LLMs meet Molecular Search

Yicheng Wang, Yuchao Lin,
Kai Yin, Zhitong Chen

Motivation

Efficiently search and analyze molecular data is crucial to computational chemistry and drug discovery

LLMs have shown remarkable capabilities in various domains

When it comes to **highly specialized** fields such as molecular science, LLMs encounter challenge

LLMs for Molecular Search

Search molecules by their properties

Hallucination Problem

Method:

Retrieval-Augmented Generation (RAG)

Test set:

Build a test set from QM9 dataset

50 molecules, SMILES, properties

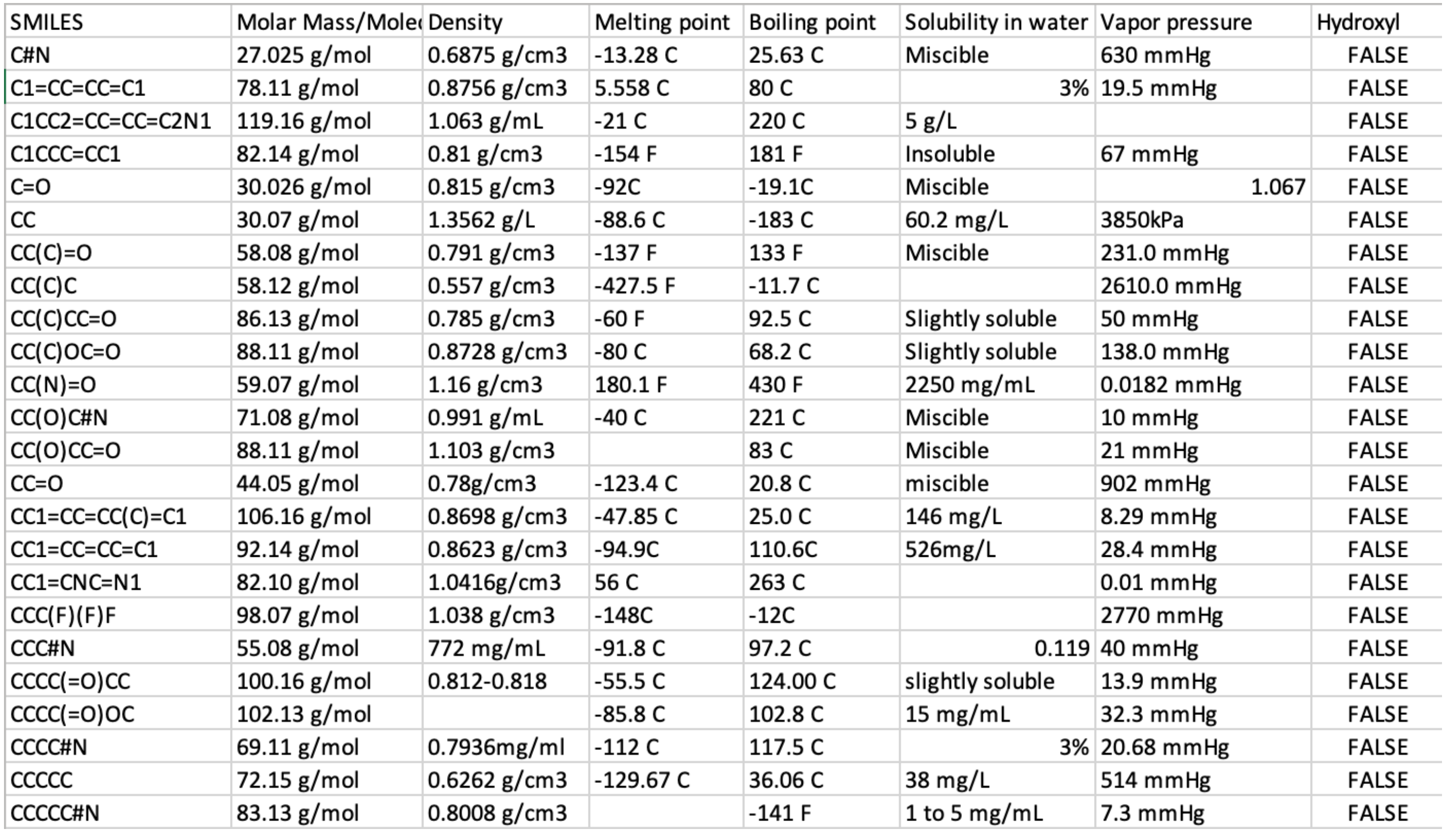
👤 : Please give me a SMILES expression for a molecule containing a benzene ring. The molecule should be non-toxic to humans and easily soluble in water.

