CSCE 670 Project Paper: LLMs Meet Molecular Search

Yicheng Wang wangyc@tamu.edu
Yuchao Lin kruskallin@tamu.edu
Kai Yin kai_yin@tamu.edu
Zhitong Chen zhitong.chen18@tamu.edu

Abstract

This project proposes an innovative method to enhance Large Language Models (LLMs) for precise molecular searches within structured datasets, specifically using the Simplified Molecular-Input Line-Entry System (SMILES) format. By integrating LLMs with retrieval-augmented generation (RAG) techniques and leveraging molecular datasets, we aim to reduce inaccuracies and improve search outcomes. We develop four approaches and conduct intensive experiments with multiple LLMs. The result shows that GPT3.5 with RAG and additional rules provides the best results. This methodology seeks to improve the intersection of LLMs with the specialized field of molecular science, offering a more accessible and accurate molecular search tool for both professional researchers and the public, thereby potentially accelerating advancements in drug discovery and materials science. The code can be found in https://github.com/YCWangVince/LLMs_meet_molecular_search and the introduction video is here: https://youtu.be/HLZpHQFBlJ0.

1. Introduction

In the rapidly evolving field of computational chemistry and drug discovery, the ability to efficiently search and analyze molecular data has become increasingly crucial. Large Language Models (LLMs) have shown remarkable capabilities in various domains, including science and engineering by leveraging vast amounts of data to generate, understand, and predict complex patterns. However, when it comes to highly specialized fields such as molecular science, LLMs often encounter challenges, primarily due to their propensity for generating erroneous information, a phenomenon known as "hallucination" [5]. This project seeks to address this gap by integrating LLMs with structured molecular datasets to enhance the accuracy and reliability of molecular searches.

To evaluate the effectiveness of our methods, we create a test set consisting of 50 molecular SMILES and associated properties, which serve as queries during the evaluation. We implement four incremental search strategies: 1. Naive search, 2. Search with Retrieval-Augmented Generation (RAG) on the QM9 dataset, 3. Search with RAG and additional rules, 4. Search with RAG, additional rules, and Auto-Chain-of-Thought (Auto-CoT) prompts [10]. We employ a series of Large Language Models (LLMs), and our results indicate that GPT-3.5 with RAG and rules performs better than other models.

This project is designed to enhance the efficiency and accuracy of molecular searches, potentially speeding up drug discovery and advancing materials science research. It also aims to make advanced molecular search tools more accessible to researchers and the general public.

2. Related Dataset

The study of molecular datasets plays an important role in computational chemistry, materials science, and drug discovery. These datasets serve as foundational benchmarks for machine learning (ML) models, enabling researchers to predict molecular properties, understand chemical reactions, and design new molecules with desired functionalities. Here, we first introduce common molecular data representations by SMILES (Simplified Molecular-Input Line-Entry System) and then delve into an overview of prominent molecular datasets/databases, specifically focusing on QM9 [6], GEOM-DRUGS [1], ZINC15 [7], and others.

2.1. Molecular Data Representation by SMILES

SMILES is a notation that allows the representation of a chemical structure in a linear text string, which can represent molecules succinctly, including information about the arrangement of atoms, the type of chemical bonds, branching, and ring structures, among other features. Specifically, in a SMILES string, the below rules are followed.

1. Atoms are represented by their chemical symbols, e.g., C for carbon, and H for hydrogen.
2. Single bonds are not explicitly shown, while double, triple, and aromatic bonds are represented by '=', '#', and ':', respectively.
3. Branching is denoted by parentheses. For example, in ethyl alcohol (ethanol), CH3CH2OH can be written as

CC(O), indicating a branch (OH) attached to the second carbon.

4. Rings are indicated by using numbers to connect the atoms forming the ring. The same number is used at the beginning and end of the ring structure.

For example, the SMILES notation of methane (CH4) is C, the notation of acetic acid (CH3COOH) is CC(=O)O, and the notation of benzene ring (C6H6) is C1CCCCC1. Overall, SMILES provides a compact and human-readable way to represent molecules, as well as a flexible way to integrate with machine learning applications by language models.

