

---

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

---

Kehan Guo<sup>1\*</sup>, Bozhao Nan<sup>2\*</sup>, Yujun Zhou<sup>1</sup>, Taicheng Guo<sup>1</sup>, Zhichun Guo<sup>1</sup>, Mihir Surve<sup>2</sup>,  
Zhenwen Liang<sup>1</sup>, Nitesh V. Chawla<sup>1</sup>, Olaf Wiest<sup>2</sup>, Xiangliang Zhang<sup>1†</sup>

<sup>1</sup>Department of Computer Science and Engineering, University of Notre Dame

<sup>2</sup>Department of Chemistry and Biochemistry, University of Notre Dame

{kguo2, bnan, xzhang33}@nd.edu

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

## Abstract

1 Large Language Models (LLMs) have shown significant problem-solving capabilities  
2 across predictive and generative tasks in chemistry. However, their proficiency  
3 in multi-step chemical reasoning remains underexplored. We introduce a new  
4 challenge: molecular structure elucidation, which involves deducing a molecule’s  
5 structure from various types of spectral data. Solving such a molecular puzzle,  
6 akin to solving crossword puzzles, poses reasoning challenges that require inte-  
7 grating clues from diverse sources and engaging in iterative hypothesis testing. To  
8 address this challenging problem with LLMs, we present **MolPuzzle**, a benchmark  
9 comprising 217 instances of structure elucidation, which feature over 23,000 QA  
10 samples presented in a sequential puzzle-solving process, involving three inter-  
11 linked sub-tasks: molecule understanding, spectrum interpretation, and molecule  
12 construction. Our evaluation of 12 LLMs reveals that the best-performing LLM,  
13 GPT-4o, performs significantly worse than humans, with only a small portion  
14 (1.4%) of its answers exactly matching the ground truth. However, it performs  
15 nearly perfectly in the first subtask of molecule understanding, achieving accuracy  
16 close to 100%. This discrepancy highlights the potential of developing advanced  
17 LLMs with improved chemical reasoning capabilities in the other two sub-tasks.  
18 Our MolPuzzle dataset and evaluation code are available at this [link](#).

## 19 1 Introduction

20 Artificial intelligence (AI) is revolutionizing the field of chemistry, influencing diverse sectors such as  
21 industrial chemical engineering [1, 2], drug discovery [3], and chemistry education [4]. In particular,  
22 recent studies have highlighted the success of large language models (LLMs) in addressing predictive  
23 challenges in chemistry, including molecular property prediction [5], reaction prediction [6], and  
24 experiment automation [7]. These advancements suggest significant potential for AI to enhance  
25 efficiency and innovation across these critical areas.

26 We introduce a new chemical challenge to AI, **molecular structure elucidation**. While this critical  
27 task has been explored in other contexts, it remains unexplored for large language models (LLMs),

---

\*Both authors contributed equally to this work, supported by the NSF Center for Computer-Assisted Synthesis (C-CAS), <https://ccas.nd.edu>

†Corresponding author.

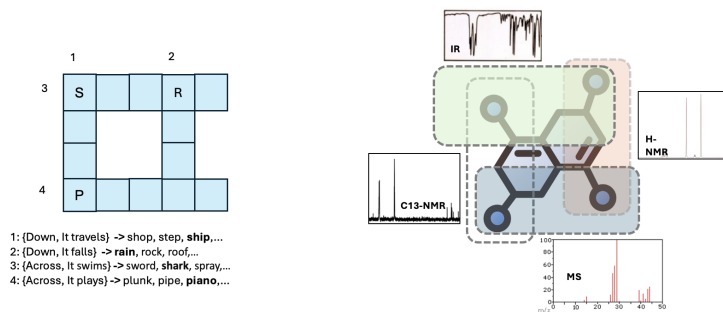


Figure 1: A crossword puzzle (left), and a molecular structure elucidation puzzle (right)

28 extending beyond familiar predictive and generative domains such as property or reaction prediction,  
 29 and representing a shift toward complex problem-solving. Analogous to solving a detailed cross-  
 30 word puzzle, **molecular structure elucidation** can be seen as a **molecular puzzle**. It requires the  
 31 integration of multifaceted data, iterative hypothesis testing, and a deep understanding of chemical  
 32 cues, much like piecing together clues across a crossword grid to form a coherent solution. Fig. 1  
 33 illustrates the problem of molecular structure elucidation alongside its analogical counterpart, the  
 34 crossword puzzle, highlighting the parallels in strategy and complexity between these two intellectual  
 35 challenges.