: A molecule that fits your criteria of containing a benzene ring, being non-toxic to humans, and easily soluble in water could be Nicotinamide (also known as Niacinamide). Nicotinamide is a form of vitamin B3, which is essential for human health and has various benefits, including skin care applications. It's water-soluble, generally considered safe and non-toxic at recommended doses, and contains a benzene ring as part of its structure.
The SMILES (Simplified Molecular Input Line Entry System) notation for Nicotinamide is:
C1=CC(=CN=C1)CON

👤 : The SMILES expression for a Nicotinamide should be C1=CC(=CN=C1)C(=O)N. Nicotinamide contains a pyridine but not a benzene ring. ☹

LLMs for Molecular Search

Test Set Examples



SMILES	Molar Mass/Mole	Density	Melting point	Boiling point	Solubility in water	Vapor pressure	Hydroxyl
C#N	27.025 g/mol	0.6875 g/cm3	-13.28 C	25.63 C	Miscible	630 mmHg	FALSE
C1=CC=CC=C1	78.11 g/mol	0.8756 g/cm3	5.558 C	80 C	3%	19.5 mmHg	FALSE
C1CC2=CC=CC=C2N1	119.16 g/mol	1.063 g/mL	-21 C	220 C	5 g/L		FALSE
C1CCC=CC1	82.14 g/mol	0.81 g/cm3	-154 F	181 F	Insoluble	67 mmHg	FALSE
C=O	30.026 g/mol	0.815 g/cm3	-92C	-19.1C	Miscible	1.067	FALSE
CC	30.07 g/mol	1.3562 g/L	-88.6 C	-183 C	60.2 mg/L	3850kPa	FALSE
CC(C)=O	58.08 g/mol	0.791 g/cm3	-137 F	133 F	Miscible	231.0 mmHg	FALSE
CC(C)C	58.12 g/mol	0.557 g/cm3	-427.5 F	-11.7 C		2610.0 mmHg	FALSE
CC(C)CC=O	86.13 g/mol	0.785 g/cm3	-60 F	92.5 C	Slightly soluble	50 mmHg	FALSE
CC(C)OC=O	88.11 g/mol	0.8728 g/cm3	-80 C	68.2 C	Slightly soluble	138.0 mmHg	FALSE
CC(N)=O	59.07 g/mol	1.16 g/cm3	180.1 F	430 F	2250 mg/mL	0.0182 mmHg	FALSE
CC(O)C#N	71.08 g/mol	0.991 g/mL	-40 C	221 C	Miscible	10 mmHg	FALSE
CC(O)CC=O	88.11 g/mol	1.103 g/cm3		83 C	Miscible	21 mmHg	FALSE
CC=O	44.05 g/mol	0.78g/cm3	-123.4 C	20.8 C	miscible	902 mmHg	FALSE
CC1=CC=CC(C)=C1	106.16 g/mol	0.8698 g/cm3	-47.85 C	25.0 C	146 mg/L	8.29 mmHg	FALSE
CC1=CC=CC=C1	92.14 g/mol	0.8623 g/cm3	-94.9C	110.6C	526mg/L	28.4 mmHg	FALSE
CC1=CNC=N1	82.10 g/mol	1.0416g/cm3	56 C	263 C		0.01 mmHg	FALSE
CCC(F)(F)F	98.07 g/mol	1.038 g/cm3	-148C	-12C		2770 mmHg	FALSE
CCC#N	55.08 g/mol	772 mg/mL	-91.8 C	97.2 C	0.119	40 mmHg	FALSE
CCCC(=O)CC	100.16 g/mol	0.812-0.818	-55.5 C	124.00 C	slightly soluble	13.9 mmHg	FALSE
CCCC(=O)OC	102.13 g/mol		-85.8 C	102.8 C	15 mg/mL	32.3 mmHg	FALSE
CCCC#N	69.11 g/mol	0.7936mg/ml	-112 C	117.5 C	3%	20.68 mmHg	FALSE
CCCCC	72.15 g/mol	0.6262 g/cm3	-129.67 C	36.06 C	38 mg/L	514 mmHg	FALSE
CCCCC#N	83.13 g/mol	0.8008 g/cm3		-141 F	1 to 5 mg/mL	7.3 mmHg	FALSE