2.2. QM9

The QM9 dataset is a widely used benchmark in the field of computational chemistry and machine learning. It consists of about 134,000 small molecules with up to 9 heavy atoms (C, O, N, F), excluding isomers. Each molecule in QM9 is represented by its SMILES and 3D coordinates. The dataset was generated from quantum chemical calculations, specifically density functional theory (DFT) at the B3LYP/6-31G(2df,p) level of theory [8]. QM9 is often used to develop and test ML algorithms capable of predicting a wide range of molecular properties, including but not limited to, atomization energies, molecular orbitals, polarizabilities, and zero-point vibrational energies.

2.3. GEOM-DRUGS

The GEOM-DRUGS dataset is part of the broader GEOM dataset series, focusing on medium-size molecules relevant to pharmaceutical research. It comprises 118,434,901 molecular conformations of 304,466 unique molecules. These molecules were generated by advanced sampling and semi-empirical density functional theory (DFT). It provides high-quality geometrical conformations obtained from high-level quantum mechanical calculations, offering a resource for benchmarking conformational analysis, molecular docking, and virtual screening algorithms.

2.4. ZINC15

The ZINC15 database is a comprehensive and publicly accessible database designed for the discovery of chemical compounds for drug development and biological research. It contains over 750 million chemical compounds, including a wide range of small organic molecules. The database is designed to be easily searchable by structure, property, or purchase availability. Users can query compounds based on molecular weight, charge, lipophilicity (logP), and other physicochemical parameters.

Besides the above QM9, GEOM-DRUGS, and ZINC15, datasets/databases like PubChem [4] and The Materials Project [3] are also well-known for molecular and material searching and discovery. In our work, we exclusively use the QM9 dataset as our foundational data source, owing to its renowned accessibility and ease of use.

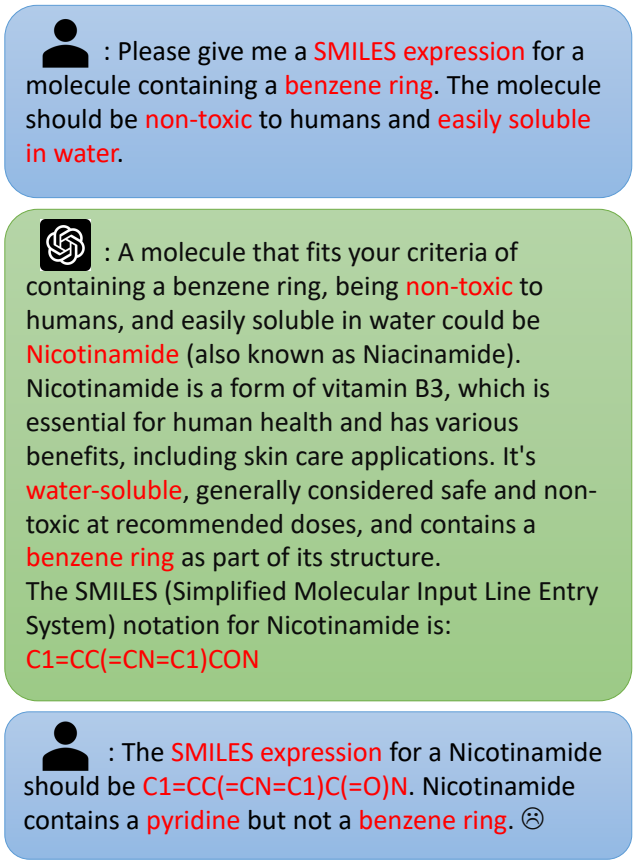

Figure 1. A hallucination example of a molecule given by GPT-4 upon request. GPT-4 returns a molecule that does not contain desired structure and a wrong SMILES expression.

3. Methodology

LLMs trained on large corpora possess general language understanding and generation capabilities. However, LLMs are prone to issues of hallucination when dealing with domain-specific tasks or processing highly specialized queries, i.e., generating incorrect information, especially for up-to-date or specialized content. In tasks involving molecular search, LLMs can easily fall into hallucination if they do not have access to a specialized dataset for guidance. A hallucination example is illustrated in Figure 1. Therefore, a method enhancing the LLMs' generation considering the real molecular dataset is required.