36 Just as a crossword puzzle requires figuring out words based on given clues and fitting them together  
 37 in a grid, molecular structure elucidation involves deducing a molecule’s structure from various types  
 38 of data such as nuclear magnetic resonance (NMR), infrared spectroscopy (IR), mass spectrometry,  
 39 and others. Each type of data provides clues about different aspects of the molecular structure. In  
 40 a crossword, we integrate clues from across different directions and hints to form words that fit  
 41 together correctly. Similarly, in molecular structure elucidation, we need to integrate information  
 42 from different spectroscopic methods to form a consistent picture of the molecule. For example,  
 43 IR spectra reveal molecular vibrations and functional groups, NMR provides information about  
 44 the framework of hydrogen and carbon atoms, while mass spectrometry can offer insights into the  
 45 molecular weight and possible fragmentations.

46 Nevertheless, molecular structure elucidation is a challenging and time-consuming task. Training  
 47 undergraduate students in chemistry to solve these puzzles has been a part of the curriculum because  
 48 determining the structure of molecules is a fundamental skill in the field. Typically, even a single  
 49 molecule puzzle question on a final exam can take 10 to 15 minutes to solve [8], demanding  
 50 considerable memory and processing skills from the students. In the domain of complex molecule  
 51 research, the process of molecular deduction can become even more complex and time-consuming.  
 52 Therefore, fully automating this process is highly beneficial for accelerating the design of new  
 53 materials and drugs, as well as enhancing the efficiency of chemical research [9, 10]. However, it  
 54 remains a challenging task due to the complexities involved in interpreting spectral data and solving  
 55 intricate reasoning problems associated with molecular structures [11].

56 In this work, we aim to present molecular structure elucidation in formats that LLMs can effectively  
 57 process. By adapting this complex task to be compatible with LLMs, we explore their potential as  
 58 promising tools in chemical research. If successful, LLMs could significantly accelerate scientific  
 59 discovery in chemistry, transforming how we approach and solve intricate molecular puzzles.

60 To achieve our objectives, we first introduce a novel dataset named **MolPuzzle**, which includes  
 61 234 instances of structure elucidation challenges inspired by common chemistry tasks. Unlike  
 62 datasets used in predictive or generative tasks, which typically consist of a collection of independent  
 63 samples and are relatively straightforward to construct, each instance in the MolPuzzle dataset is  
 64 uniquely complex. It is structured as a sequential process involving three interlinked sub-tasks:  
 65 **molecule understanding**, **spectrum interpretation**, and **molecule construction**. These instances  
 66 are accompanied by multimodal data, including images of IR, MASS, H-NMR, and C-NMR spectra,  
 67 alongside their corresponding molecular formulas. Presenting such a complex, multimodal problem in  
 68 a format that LLMs can effectively process presents a unique challenge. We, a team of AI researchers

69 and chemists, are dedicated to formulating the molecule puzzle instances in descriptive languages  
70 that are accessible to LLMs. Our focus is on ensuring the utility of these instances, as well as their  
71 comprehensive coverage over various scenarios and challenges that mimic real-world conditions. By  
72 doing so, **MolPuzzle** opens the door for LLMs to contribute meaningfully to the field of chemistry,  
73 potentially accelerating scientific discoveries and innovations.

74 Second, we present our effort to automate the solving of molecular structure elucidation using LLMs.  
75 While certain sub-tasks, such as translating an IR spectrum into a molecular formula, may be solvable  
76 by encoder-decoder models [12], the comprehensive resolution of the entire molecular puzzle likely  
77 requires the advanced planning and reasoning capabilities of LLMs. We tested 11 state-of-the-art  
78 LLMs including GPT-4o, Gemini-pro, and Claude-3-opus. We also conducted a human baseline to  
79 compare the performance of humans and LLMs in solving the same puzzles. The **key findings** are:  
80 1) GPT-4o significantly outperforms other LLMs; 2) The best-performing LLM, GPT-4o, performs  
81 significantly worse than humans, with only a small portion (1.4%) of its answers exactly matching  
82 the ground truth; and 3) GPT-4o’s performance primarily collapses in the Stage-2 of spectrum  
83 interpretation and gets worse in the Stage-3 of molecule construction, although it performs nearly  
84 perfectly in Stage-1 of molecule understanding (with accuracy close to 100%).

