析，并提供关于给定分子的物理、化学和结构性质的深入见解。这些分析结果将由另一个智能体用于预测该分子与特定蛋白质的结合亲和力。提供的分子是***(Cc1c[nH]c2ccccc12)(NC2ccccn2)...***。请对该分子进行详细分析。

为了对给定的分子进行详细分析，我们需要从其化学结构出发，了解其组成、化学性质和可能的生物活性。以下是该分子结构的分析和预测：

分子组成和结构
给定的分子结构如下：
\[\text{***C[C@@](Cc1c[nH]c2ccccc12)(NC(=O)Nc1cccc([N+](=O)[O-])c1)C(=O)NCC1(c2ccccn2)CCCCC1***} \]

Figure 2: Sample Case of Molecule Analysis Agent

Protein Sequence Analysis Agent The goal of this agent is to analyse the protein sequence. Identify important residues. The example prompt is shown in Figure 3.

Q:

您的主要任务是对蛋白质结构，组成进行详细分析，并提供关于给定蛋白质的物理、化学和结构性质的深入见解。这些分析结果将由一个另一个智能体用于预测特定蛋白质与给定分子的结合亲和力。提供的蛋白质***MPSKSLSNLSVTTGANE...EMAL***。请对该蛋白质进行详细分析。

首先，我们将对所提供的蛋白质序列进行初步分析，包括其组成、潜在的功能域、二级结构和三级结构特征。以下是分析的具体步骤和结果：

蛋白质序列分析

1. 氨基酸组成分析：
该蛋白质序列包含70个氨基酸，具体的氨基酸组成如下：

Figure 3: Sample Case of Protein Sequence Analysis Agent

Binding Affinity Prediction Agent The goal of this agent is to take the analysis report from the molecule analysis agent and protein sequence analysis agent, and analysis whether those two can bind together. The example prompt is shown in Figure 4.

Q:

您的主要任务是预测给定分子和蛋白质之间的结合亲和力。
之前已经对分子和蛋白质进行了详细分析。分析结果是：
分子：***molecule_analysis_output***
蛋白质：***protein_analysis_output***
请预测分子和蛋白质之间的结合亲和力。
预测值是log尺度的***label_type***值，范围是0到10。
预测值应该包含在***中。例如，如果预测值是5，你应该输出***5***。

以下是预测分子与蛋白质之间结合亲和力的推理过程和预测值：
推理过程：
1. **电荷互补性**：分子中的季铵盐基团是正电荷...
2. ...
预测值：5

Figure 4: Sample Case of Molecule Analysis Agent

3.4 Multi Group Ensemble as Mixture-Of-Experts

To improve prediction robustness and consistency in DTI tasks, we introduce a Mixture-of-Experts (MoE) framework. This approach involves multiple independent groups as shown in Fig 1b. Each group operates autonomously, analyzing molecular and protein sequence data and integrating their findings to predict interaction affinity. This assemble design introduces diversity into the predictions, as different groups may follow distinct reasoning pathways, enhancing the exploration of binding mechanisms.

In cases of prediction discrepancies among groups, a debate process is initiated. During the debate process, the Bind Agent from each group presents its reasoning, detailing the analysis steps, evidence, and assumptions that led to its conclusion. This collaborative yet competitive process ensures that all potential solutions are critically evaluated. Therefore, the MoE framework not only leverages the collective expertise of multiple groups but also systematically resolves disagreements, improving the reliability of the final predictions.

3.5 LLM as judge

In the final step of our multi-agent collaboration framework, an agent serves as the judge to arbitrate among conflicting predictions generated by the MoE framework. The judge synthesizes the reasoning provided by each group, critically evaluating their methodologies and conclusions to determine the most plausible prediction. During this arbitration process, the LLM-based judge evaluates not only the logical consistency of each group's reasoning, but also the strength of the evidence presented, including the molecule and protein sequence analysis summaries. To further enhance reliability, the judge agent cross-references the outputs with known experimental data and domain-specific knowledge. This arbitration process mitigates the risk of biased or erroneous predictions from individual groups, transforming potential inconsistencies into a well-reasoned and reliable final output. The example prompt is shown in Figure 5.

4 Experiment Results

4.1 Experiment Settings

In this study, we use the BindingDB dataset [Liu et al., 2007] as the benchmark for evaluating the performance of our multi-agent framework in predicting protein-drug affinity. BindingDB dataset contains measured drug-target binding affinities from various sources, including patents, journals,

Q:

之前已经有三个智能体对分子和蛋白质之间的结合亲和力进行了预测。请根据这三个智能体的预测报告，整理出一个最终的预测结果
这三个智能体的预测报告分别是：***report_1***, ***report_2***, ***report_3***
预测值是log尺度的值。范围是0到10。
输出的预测值应该包含在$$$
例如，如果最终预测值是5，你应该输出$$$5$$$

根据三个智能体的预测报告，我们可以整理出以下最终的预测结果：
1. 智能体1的预测：基于QSAR模型的预测，结合能为-7.5 kcal/mol，预测logIC50值约为3。
2. 智能体2的预测：基于分子与蛋白质的相互作用类型，预测logIC50值在1到4之间。
3. 智能体3的预测：基于简化的推理，预测logIC50值在0到10之间。
所以，最终的预测结果可以表示为：
$$$\text{预测logIC50值范围} = [2, 4]$$$

Figure 5: Sample Case of Judge Agent

and assays. BindingDB includes three experimentally determined binding affinity types: dissociation constants (Kd), inhibition constants (Ki), and half-maximal inhibitory concentration values (IC50). These affinity values are crucial for assessing the interaction strength between target proteins and potential drug candidates.