3.1. Prompt Design

We develop four types of prompts to explore the searching capability of LLMs.

Naive Search: We directly give the molecular property

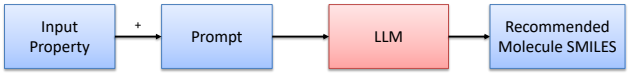

Figure 2. Naive Search

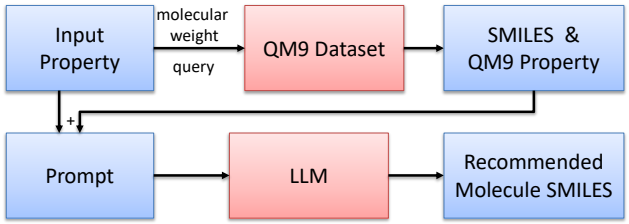

Figure 3. Search with RAG

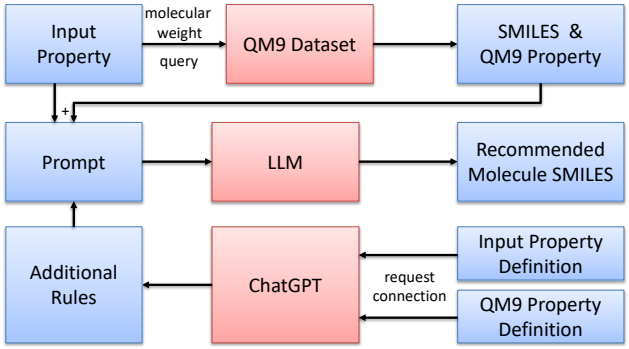

Figure 4. Search with RAG and additional rules

queries to the LLMs and ask the LLMs shown in Figure 2 to return the top 10 SMILES with the following prompts:

User: *You are a professional chemist and familiar with the properties of all kinds of organic molecules. I will provide you with properties of an organic molecule, which may include the molecular weight, density, melting point, boiling point, vapor pressure, and the functional groups it contains. Your task is recommending the ten most likely organic molecules that meet these properties, along with their SMILES expressions.*

Properties: {*PROPERTIES*}

The top 10 recommended SMILES:

This simple method can show the basic chemical and molecular knowledge level of a LLM.

Search with RAG: In this approach, we firstly retrieve related molecules the QM9 dataset by using the molecular weight property shown in Figure 3. Only those with a molecular weight within $\pm 5 g/mol$ of the query can be retrieved and used. We use the following prompt:

User: *You are a professional chemist and familiar with the properties of all kinds of organic molecules. I will provide you with properties of an organic molecule, which may include the molecular weight, density, melting point, boiling point, vapor pressure, and the functional groups it contains. Your task is recommending the ten most likely organic molecules that meet these properties, along with their SMILES expressions.*

Properties:

{*PROPERTIES*}

System: *I referred to a large molecular dataset and I found some molecules and their properties that might qualify:*

{*QM9_REFERENCE*}

Based on my chemical knowledge and the references, here are the top 10 recommended SMILES: Using RAG to provide the LLMs with retrieved and potentially relevant molecules with their properties will help improve the accuracy of molecular searches.

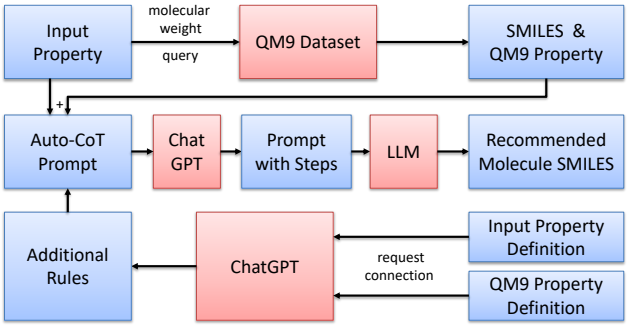

Figure 5. Search with RAG, additional rules, and Auto-CoT

Search with RAG and additional rules: In this approach, we give additional rules (generated by ChatGPT) that can connect the queried properties with the chemical properties from the retrieved molecules. The pipeline is shown in Figure 4.