85 To summarize, our key contributions in this work are the presentation of:

- 86 • **A new reasoning problem for AI community.** As the focus of AI development has evolved  
87 from solving predictive tasks and generative tasks to engaging in complex reasoning tasks—akin  
88 to system 2 level thinking—we introduce a reasoning task centered around molecular structure  
89 elucidation. This crucial problem from the field of chemistry sets a high benchmark for AI models  
90 to reach. Solving this task requires AI models to possess the ability to interpret spectral images,  
91 engage in complex reasoning, and plan effectively across extended workflows. This not only  
92 challenges the current capabilities of AI but also pushes the boundaries of what AI can achieve in  
93 scientific domains, particularly in understanding and manipulating molecular structures.
- 94 • **A new light of AI solutions for chemistry community.** By proposing the **MolPuzzle dataset**,  
95 we establish another bridge between the fields of AI and chemistry. This initiative leverages the  
96 important capabilities of multimodal LLMs, providing the chemistry community with innovative  
97 solutions to accelerate the process of structure elucidation. Our initial exploration serves as a  
98 demonstration of the potential for these technologies. It sets the stage for further collaborative  
99 efforts, inspiring researchers from both domains to collaboratively explore new frontiers in scientific  
100 discovery.

101 The paper is organized as follows. Section 2 presents the related work. In Section 3, we elaborate  
102 on the curation of the MolPuzzle dataset. In Section 4, we report the usage of multimodal LLMs in  
103 solving MolPuzzle. In Section 5, we discuss the main findings and directions opened by this work. In  
104 section 7, we discuss the broader impact of our work. Last, we summarize the study in Section 8 and  
105 offer our conclusions.

## 106 2 Related Work

107 **Molecular Structure Elucidation.** Automated molecular structure determination has been re-  
108 searched for decades [13, 14, 15, 16, 17], initially focusing on rule-based systems [18, 19] that  
109 interpret spectral data using predefined chemical rules and expert knowledge. Notable examples  
110 include SENECA [20], employing genetic algorithms on NMR data, and ACD/Structure Elucidator  
111 [21], a commercial software integrating various spectral data. While effective for well-characterized  
112 compounds, rule-based methods struggle with complex or novel molecules that deviate from es-  
113 tablished patterns, and their proprietary nature limits benchmarking accessibility. Machine learning  
114 approaches [22, 23, 24, 25, 26, 27, 28, 29] have also been explored. Early studies utilized neural  
115 networks to assign infrared spectra to molecular structures [30], and recent advancements leverage  
116 deep learning for complex datasets [31]. For example, Alberts et al. [12] used a transformer-based  
117 model to predict SMILES strings from IR spectra, later extending this to NMR data analysis [27].  
118 However, most existing research focuses on molecule elucidation using single-type spectrum data,  
119 sufficient for simple molecules but inadequate for complex ones since each spectrum provides only

120 partial structural information. Our study aims to leverage the reasoning and planning capabilities of  
121 multimodal large language models (MLLMs) to integrate diverse spectral data, addressing challenges  
122 in complex real-world chemistry tasks. We focus on solving the entire puzzle using multiple clues  
123 rather than deciphering one word from a single clue.

124 **Multimodal Benchmarks for LLMs.** With the advancements in developing multimodal LLMs  
125 [32, 33, 34, 35, 36], a number of multimodal benchmarks have been curated. These benchmarks are  
126 crucial for evaluating and refining the capabilities of MLLMs to process and integrate diverse data  
127 types, such as text, images, and audio, for a cohesive understanding. Notably, a benchmark proposed  
128 by Yue et al. [37] assesses the reasoning abilities of MLLMs in various college-level subjects.  
129 Similarly, MathVista [38] explores MLLMs’ multimodal reasoning capabilities in mathematics,  
130 while Yin et al. [39] introduced LAMM, a dataset focusing on multimodal instruction tuning  
131 and the LabSafetyBench [36] assessed the reliability and safety awareness of LLMs in laboratory  
132 environments. Our research shifts the focus to the chemistry domain [6, 40]. To our knowledge, this  
133 study is the first to adopt a realistic chemistry task for MLLM processing and to conduct a thorough  
134 evaluation of these models’ proficiency in chemistry-related reasoning and image analysis. This  
135 specialized focus will enhance our understanding of MLLMs’ capabilities within a specific scientific  
136 domain.