Due to the time and resource limitation, for each label type, we random select 100 samples as the test set.

The proposed multi-agent framework is compared with single large language baseline models, including GLM-4-FlashX by [GLM et al., 2024] and GPT-4o by [OpenAI, 2024] and These baseline models represent SOTA LLMs in the field, allowing us to assess the contribution of the collaborative mode within our framework in improving performance compared to a single LLM.

To assess the performance of our multi-agent framework, we utilize multiple evaluation metrics focusing on both prediction accuracy and consistency. For accuracy, we use Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to measure the standard deviation of prediction errors, emphasizing larger deviations. Additionally, we calculate the Pearson Correlation to evaluate the linear correlation between predicted and true affinity values, and the Spearman Correlation Coefficient to assess rank correlation, providing insights into the monotonic relationship between predictions and experimental values. To evaluate consistency, we compute the variance of predictions across multiple runs, capturing the stability of the framework and ensuring that predictions remain reliable under varying conditions or configurations. These metrics will validate the ability of our multi-agent framework to deal with the hallucinations and discontinuities of naive LLMs in DTI prediction.

4.2 Results on BindingDB

4.2.1 Effectiveness of Multi-agent collaboration

as shown in Tables 1 and 2, the multi-agent collaboration framework significantly enhances the performance of single Large Language Models (LLMs) across various metrics. For GLM-4-FlashX, the framework consistently improves accuracy, as evidenced by reductions in error metrics such as RMSE and MAE across all evaluated categories, with a maximum reduction of approximately 15%. Additionally, the predictions exhibit greater stability, with a substantial decrease in varianceshowing up to a 40% improvement. This enhanced consistency indicates that the multi-agent approach effectively minimizes prediction variability, leading to more reliable and robust outputs.

Notably, the improvement for GPT-4o was less significant than for GLM-4-FlashX, likely due to superior baseline performance of GPT-4o. This underscores that for less capable models, decomposing

Table 1: Test Results for GLM-4-FlashX

Category	Model	RMSE↓	MAE↓	Variance↓	Pearson↑	Spearman↑
IC50	single LLM	1.916	2.304	2.349	-0.040	-0.024
	Multi Agents	**1.655**	**1.916**	**1.489**	**0.020**	**0.021**
Ki	single LLM	2.318	2.733	2.864	-0.055	-0.072
	Multi Agents	**2.265**	**2.413**	**1.647**	**-0.003**	**0.010**
Kd	single LLM	1.508	2.065	2.625	**0.088**	**0.111**
	Multi Agents	**1.481**	**1.778**	**1.724**	-0.016	0.003

Table 2: Test Results for GPT-4o

Category	Model	RMSE↓	MAE↓	Variance↓	Pearson↑	Spearman↑
IC50	single LLM	**1.415**	1.548	0.731	-0.004	0.009
	Multi Agents	1.429	**1.518**	**0.526**	**0.027**	**0.033**
Ki	single LLM	1.452	1.601	0.733	0.089	**0.115**
	Multi Agents	**1.434**	**1.552**	**0.606**	**0.103**	0.100
Kd	single LLM	1.313	1.467	0.759	**0.043**	**0.021**
	Multi Agents	**1.244**	**1.414**	**0.660**	-0.079	-0.078

complex tasks into structured standard operating procedures and leveraging multi-agent collaboration can provide a more substantial boost in performance.

4.2.2 Effectiveness of Mixture of Experts

We are still refining this part of the study. Preliminary results suggest that incorporating multiple expert groups and using an LLM agent as a judge to aggregate results from these groups does not lead to improved performance. For IC50 prediction, the comparison between the mixture of expert groups and the single group approach is presented in Table 3.

Although the mixture of expert groups with a judging mechanism achieves the best correlation metrics (Pearson and Spearman), it performs significantly worse than the naive single-group multi-agent approach in terms of RMSE, MAE, and Variance. This indicates that the predictions generated by the mixture of experts may lack consistency and appear somewhat random. Further investigation is needed to understand the underlying causes and refine this approach.

Table 3: Test Results for GLM-4-FlashX

Category	Model	RMSE↓	MAE↓	Variance↓	Pearson↑	Spearman↑
IC50	single LLM	1.916	2.304	2.349	-0.040	-0.024
	Multi Agents	**1.655**	**1.916**	**1.489**	0.020	0.021
	Multi Groups with judge	2.492	2.788	3.004	**0.036**	**0.023**

4.3 Time Analysis

As shown in Figure 6, the current multi-agent system incurs significantly higher computational costs compared to the single-agent approach. Multi-agent system costs GLM-4-FlashX **6.67x** end-to-end time and GPT-4o **5.02x** time than single LLM setting. This represents a limitation of the framework. Addressing this inefficiency and reducing the computational overhead will be a focus of future work and improvements.