User: *You are a professional chemist and familiar with the properties of all kinds of organic molecules. I will provide you with properties of an organic molecule, which may include the molecular weight, density, melting point, boiling point, vapor pressure, and the functional groups it contains. Your task is recommending the ten most likely organic molecules that meet these properties, along with their SMILES expressions.*

Properties:

{*PROPERTIES*}

System: *I referred to a large molecular dataset and I found some molecules and their properties that might qualify:*

{*QM9_REFERENCE*}

{*QM9_PROPERTY_DEFINITION*}

Based on my chemical knowledge I will use the following rules to make molecule recommendations:

{*RULES_GENERATED_BY_CHATGPT*}

Here are the top 10 recommended SMILES:

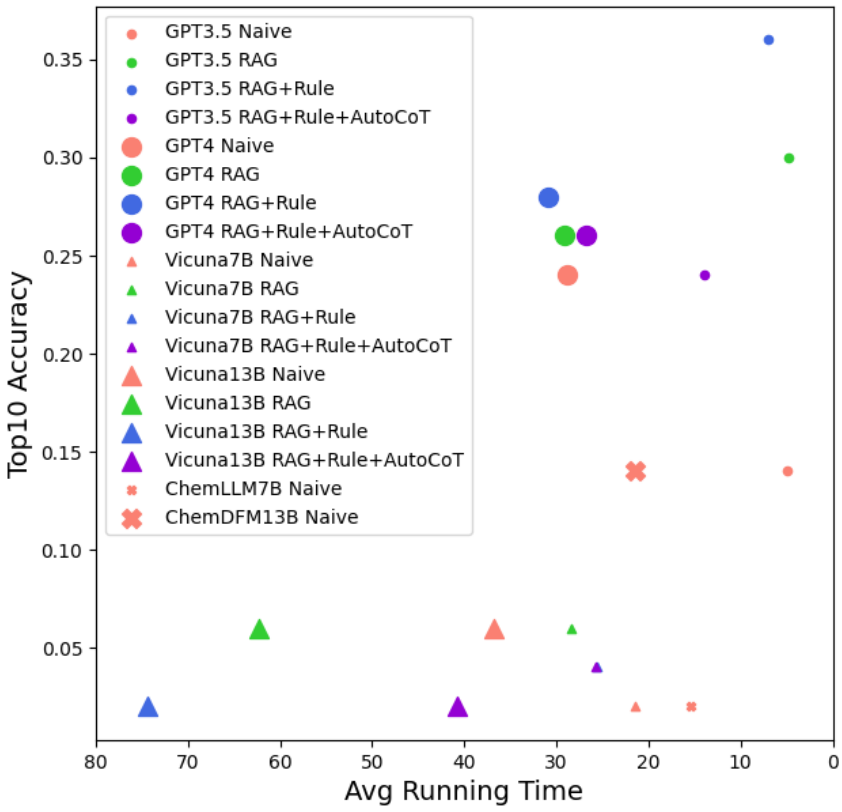

Figure 6. Experimental results. The x-axis shows the average running time (sec) per search, and the y-axis is the Top-10 Accuracy.

Search with RAG, and additional rules, and Auto-CoT: Furthermore, we apply the Auto-Chain-of-Thought [10] prompting method shown in Figure 5.

User: *You are a professional chemist and familiar with the properties of all kinds of organic molecules. I will provide you with properties of an organic molecule, which may include the molecular weight, density, melting point, boiling point, vapor pressure, and the functional groups it contains. Your task is recommending the ten most likely organic molecules that meet these properties, along with their SMILES expressions.*

Properties:

{PROPERTIES}

System: *I referred to a large molecular dataset and I found some molecules and their properties that might qualify:*

{QM9_REFERENCE}

{QM9_PROPERTY_DEFINITION}

Based on my chemical knowledge I will use the following rules to make molecule recommendations:

{RULES_GENERATED_BY_CHATGPT}

To recommend the ten most likely organic molecules that meet specific properties, I would follow a structured approach using the information you provided and applying some rules of organic chemistry. Here's how I would proceed:

{STEPS_GENERATED_BY_CHATGPT}

Here are the top 10 recommended SMILES:

3.2. LLMs

We pick three types of model to evaluate their molecular search capability.