### 137 3 The MolPuzzle Dataset

138 Existing benchmarks of chemical tasks primarily focused on predictive or generative tasks involving  
139 collections of independent samples that were relatively straightforward to construct. In contrast,  
140 our dataset, MolPuzzle, aims to characterize an intertwined assessment of chemistry reasoning and  
141 visual understanding, testing the application of AI-assisted technology towards broader scientific  
142 discovery. Our data collection process is rigorously designed and implemented by a team uniquely  
143 qualified for this task, consisting of esteemed researchers in chemistry and experienced AI specialists  
144 who have previously tackled complex chemistry problems. This collaboration ensures that the  
145 MolPuzzle dataset not only accurately reflects real-world chemical phenomena and challenges but is  
146 also structured in a way that optimally facilitates access and usability for LLMs.

147 The basic principles guiding our data curation for the MolPuzzle dataset are: 1) ensuring compre-  
148 hensive coverage by including a wide range of tasks that synthesize visual context with chemical  
149 knowledge, facilitating thorough evaluations; 2) varying levels of difficulty to challenge LLMs  
150 and highlight their potential limitations; 3) ensuring robust assessment outcomes, i.e., the results  
151 are definitive and reliable; and 4) incorporating human expert analysis to identify strengths and  
152 weaknesses in model performance, significantly enhancing our understanding of LLMs capabilities.

153 In this section, we outlined the construction process for the MolPuzzle dataset. We detailed the  
154 creation of puzzle tasks in three stages (3.1), as well as the QA pairs involved in these tasks (3.2).  
155 Examples are presented in Fig. 2.

#### 156 3.1 Task Construction

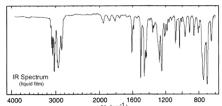
157 Just like a word puzzle where each clue progressively reveals the final answer, the solution to a  
158 molecule puzzle is a SMILES string that captures the interconnected substructures of a molecule. We  
159 design our molecule puzzles so that solving one requires the accurate identification and integration of  
160 each substructural clue, gradually unveiling the complete SMILES representation of the molecule.  
161 This approach is inspired by the analytical strategies employed by chemists in the real world, who  
162 interpret spectral data and chemical properties to deduce the structures of unknown molecules. Our  
163 puzzle-building process mirrors this scientific exploration, arranging clues in a sequence from simple  
164 to complex, where each clue builds upon the insights gained from the previous one, requiring precision  
165 and careful thought at every stage. We next provide more details on our clue design methodology.

166 **The Initial Stage (Molecule Understanding).** In designing a molecule puzzle, the first stage involves  
167 determining how many building blocks, or substructures, are available. This foundational step is  
168 crucial as it sets the stage for constructing the molecule’s complete structure, akin to identifying the  
169 key pieces in a complex jigsaw puzzle. Starting with the initial hint: A molecular formula, derived  
170 from a mass spectrum, indicates the exact types and numbers of atoms in a molecule (e.g., C<sub>15</sub>H<sub>22</sub>O<sub>2</sub>,

171 representing carbon, hydrogen, and oxygen), chemists can begin to deduce possible structures from  
172 the degree of saturation which is calculated based on the number of rings and multiple bonds  
173 present in the molecule, the potential for forming aromatic rings, or the presence of functional  
174 groups. The initial information provides a preliminary range of building blocks, which can later be  
175 selected and assembled to solve the molecular puzzle. To benchmark the capability of LLMs in this  
176 stage, we developed 26 unique templates (see Appendix A.2 for details), targeting key analytical tasks  
177 such as saturation identification, aromatic ring identification, functional group identification, and  
178 saturation degree calculation. This initiative produced 5,859 QA-format pairs, effectively evaluating  
179 the models' capacity to understand and process molecular data. Details of these samples are reported  
180 in Appendix A.3.