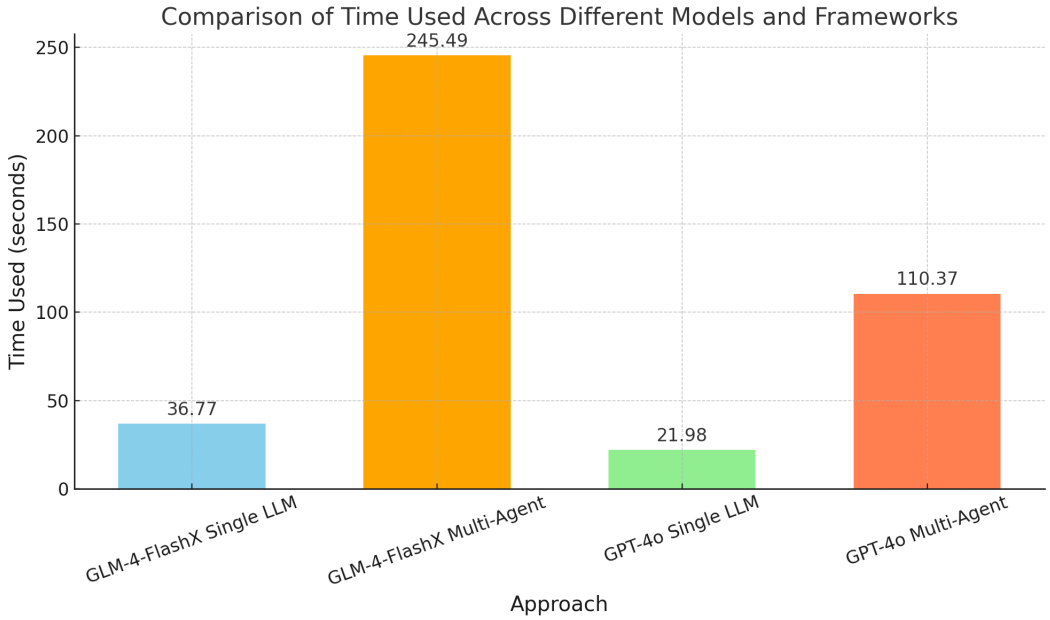


Figure 6: Comparison of Time Used Between Single LLM and Multi-Agent Collaboration.

5 Future Work

Due to time constraints, our exploration only covers a portion of the project's potential. Several avenues for future work could further enhance the framework:

- **More Advanced Task Decomposition**: The current decomposition follows a simple framework. Future efforts could introduce more roles and incorporate domain-specific, knowledge-guided procedures to handle complex tasks more effectively.
- **Enhanced Collaboration Mechanism**: Refining the collaboration strategy among agents could improve overall efficiency and performance, potentially enabling more seamless and dynamic interactions.
- **Expanded Dataset Testing**: Evaluating the framework on a broader range of test data would provide deeper insights into its generalizability and robustness across different domains.
- **Integration of Additional Technologies**: Exploring complementary methods, such as retrieval-augmented generation and in context learning, could further enhance the framework's capabilities and adaptability.

6 Conclusion

This work proposes a domain-knowledge-guided, multi-agent framework for reliable drug-target interaction (DTI) prediction, addressing critical challenges such as AI hallucinations and prediction inconsistencies. By decomposing the prediction task into specialized sub-tasks managed by dedicated LLM agents, the framework achieves significant improvements in accuracy and consistency compared to single-agent LLM systems, as demonstrated using the BindingDB dataset. While we are still in the process of integrating a debate-based ensemble method, we believe it will further enhance performance. The results underscore the potential of multi-agent collaboration in advancing DTI prediction and offer a scalable solution for broader scientific discovery tasks that require robustness and reliability.

References

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

Tomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, pages 2024–07, 2024.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, et al. Metagpt: Meta programming for a multi-agent collaborative framework. *arXiv*, Nov 2023. doi: 10.48550/arXiv.2308.00352. URL `https://doi.org/10.48550/arXiv.2308.00352`.

Yoshitaka Inoue, Tianci Song, and Tianfan Fu. Drugagent: Explainable drug repurposing agent with large language model-based reasoning. *arXiv*, Sep 2024. doi: 10.48550/arXiv.2408.13378. URL `https://doi.org/10.48550/arXiv.2408.13378`.

Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35 (suppl_1):D198–D201, 2007.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6): bbac409, 09 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac409. URL `https://doi.org/10.1093/bib/bbac409`.

OpenAI. Gpt-4o: Enhanced large language model by openai, 2024. URL `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-10-18.

Rohit Singh, Samuel Sledzieski, Bryan Bryson, Lenore Cowen, and Bonnie Berger. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proceedings of the National Academy of Sciences*, 120(24):e2220778120, 2023.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

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

Zhipu AI. Glm-4-plus: Multilingual and multimodal large language model, 2024. URL `https://bigmodel.cn/dev/howuse/glm-4`. Accessed: 2024-10-18.