LLMs meet Molecular Search

1. Naïve Search
2. Search with RAG
3. Search with RAG + additional rules (generated by ChatGPT)
4. Search with RAG + additional rules + Auto-CoT

LLMs:

GPT3.5/GPT4, Vicuna 1.5 7B/13B

Two Chemical LLMs: ChemLLM 7B, ChemDFM 13B

LLMs meet Molecular Search

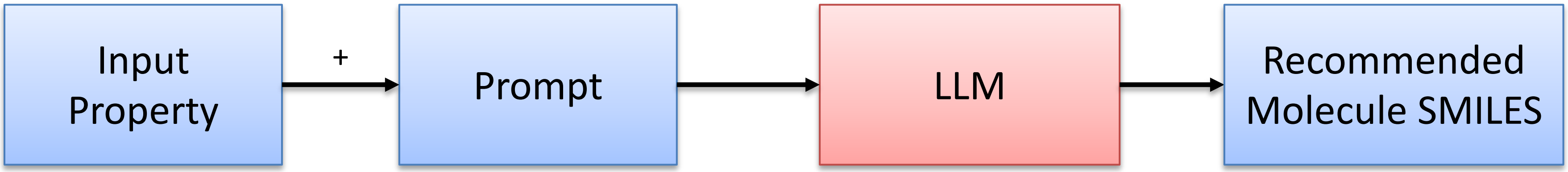
1. Naïve Search

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:

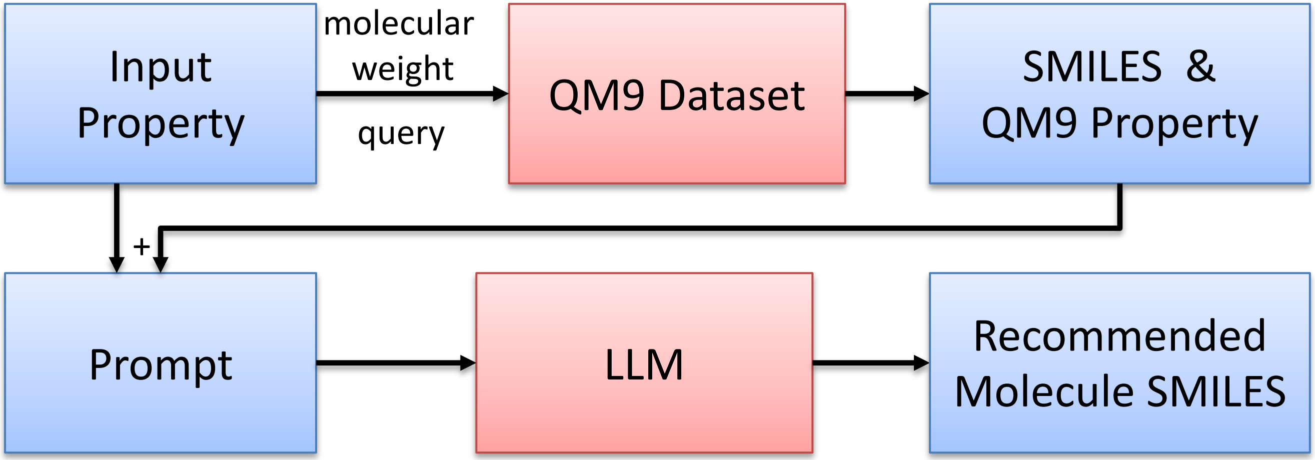
{PROPERTIES}

The top 10 recommended SMILES:



LLMs meet Molecular Search

2. Search with RAG on QM9



Input Property

molecular weight query

QM9 Dataset

SMILES & QM9 Property

+

Prompt

LLM

Recommended Molecule SMILES

LLMs meet Molecular Search

2. Search with RAG on QM9

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}

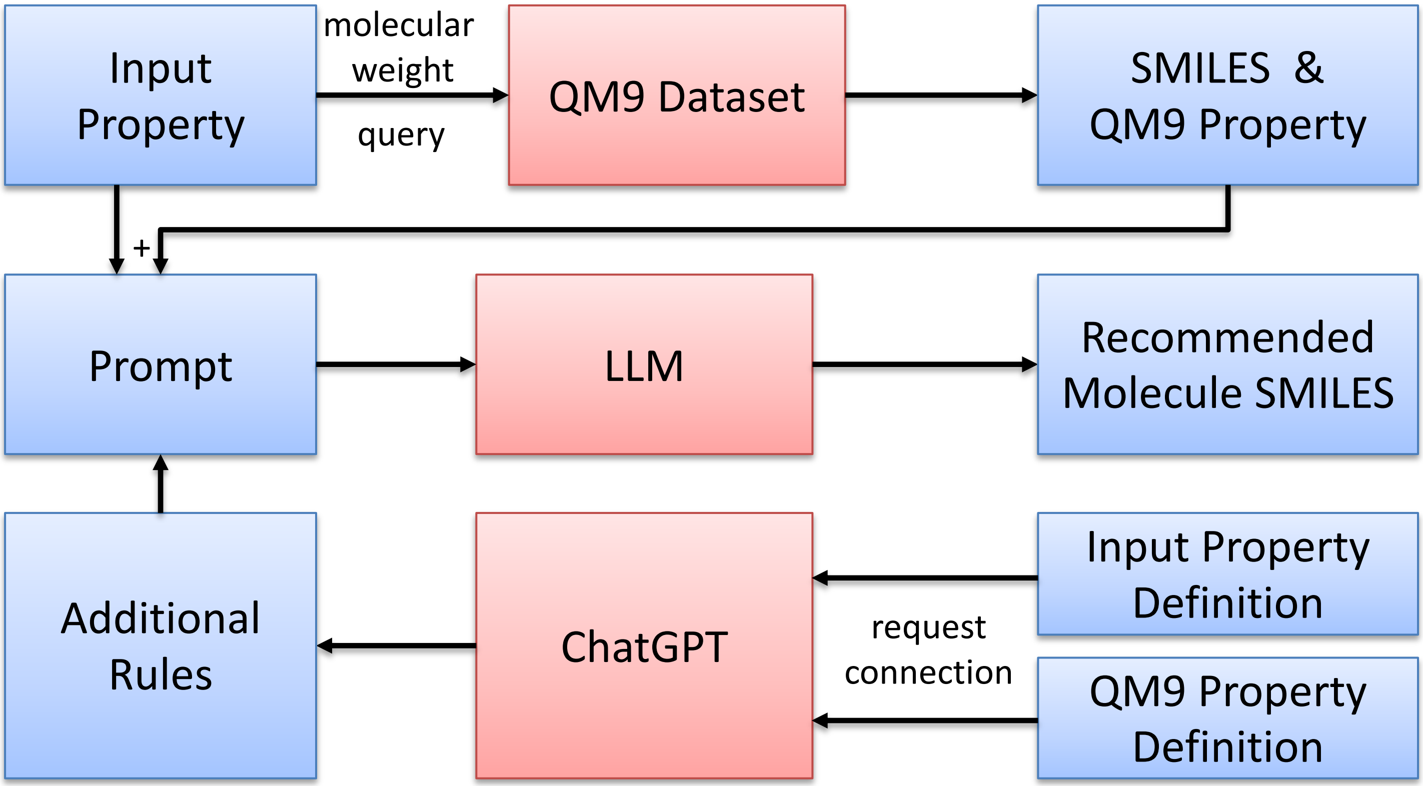
***System:** I referred to a large molecular dataset and I found some molecules and their properties that might qualify:*