181 **The Second Stage (Spectrum Interpretation).** With the initial building blocks of the molecule  
182 identified from the molecular formula, the next critical step involves refining these components  
183 through detailed spectral analysis. Spectrum images such as IR, MASS,  $^1\text{H-NMR}$ , and  $^{13}\text{C-NMR}$   
184 serve as new hints, each adding layers of information akin to clues in a complex puzzle. These  
185 spectral images are pivotal in confirming or revising the initial hypotheses about the molecule's  
186 structure. For example, IR spectroscopy can verify the presence of specific functional groups, MASS  
187 spectrometry can provide the molecular MASS, molecule mass, and fragmentation patterns, and  
188 NMR techniques detail the arrangement of hydrogen and carbon within the molecule. By integrating  
189 these new hints, researchers can construct a more robust and experimentally accurate model of the  
190 molecule. This process not only theoretically validates each building block but also ensures they align  
191 perfectly with empirical data, leading to a comprehensive understanding of the molecular structure.  
192 Given the importance of spectral images in this analysis, we have developed specialized question  
193 templates to evaluate the proficiency of LLMs in interpreting these images. For instance, we created  
194 17 templates for IR and 12 for each of H-NMR, and C-NMR. Each template, such as 'Analyze the  
195 IR spectrum' includes specific queries designed to extract detailed insights, such as 'What does  
196 the absorption in 3200-3600 suggest?' This structure enables us to format the questions for Visual  
197 Question Answering (VQA), facilitating a systematic approach to query handling. Our method has  
198 successfully generated a significant repository of VQA format examples, comprising 3,689 for IR  
199 and 2,604 for each of MASS, H-NMR, and C-NMR. A detailed analysis of these tasks is available in  
200 Appendix A.4.

201 **The Final Stage (Molecule Construction).** After completing the first two stages, we can assert that  
202 we have gathered the necessary building blocks to assemble the molecule. The assembly process will  
203 be guided by insights derived from NMR data. Specifically,  $^1\text{H-NMR}$  provides information about  
204 the hydrogen environment in the molecule, such as the number of hydrogen atoms, their types (e.g.  
205 aromatic), and their connectivity. Meanwhile, C-NMR provides detailed insights into the carbon  
206 framework, indicating whether carbon atoms are part of an aromatic ring or not. The assembly of the  
207 final molecular structure is an iterative process, during which functional groups are uncovered based  
208 on the specific hydrogen and carbon environments. The approach to assembling the final molecular  
209 structure is iterative. Starting with initial building blocks selected from the identified fragment pool,  
210 LLMs are prompted to select one structure from the pool step by step, based on the NMR guidance,  
211 until the maximum number of iterations is reached or the fragment pool is exhausted. This systematic  
212 addition ensures that each step in the assembly process not only fits with the previous structure  
213 but also aligns perfectly with the latest spectral data, driving us closer to the accurate molecular  
214 configuration. We created 27 task templates for each molecule to assess the capability of LLMs in  
215 comprehending NMR spectra. These templates include 5 questions about atom numbers and 22 tasks  
216 centered on functional groups, generating a total of 6,318 question-answer pairs. We sample both  
217 atom-related questions concerning the number of hydrogens and carbons, as well as those targeting  
218 functional groups. To reduce bias and ensure more balanced performance, we balance the distribution  
219 of labels in the answers—whether indicating the presence or absence of a functional group or specific  
220 environment. This ensures a more unbiased evaluation across the sampled tasks.

<p><b>1. Identify molecule substructures based on molecule formula</b></p> <p>Prompt: As an expert organic chemist, your task is to analyze the chemical formula C<sub>6</sub>H<sub>10</sub>O<sub>6</sub> and determine the potential molecular structures and the degree of unsaturation. Utilize your knowledge to systematically explore and identify plausible molecular substructure.</p>	<p><b>2. Refine the substructure pools based on Spectrum images.</b></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><b>3. Select fragments from the pools and assemble molecule iteratively</b></p> <p><b>Initial selection:</b> 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><b>Iteration:</b> 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><b>End:</b> 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>

(a). The Initial Stage

(b). The Second Stage

(c). The Final Stage

Figure 2: Examples of QA pairs in the 3 stages of MolPuzzle

### 221 3.2 QA Sample Derivation

222 The QA samples for Stage 1 and Stage 2 are automatically generated using their respective question  
 223 templates (see Appendix A.2) and RDKit [41]. RDKit is an open-source cheminformatics toolkit  
 224 widely employed for handling chemical informatics data, including molecular structures and finger-  
 225 prints. This toolkit plays a role in ensuring that the responses, based on the SMILES strings from  
 226 each molecule puzzle, are accurate and chemically valid. The distribution of these QA samples across  
 227 different categories is illustrated in Fig. 4. They form a diverse collection of samples for evaluating  
 228 LLMs' ability to understand molecular formulas and spectra.