1. GPT series, including GPT3.5-Turbo and GPT4-Turbo from OpenAI.

2. Open-sourced Llama2-based finetuned models. We use Vicuna-v1.5 7B and 13B versions [2].

3. Open-sourced models that are finetuned in chemistry context. We choose two models: ChemLLM [9] and ChemDFM [11]. However, these two models only accept 4k context so we only test the naive search on the two models.

4. Experiment

In this section, we present our experimental results. We first detail the construction of our test set, followed by the

performance evaluation of various methods applied to this set.

4.1. Constructing Test Set

To assess the accuracy of the search results, we construct a test set to evaluate the Large Language Models (LLMs). First, we randomly select a SMILES from the QM9 dataset. We then search for this SMILES on the PubChem website (https://pubchem.ncbi.nlm.nih.gov/) to retrieve its properties, such as molecular weight, density, solubility, boiling and melting points, and vapor pressure. We also incorporate functional group information from the QM9 dataset. We label 50 molecules and establish property-SMILES pairs for testing. The evaluation metric is Top-10 Accuracy, which considers a search successful if any of the 10 generated SMILES matches the labeled one.

4.2. Evaluation Result

We record all test results and average running times to demonstrate the effectiveness and efficiency of the LLMs. As illustrated in Figure 6, each data point represents a result from LLM prompting.

Remarkably, GPT-3.5 not only achieves the highest performance but also records the lowest running time. The addition of RAG and supplementary rules significantly enhances GPT-3.5's performance, increasing accuracy from 14% to 36%. This enhancement validates the utility of RAG and the additional rules.

GPT-4, which possesses considerable prior chemical knowledge, attains 24% accuracy using naive search. However, this advantage may prove counterproductive when integrating RAG and additional strategies. GPT-4 might overrely on its intrinsic knowledge, thus overlooking retrieved data and supplementary rules.

Moreover, the chemistry-specific models and open-source models do not perform as well as the GPT series. The ChemDFM model matches the performance of GPT-3.5 but cannot benefit from further strategies due to its limited context window size. The Vicuna models perform poorly, with notably low accuracy and extended running times.

Additionally, we observe that including prompts with Auto-CoT does not further enhance the models' performance. It appears that the prior integration of RAG and rules provides sufficient information, rendering the Auto-CoT steps superfluous in this context.

5. Conclusion

In this project, we have demonstrated that integrating Large Language Models (LLMs) with domain-specific datasets significantly enhances their performance in molecular search tasks. Our approach, which combines Retrieval-Augmented Generation (RAG) and rule-based enhancements, has shown promising results, particularly in improving search accuracy and reducing hallucinations typical of LLMs in specialized domains. The findings suggest that while LLMs are powerful, their utility in fields like computational chemistry can be maximized by tailoring their application with targeted data interventions. Future work could focus on exploring dynamic data retrieval mechanisms and developing domain-specific LLMs to further reduce the need for external prompts and improve the efficiency of scientific discovery processes.

References

[1] Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022. 1

[2] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. 4

[3] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013. 2

[4] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016. 2

[5] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023. 1

[6] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014. 1

[7] Teague Sterling and John J Irwin. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337, 2015. 1

[8] Julian Tirado-Rives and William L Jorgensen. Performance of b3lyp density functional methods for a large set of organic molecules. *Journal of chemical theory and computation*, 4(2):297–306, 2008. 2

[9] Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, Shufei Zhang, Mao Su, Hansen Zhong, Yuqiang Li, and Wanli Ouyang. Chemllm: A chemical large language model, 2024. 4

[10] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022. 1, 4

[11] Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen,

Xin Chen, and Kai Yu. Chemdfm: Dialogue foundation model for chemistry, 2024. 4