

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

Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve
Zhenwen Liang, Nitesh V. Chawla, Olaf Wiest, Xiangliang Zhang

[Department of Computer Science and Engineering, University of Notre Dame, US](#)

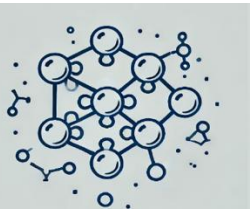


Paper Website



NSF Center for Computer
Assisted Synthesis

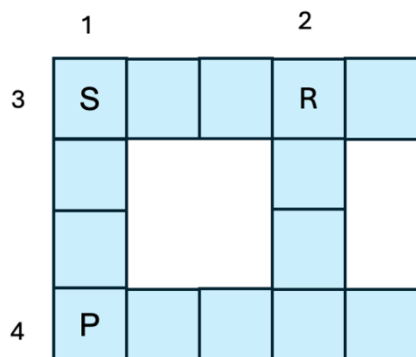




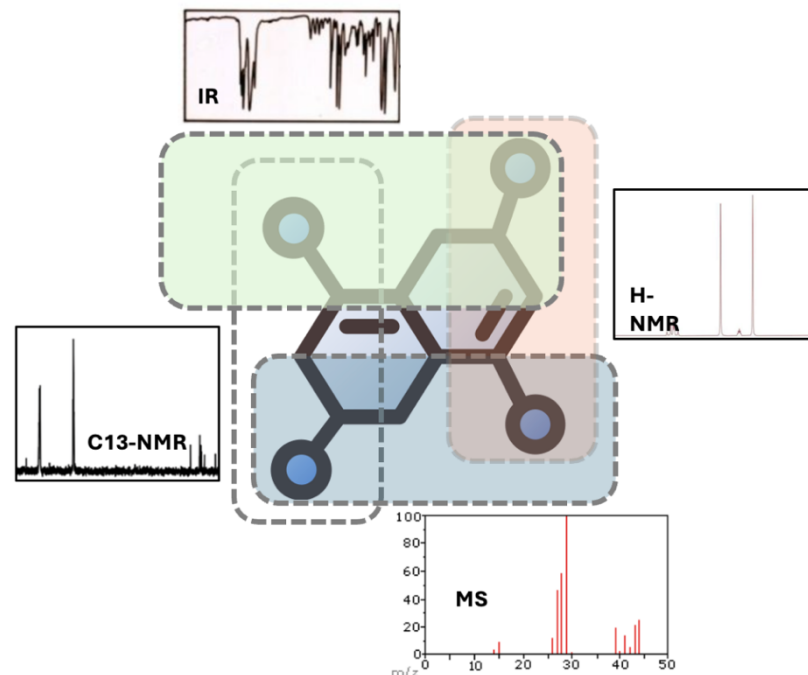
What is Structure Elucidation?

Definition: The process of determining the molecular structure of a compound based on spectroscopic data such as Nuclear Magnetic Resonance (NMR), Infrared (IR), and Mass Spectrometry (MS).

Solving Structure Elucidation is like solving a word puzzle

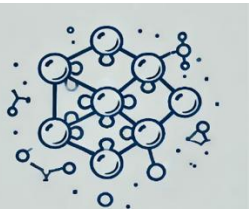


- 1: {Down, **hints:** It travels} -> shop, step, ship,...
- 2: {Down, **hints:** It falls} -> rain, rock, roof,...
- 3: {Across, **hints:** It swims} -> sword, shark, spray,...
- 4: {Across, **hints:** It plays} -> plunk, pipe, piano,...



Word Puzzle

Molecular Puzzle



Introduction & Motivation

For all Chemistry undergraduate students, Structure Elucidation

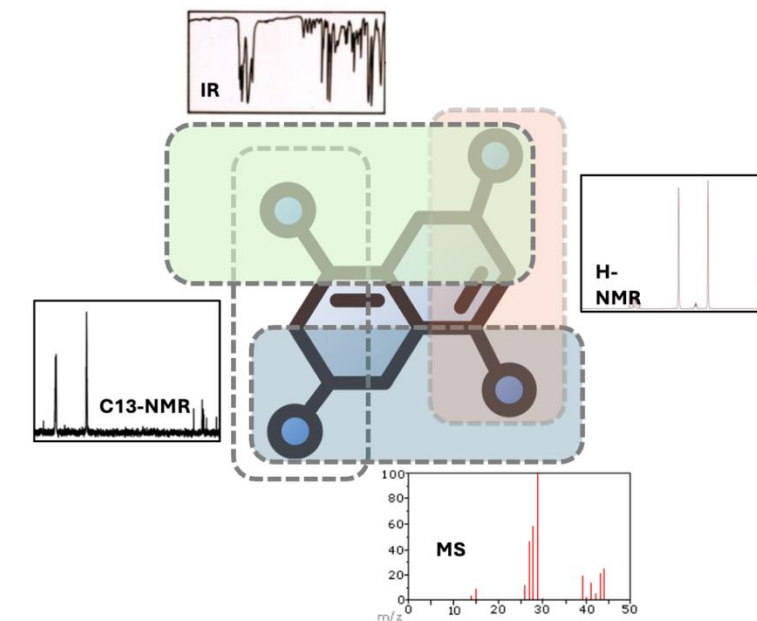
- is a long-standing problem and fundamental part of their curriculum.
- is a key skill learned in organic chemistry courses, allowing them to identify the structure of unknown compounds based on their spectral characteristics.