229 The fragment of each QA pair at Stage 3 is initially generated by LLMs, i.e., responding to the  
 230 prompt 'select one fragment...'. To validate the reliability of these automated generations of QA  
 231 pairs, experts—two Ph.D. candidates from the chemistry department—manually and independently  
 232 verified 50 samples, labeling the generated fragments as 'correct' or 'wrong'. Their verification  
 233 was consistent and demonstrated that 67.4% of examples have correct fragment pools in automated  
 234 generation. To ensure the quality of derived QA pairs in Stage 3, these chemists manually corrected  
 235 the fragments pool for each instance in the benchmark.

236 Fig.3 reports the statistical distribution for the MolPuzzle dataset, which includes 217 puzzle instances  
 237 (the reasoning of 217 different molecules). Since one puzzle can be solved by different paths, different  
 238 numbers of QA samples are derived in three stages. We will next evaluate LLMs' performance in  
 239 solving each puzzle, as well as their capability to solve individual questions.

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

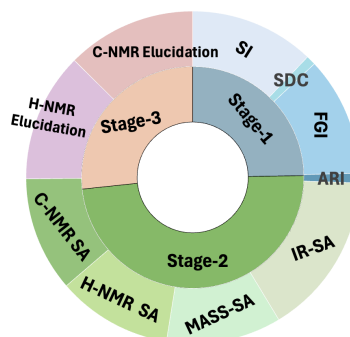


Figure 4: Inner ring: sample distribution in 3 stages. Outer ring: sample distribution across categories in each stage. SI: saturation identification, SDC: saturation degree calculation, FGI: functional group identification, ARI: aromatic ring identification, SA: spectrum analysis.

## 240 4 Solving MolPuzzle by Multimodal Large Language Models

241 The reasoning capabilities of foundation models in the chemistry domain remain underexplored.  
242 Thus, our aim is to perform both qualitative and quantitative evaluations to systematically assess the  
243 reasoning and planning abilities of these models in visual chemistry contexts, using the MolPuzzle  
244 benchmark. We first conducted evaluation of a variety of LLMs for completing the individual tasks  
245 in each stage, including GPT-4o [42], GPT-3.5-turbo [43], Claude-3-opus [44], Gemini-pro [45],  
246 Galactica-30b [46], LLama-3-8B-Instruct [47], Vicuna-13B-v1.5 [48], Mistral-7B-Instruct-v0.3 [49],  
247 and in particular multimodal LLMs such as Gemini-pro-vision [45], LLava-LLama-3-8B [50], Qwen-  
248 VL-Chat [51], and InstructBlip-Vicuna-7B/13B [32]. Due to space limits, we present only selected  
249 results in Table 1 and report the complete list of results in Appendix B. We then assess LLMs’  
250 capability to solve the entire puzzles, specifically focusing on how effectively these models can derive  
251 the final molecular structure from provided hints (the questions in QA samples). The results are  
252 reported in Table 2.

253 All tasks are evaluated in a zero-shot setting to determine the problem-solving capabilities of LLMs  
254 without prior fine-tuning on specific task data. The evaluation process consists of three steps:  
255 response generation, answer extraction, and score calculation. More details of the experimental  
256 settings including prompts and hyperparameters are presented in Appendix B.1.

257 To gain an in-depth understanding of the performance of LLMs in comparison with human experts,  
258 particularly their failed cases, we invited six Ph.D. candidates in chemistry to solve the puzzles in  
259 MolPuzzle, and also assess LLMs’ results. More comprehensive details of this **human baseline**  
260 and evaluation process are presented in Appendix B.2. The reported performance, including human  
261 baselines, is presented as an average with standard deviation over all samples.

### 262 4.1 LLMs’ Performance on Solving Molecule Puzzles

#### 263 4.1.1 Addressing individual QA tasks in three stages

264 In Table 1, we report the performance of selected LLMs on conducting individual QA tasks in the three  
265 stages, including GPT-4o, GPT-3.5-turbo, Claude-3-opus (three top-performing proprietary models),  
266 Llama-3-8B-Instruct (the best performing open-source model), and the reference human baseline  
267 performance. In stage 2, the variant of Llama3 for a multimodal setting, LLava-LLama-3-8B, is used  
268 for handling spectrum image analysis. Since each task involves performing a question-answering  
269 task, we evaluate the performance using F1 and accuracy by comparing the LLMs’ answers with the  
270 ground truth. F1 scores are reported in Table 1, while the accuracy and performance of more LLMs  
271 can be found in Appendix B.