{QM9_REFERENCE}

Based on my chemical knowledge and the references, here are the top 10 recommended SMILES:

LLMs meet Molecular Search

3. Search with RAG on QM9 with additional rules



LLMs meet Molecular Search

3. Search with RAG on QM9 with additional rules

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}

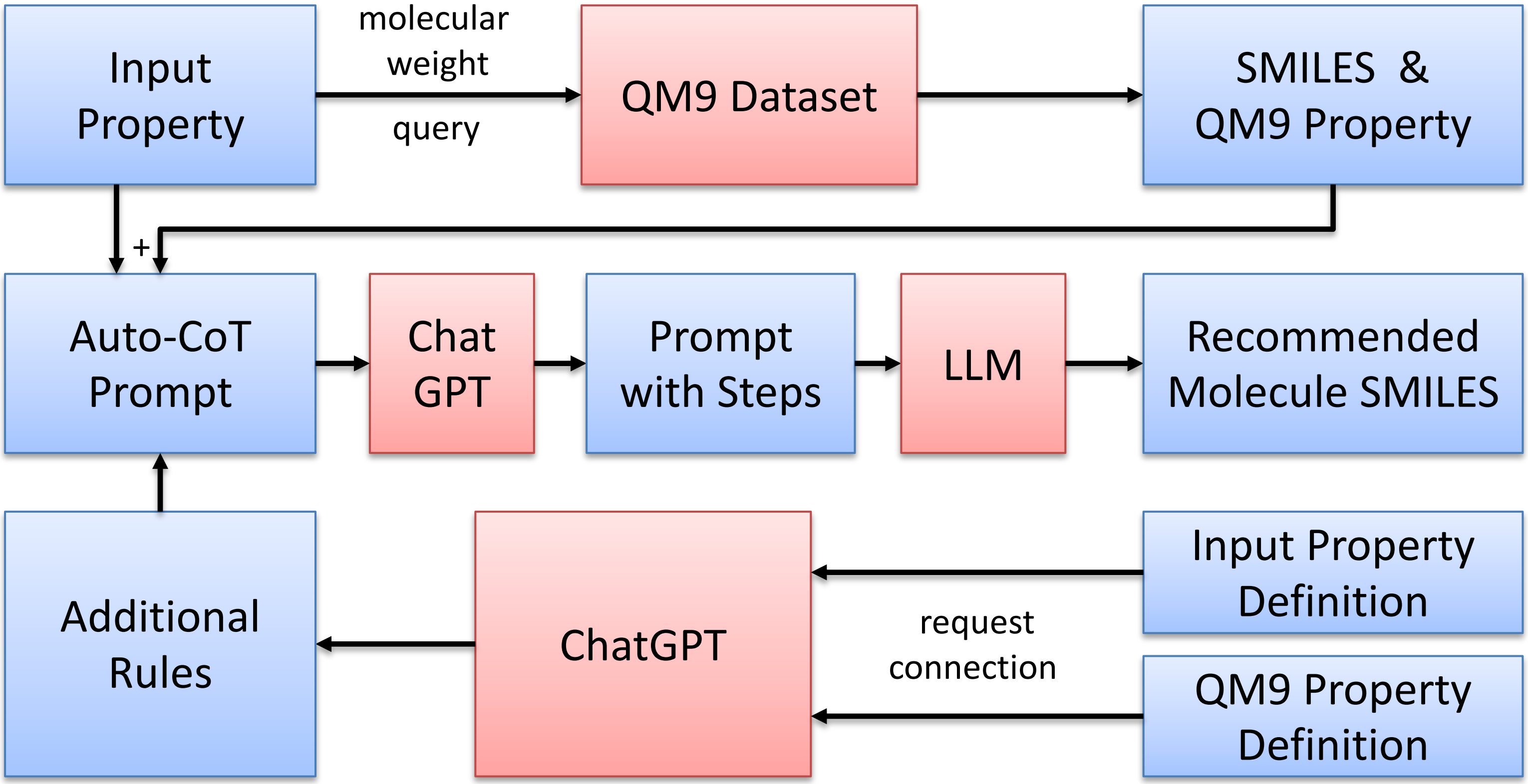
Based on my chemical knowledge I will use the following rules to make molecule recommendations:

{RULES_GENERATED_BY_CHATGPT}

Here are the top 10 recommended SMILES:

LLMs meet Molecular Search

4. Search with RAG + additional rules + Auto-CoT



Input Property

molecular weight

query

QM9 Dataset

SMILES & QM9 Property

+

Auto-CoT Prompt

Chat GPT

Prompt with Steps

LLM

Recommended Molecule SMILES

Additional Rules

ChatGPT

request connection

Input Property Definition

QM9 Property Definition

LLMs meet Molecular Search

4. Search with RAG + additional rules + Auto-CoT

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：

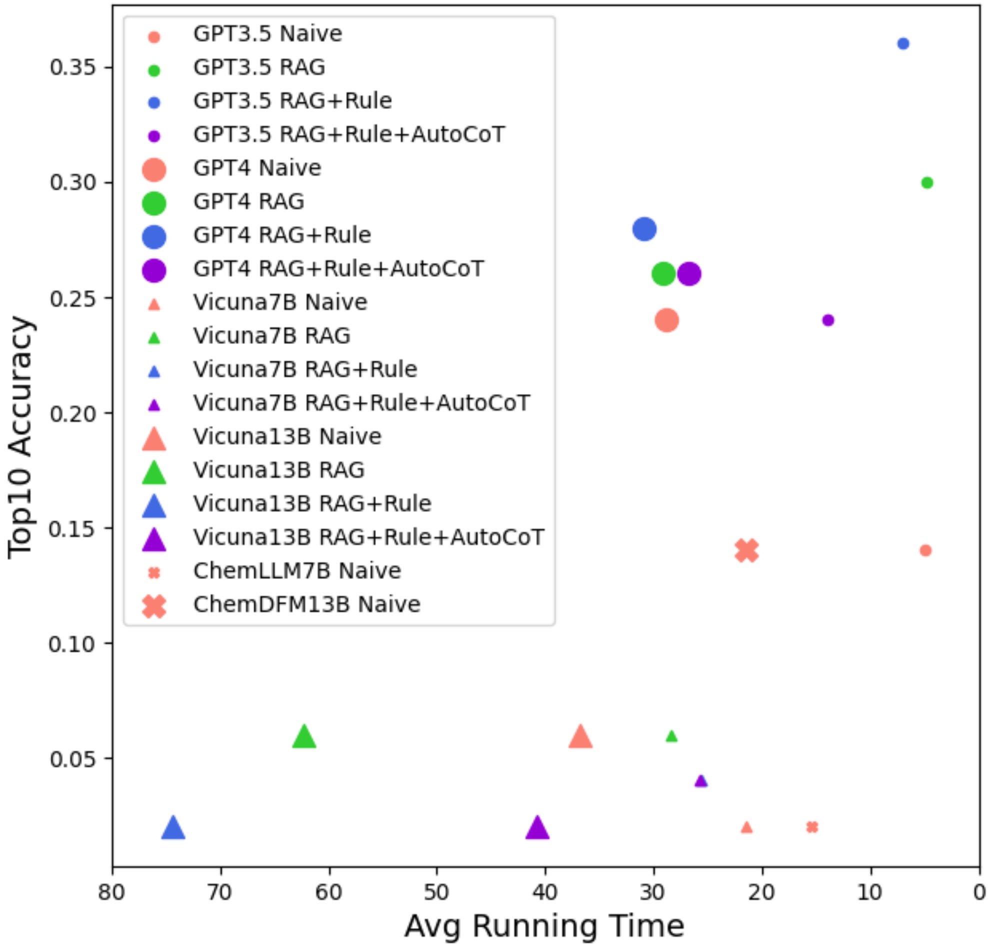
{RULES_GENERATED_BY_CHATGPT}

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:

Experiment Results



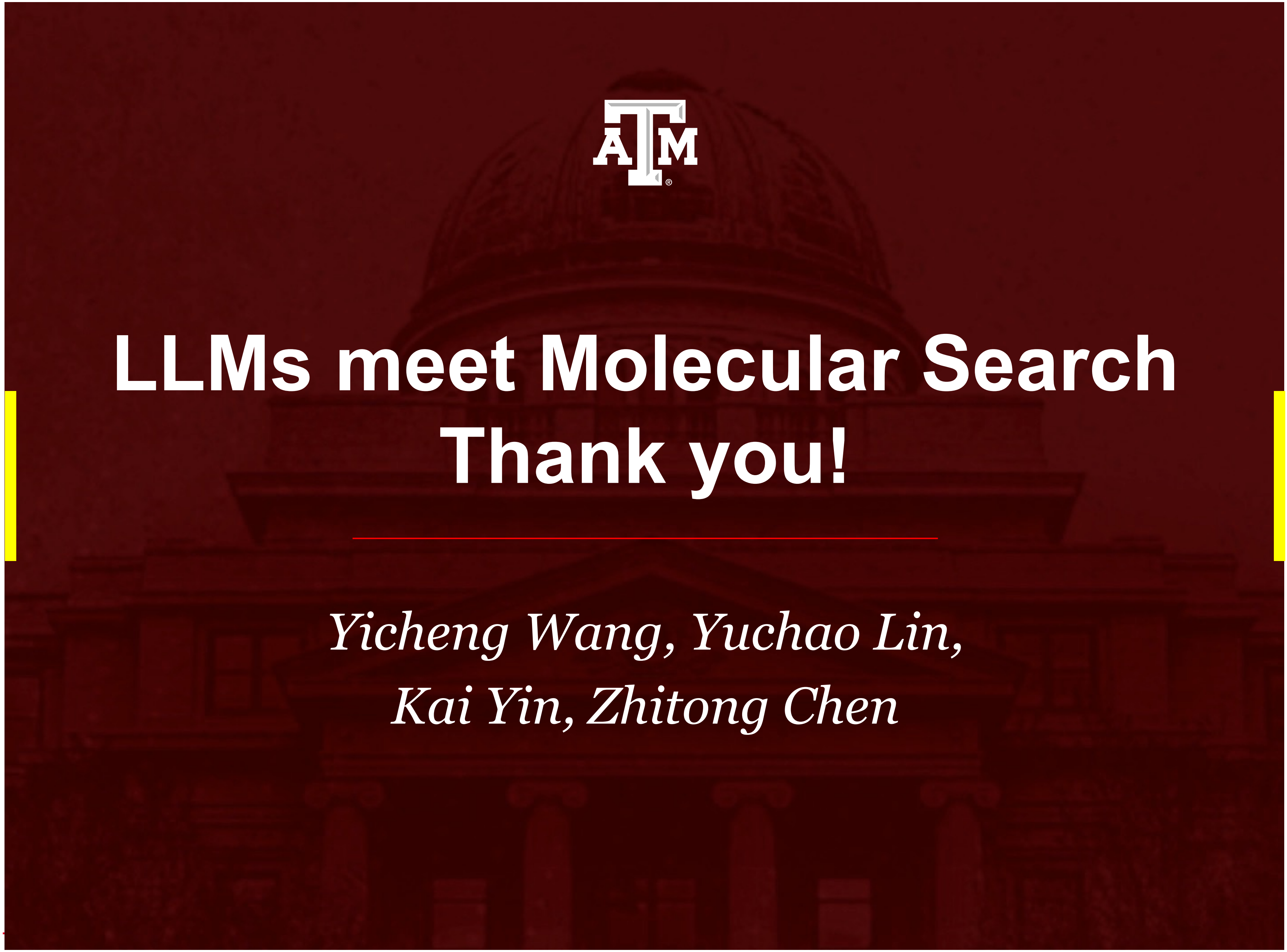
Best approach:

GPT3.5 with RAG + additional rules

GPT4 is less effective than GPT3.5

Open-sourced models (Vicuna) are the worst

Chemical LLMs do not have the performance they clam (better than GPT4/3.5)



LLMs meet Molecular Search
Thank you!

Yicheng Wang, Yuchao Lin,
Kai Yin, Zhitong Chen