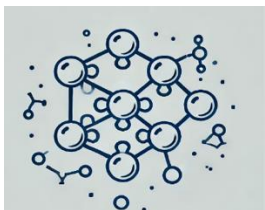
Our research goal:

- Introduce this challenging reasoning problem to the AI community: MolPuzzle, a benchmark for this problem (217 instances)
- Answer the question: Can LLMs perform better than Chemistry students on solving these puzzles?



VS





Our MolPuzzle Benchmark

- **Dataset Overview:**

- 217 instances with over 23,000 QA samples.
- Three interlinked tasks: Molecule understanding, spectrum interpretation, molecule construction.

- **Unique Aspect:** Multimodal reasoning tasks incorporating IR, MASS, NMR data.

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

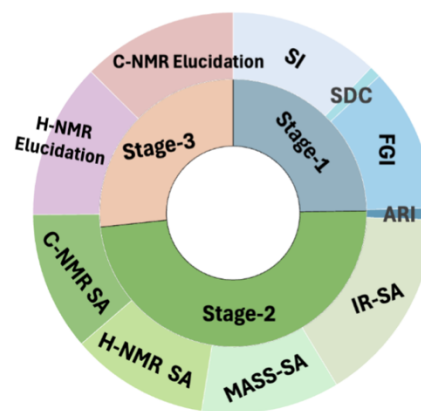
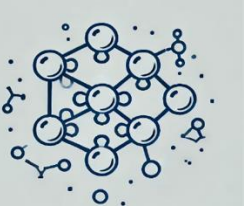
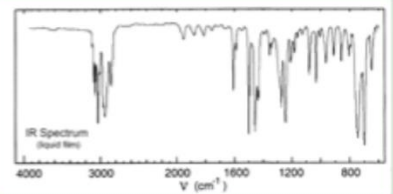


Figure 4: Inner ring: sample distribution in 3 stages. Outer ring: sample distribution across categories in each stage. SI: saturation identification, SDC: saturation degree calculation, FGI: functional group identification, ARI: aromatic ring identification, SA: spectrum analysis.



Reasoning Tasks in MolPuzzle

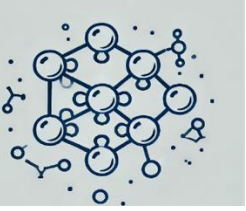
- **Stage 1:** Molecule Understanding (e.g., functional group identification)
- **Stage 2:** Spectrum Interpretation (analyzing IR, H-NMR, C-NMR data)
- **Stage 3:** Molecule Construction (assembling based on spectra)
- **Data:** Derived from curated spectra, RDKit validation for molecule accuracy.

<p>1. Identify molecule substructures based on molecule formula</p> <p>Prompt: As an expert organic chemist, your task is to analyze the chemical formula C₆H₁₀O₆ and determine the potential molecular structures and the degree of unsaturation. Utilize your knowledge to systematically explore and identify plausible molecular substructure.</p>	<p>2. Refine the substructure pools based on Spectrum images.</p>  <p>Prompt: As an expert in organic chemistry, you are tasked with analyzing potential molecular structures derived from IR spectral data. Given the molecular formula and an initial set of potential fragment SMILES identified, your objective is to explore and systematically determine plausible molecular substructure that are consistent with the IR spectral data.</p>	<p>3. Select fragments from the pools and assemble molecule iteratively</p> <p>Initial selection: Prompt: Selected one fragment from the list of SMILES for the Initial structure for molecular construction: Identify one specific fragment from the [pool of fragments] provided: ensuring it's consistent with both [C13-NMR] and [H-NMR].</p> <p>Iteration: Prompt: Select one fragment from the provided list of SMILES to add to the current molecule. Identify a specific fragment from the [pool of fragments]: , ensuring it is consistent with both the [C13-NMR] and [H-NMR] spectra.</p> <p>End: when run out of heavy atoms.</p>
<p>Answer: Carboxylic Acid (Yes) degree of unsaturation = 2</p>	<p>Answer: ["C(=O)O", "C(=O)OC", "C=O", "CO", "C1CO1"]</p>	<p>Answer: C1C(C(C(C(O1)O)O)O)C(=O)O</p>

Stage 1

Stage 2

Stage 3



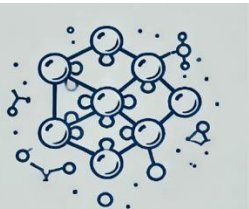
LLMs Evaluated

- **Models Tested:** GPT-4o, Claude-3-opus, etc.
- **Approach:** Zero-shot evaluations, comparisons with human baselines.



