MolPuzzle

Can LLMs Solve Molecule Puzzles? A Multimodal Benchmark for Molecular Structure Elucidation

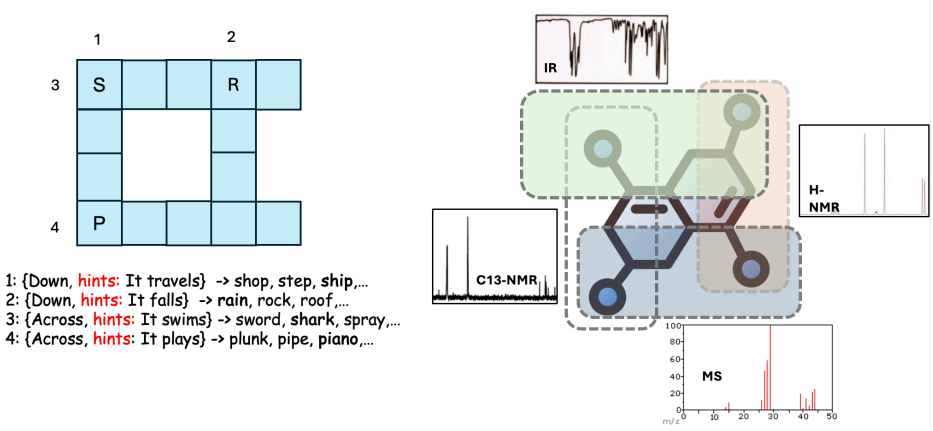
Kehan Guo[1,*], Bozhao Nan[2,*], Yujun Zhou[1,], Taicheng Guo[1], Zhichun Guo[1], Mihir Surve[2], Zhenwen Liang[1], Nitesh V. Chawla[1], Olaf Wiest[2], Xiangliang Zhang[1,††]

[1]Department of Computer Science and Engineering, University of Notre Dame,

[2]Department of Chemistry and Biochemitry, University of Notre Dame,

(* Equal contribution,†† Corresponding author)

Slides Code Data Paper



Comparison of molecular structure elucidation to solving a crossword puzzle. Just as crossword clues provide hints for fitting words into a grid, spectroscopic data such as NMR, IR, and mass spectrometry offer complementary clues about a molecule's structure. Integrating these diverse clues leads to a complete and consistent picture of the molecule, similar to how words fit together in a puzzle.

Introduction

Artificial intelligence (AI) is revolutionizing chemistry, with significant impacts on industrial chemical engineering, drug discovery, and education. Large language models (LLMs) have successfully addressed predictive tasks such as molecular property prediction, reaction prediction, and experiment automation. Here, we introduce **molecular structure elucidation**, a task that presents a new challenge for AI. **This task requires integrating diverse spectroscopic data, iterative hypothesis testing, and deep chemical reasoning to determine a molecule's structure**. Much like solving a complex crossword puzzle, it involves piecing together clues to form a coherent solution. The Figure highlights this analogy, illustrating the similarities in strategy and complexity between molecular structure elucidation and solving a crossword puzzle.

In this work, we present a novel approach to molecular structure elucidation, adapting the task for Large Language Models (LLMs) to explore their potential in chemical research. **Our primary contribution is the introduction of the MolPuzzle dataset, comprising 234 complex structure elucidation challenges involving multimodal data like IR, MASS, H-NMR, and C-NMR spectra, as well as molecular formulas.** Each instance requires LLMs to navigate three key sub-tasks: molecule understanding, spectrum interpretation, and molecule construction.