VS







LLMs Evaluated



Table 1: F1 scores (\uparrow) 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.



Stage 1 (Molecule Understanding) Tasks				
Method	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
Galactica-30b	0.826±0.248	0.347±0.000	0.467±0.005	0.000±0.000
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

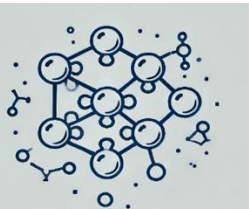
Stage 2 (Spectrum Interpretation) Tasks				
Method	IR Interpretation	MASS Interpretation	H-NMR Interpretation	C-NMR Interpretation
GPT-4o	0.656±0.052	0.609±0.042	0.618±0.026	0.639±0.010
LLava	0.256±0.026	0.101±0.021	0.118±0.008	0.254±0.015
Human	0.753±0.221	0.730±0.11	0.764±0.169	0.769±0.101

Stage-3 (Molecule Construction) Tasks		
Method	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

 As good as 

 A litter worse than 

 much worse than 



LLMs Evaluated

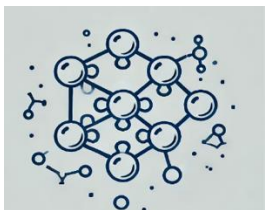
Table 2: The performance of LLMs and human baseline in solving MolPuzzle. The best LLM results are in bold font. Acc. stands for the Accuracy of Exact Match.

Method	Acc. (\uparrow)	Levenshtein (\downarrow)	Validity (\uparrow)	MACCS FTS (\uparrow)	RDKit FTS (\uparrow)	Morgan FTS (\uparrow)
GPT-4o	0.014\pm0.004	11.653\pm0.013	1.000\pm0.000	0.431\pm0.009	0.293\pm0.013	0.232 \pm 0.007
Claude-3-opus	0.013 \pm 0.008	12.680 \pm 0.086	1.000\pm0.000	0.383 \pm 0.050	0.264 \pm 0.040	0.241\pm0.037
Gemini-pro	0.000 \pm 0.000	12.711 \pm 0.196	1.000\pm0.000	0.340 \pm 0.017	0.208 \pm 0.002	0.171 \pm 0.007
Human	0.667 \pm 0.447	1.332 \pm 2.111	1.000 \pm 0.000	0.985 \pm 0.022	0.795 \pm 0.317	0.810 \pm 0.135

- **Performance Insights:**

- LLMs excel in molecule understanding but struggle with spectrum interpretation and molecule construction.
- Top LLM (GPT-4o) achieved 1.4% exact match accuracy on full tasks.

- **Gap Analysis:** Significant room for improvement, especially in complex reasoning tasks.



Challenges & Future Directions

- **Challenges:**

- Complex spectral data interpretation.
- Iterative molecule construction processes.

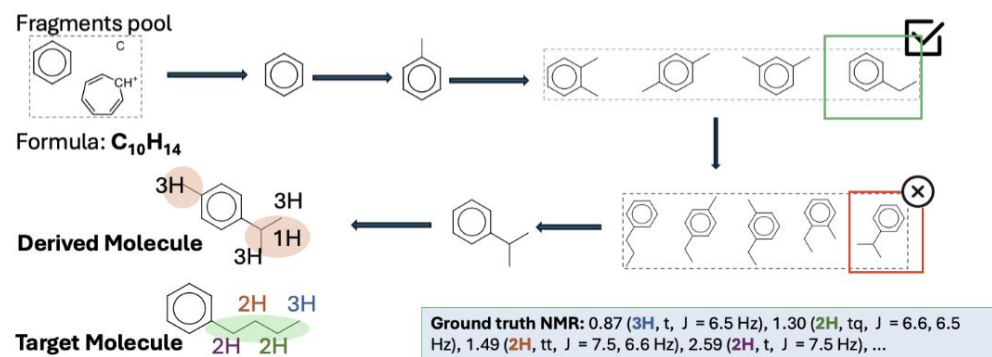


Figure 5: The target molecule contains four distinct non-aromatic hydrogen types, color-coded in the ground truth NMR. However, the model-derived molecule shows hydrogen counts of 3, 3, and 1, differing from the ground truth. The mismatch between the hydrogen types in the green section of the target molecule and the orange region of the predicted molecule results in incorrect fragment selection and assembly.

- **Future Focus:**

- Specialized LLM training for visual and chemical data.
- Advanced planning and reasoning strategies.



Questions

Thank You For Your Attention!

MolPuzzle Website:

<https://kehanguo2.github.io/Molpuzzle.io/>

Feel free to reach out to us at kguo2@nd.edu. We welcome collaborations aimed at enhancing the reasoning capabilities of LLMs in the scientific domain.