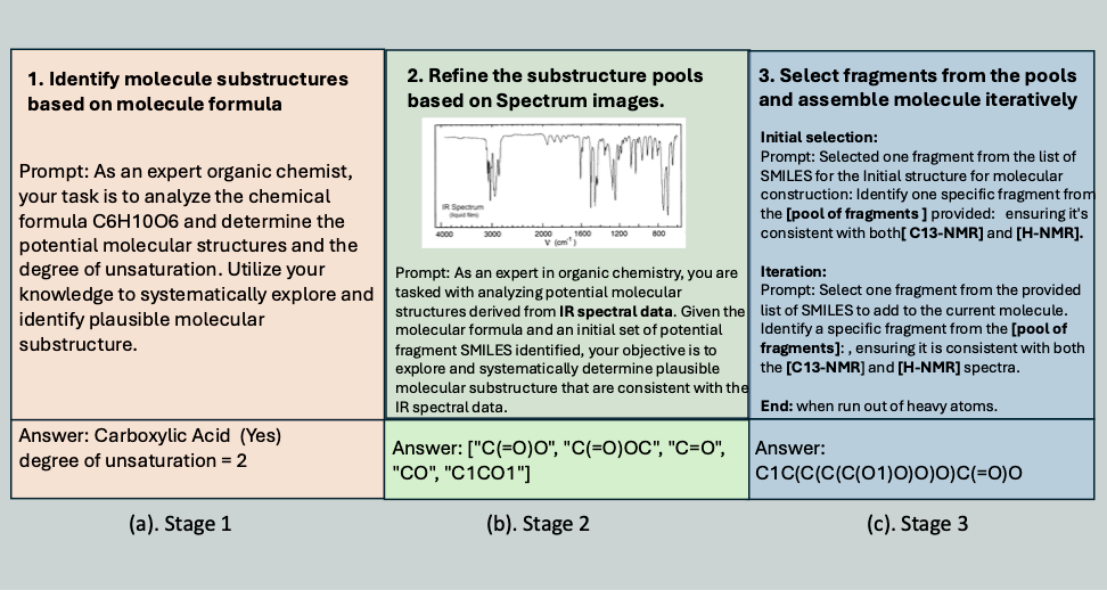
We tested 11 state-of-the-art LLMs, including GPT-4o and Claude-3-opus, alongside human benchmarks. Key findings include: **(1) GPT-4o outperforms other models but still underperforms compared to humans, with only 1.4% of its answers exactly matching the ground truth;(2) LLMs struggle particularly in spectrum interpretation and molecule construction.**

In summary, our contributions are two-fold: Our contributions are twofold: **(1) A new reasoning challenge for the AI community focused on complex problem-solving in chemistry; and (2) New AI tools for the chemistry community, showcasing LLMs' potential to accelerate molecular structure elucidation and inspire interdisciplinary collaboration.**

Overview of the MolPuzzle Benchmark

The MolPuzzle benchmark is designed to test the reasoning capabilities of Large Language Models (LLMs) in molecular structure elucidation tasks. This dataset contains 200 instances of molecular structure elucidation challenges, simulating real-world chemistry tasks. Each instance in MolPuzzle involves three interlinked sub-tasks:

- **Molecule Understanding:** This stage evaluates the model's ability to identify and understand basic molecular structures, starting from the molecular formula derived from mass spectrometry data. The dataset includes questions about the degree of saturation, aromatic rings, and functional groups, helping the model narrow down possible molecular structures.
- **Spectrum Interpretation:** This stage involves analyzing multimodal data, including IR, MASS, 1H-NMR, and 13C-NMR spectra. These spectral images provide critical information about functional groups, molecular mass, and the arrangement of atoms. The dataset challenges models to integrate these clues and refine molecular hypotheses based on the spectral data.
- **Molecule Construction:** In this final stage, the models attempt to assemble the molecule based on the information gathered from previous steps. This involves constructing a valid molecular structure that fits the constraints provided by the NMR data.



In total, Molpuzzle includes 23,678 data examples collected from each Stage.

Statistic	Number
Total MolPuzzle Instances	217
Stage-1 QA samples	5,859
- Num. of molecule formula	176
- Max question length	128
- Average question length	94
Stage-2 QA samples	11,501
- Num. of spectrum images	868
- Max question length	340
- Average question length	264
Stage-3 QA samples	6,318
- Maximum Iteration	7
- Max question length	356
- Average question length	238

Figure 3: Statistic of the MolPuzzle dataset

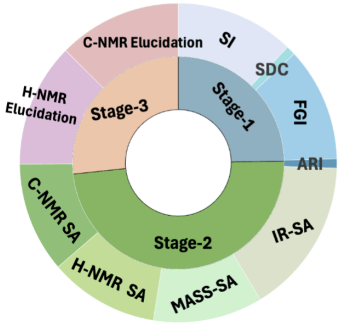


Figure 4: Inner ring: sample distribution in 3 stages. Outer ring: sample distribution across categories in each stage. SI: saturation identification, SDC: saturation degree calculation, FGI: functional group identification, ARI: aromatic ring identification, SA: spectrum analysis.

Experiment Results

Addressing individual QA tasks in three stages

We first conducted evaluation of a variety of LLMs for completing the individual tasks in each stage, including GPT-4o, GPT-3.5-turbo, Claude-3-opus, Gemini-pro, LLama-3-8B-Instruct, Vicuna-13B-v1.5, Mistral-7B-Instruct-v0.3, and in particular multimodal LLMs such as Gemini-pro-vision, LLava-Llama-3-8B, Qwen-VL-Chat, and InstructBlip-Vicuna-7B/13B.

Method	Stage 1 (Molecule Understanding) Tasks			
	SI	ARI	FGI	SDC
GPT-4o	**1.00±0.000**	0.943±0.016	0.934±0.005	0.667±0.003
GPT-3.5-turbo	0.451±0.025	0.816±0.017	0.826±0.075	0.5±0.099
Claude-3-opus	0.361±0.009	**0.988±0.015**	**0.934±0.001**	**0.856±0.016**
Llama3	0.228±0.043	0.696±0.051	0.521±0.003	0.000±0.000
Human	**1.00±0.000**	**1.000±0.000**	0.890±0.259	0.851±0.342

Method	Stage 2 (Spectrum Interpretation) Tasks			
	IR Interpretation	MASS Interpretation	H-NMR Interpretation	C-NMR Interpretation
GPT-4o	**0.656±0.052**	**0.609±0.042**	**0.618±0.026**	**0.639±0.010**
LLava	0.256±0.026	0.101±0.021	0.118±0.008	0.254±0.015
Human	0.753±0.221	0.730±0.110	0.764±0.169	0.769±0.101

Method	Stage 3 (Molecule Construction) Tasks	
	H-NMR Elucidation	C-NMR Elucidation
GPT-4o	**0.524±0.021**	**0.506±0.037**
Llama3	0.341±0.015	0.352±0.017
Human	0.867±0.230	0.730±0.220

Table 1: F1 scores (↑) of individual QA tasks in three stages. The best LLMs results are in bold font.

Tasks in stage 1 are SI: Saturation Identification, ARI: Aromatic Ring Identification, FGI: Functional Group Identification, and SDC: Saturation Degree Calculation.

Addressing entire molecule puzzles

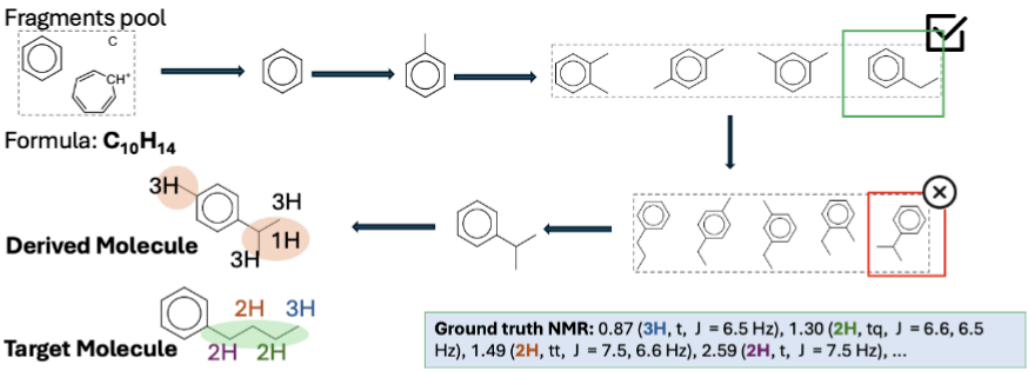
For solving the entire molecule puzzles, the evaluation is limited to the three most advanced multimodal LMMs: GPT-4o, Claude-3-opus, and Gemini-pro, due to the involvement of spectrum image analysis in Stage 2.

Method	Acc. (↑)	Levenshtein (↓)	Validity (↑)	MACCS FTS (↑)	RDK FTS (↑)	Morgan FTS (↑)
GPT-4o	**0.014±0.004**	**11.653±0.013**	**1.000±0.000**	**0.431±0.009**	**0.293±0.013**	0.232±0.007
Claude-3-opus	0.013±0.008	12.680±0.086	**1.000±0.000**	0.383±0.050	0.264±0.040	0.241±0.037
Gemini-pro	0.000±0.000	12.711±0.196	**1.000±0.000**	0.340±0.017	0.208±0.002	0.171±0.007
Human	0.667±0.447	1.332±2.111	**1.000±0.000**	**0.985±0.022**	0.795±0.317	0.810±0.135

Table 2: The performance of LLMs and human baseline in solving MolPuzzle. The best LLM results are in bold font. Acc. stands for the Accuracy of Exact

Match.

Success and Failure Analysis



Error in solving the molecule puzzle

The Figure presents case studies that illustrate the iterative steps involved in Stage 3, showcasing the most common errors made by GPT-4o: **the accumulation of errors in iterative steps, which can lead to catastrophic failures.** Note that this stage focuses on selecting the correct fragments and assembling them step by step to form the final molecular structure. We find that **GPT-4o can initially succeed in picking the correct fragment when the structure is comparatively simple. However, as the process progresses, it does no select structures that satisfy all the requirements indicated by the NMR data.**

BibTeX

```
{@inproceedings{guocan,
        title={Can LLMs Solve Molecule Puzzles? A Multimodal Benchmark for Molecular
        author={Guo, Kehan and Nan, Bozhao and Zhou, Yujun and Guo, Taicheng and Guo,
        booktitle={The Thirty-eight Conference on Neural Information Processing Syste
```