272 The results of Stage 1 (in Table 1 and Appendix Table 3) show that the GPT-4o model excels in these  
273 tasks (achieving near-perfect F1 score in 3 out of 4 tasks). The high scores in SI, AI, and FI suggest  
274 that LLMs are able to succeed in relatively straightforward chemistry analysis tasks, performing  
275 comparably to human experts. However, open-sourced models like LLama3 have limitations in  
276 addressing these tasks, possibly due to their limited reasoning abilities in chemistry text-reasoning  
277 tasks. In addition, GPT-4o’s comparative performance to humans indicates significant advancements  
278 in the use of LLMs for complex scientific tasks, suggesting a promising future for leveraging advanced  
279 LLMs to improve the efficiency of scientific analysis and discovery.

280 For the multimodal tasks of Stage 2, GPT-4o remains the top performer, though it exhibits intermediate  
281 performance in spectrum interpretation. The F1 scores for the four types of spectra average around  
282 0.6, indicating a moderate level of accuracy in this complex aspect of the challenge. This performance  
283 is notably less competitive compared to human baselines, which succeed in approximately 73-77% of  
284 the tasks across the four types of spectrum interpretation. This indicates that spectrum interpretation  
285 is inherently challenging. While GPT-4o has made significant strides in automated spectrum analysis,  
286 there remains considerable room for improvement to bridge the gap between its capabilities and  
287 human expertise. More details are presented in Appendix B.4.  
288 The results for Stage 3 indicate that the most advanced LLM, GPT-4o, significantly underperforms  
289 compared to the human baseline, with nearly a 40% difference. This might be caused by the fact that

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	<b>1.00±0.000</b>	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	<b>0.988±0.015</b>	<b>0.934±0.001</b>	<b>0.856±0.016</b>
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	<b>0.656±0.052</b>	<b>0.609±0.042</b>	<b>0.618±0.026</b>	<b>0.639±0.010</b>
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	<b>0.524±0.021</b>		<b>0.506±0.037</b>	
Llama3	0.341±0.015		0.352±0.017	
Human	0.867±0.230		0.730±0.220	

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	<b>0.014±0.004</b>	<b>11.653±0.013</b>	<b>1.000±0.000</b>	<b>0.431±0.009</b>	<b>0.293±0.013</b>	0.232±0.007
Claude-3-opus	0.013±0.008	12.680±0.086	<b>1.000±0.000</b>	0.383±0.050	0.264±0.040	<b>0.241±0.037</b>
Gemini-pro	0.000±0.000	12.711±0.196	<b>1.000±0.000</b>	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

290 the reasoning ability required for these tasks is complex and multifaceted. When information con-  
 291 verges, such as identifying equivalent hydrogen or ring arrangements, a comprehensive understanding  
 292 of the NMR peaks and their corresponding structures is essential. See more details in Appendix B.5.

#### 293 4.1.2 Addressing entire molecule puzzles

294 For solving the entire molecule puzzles, the evaluation is limited to the three most advanced mul-  
 295 timodal LLMs: GPT-4o [42], Claude-3-opus [44], and Gemini-pro [45], due to the involvement  
 296 of spectrum image analysis in Stage-2. The results of these models are reported in Table 2, along  
 297 with those from the human baseline (see complete evaluation process is reported in Appendix C). To  
 298 comprehensively evaluate the performance, we employ two different types of metrics. The first type  
 299 of metric measures the chemical similarity between the ground-truth molecules and the generated  
 300 molecules, assessed using FTS (Fingerprint Tanimoto Similarity) [52] in terms of MACCS [53],  
 301 RDKit [41], and Morgan [54]. Since the generated molecules are in SMILES string format, we also  
 302 employ natural language processing metrics including the Accuracy of Exact Match [55], and Leven-  
 303 shtein distance [56] (the minimum number of single-character editing required to transform one string  
 304 into another). Finally, to evaluate whether constructed molecules are valid, we use RDKit [41] to  
 305 check the validity of constructed molecules and report the percentage of molecules that are confirmed  
 306 as valid.

307 The results in Table 2 show that the best-performed LLM, GPT-4o, is performing much worse than  
 308 humans, indicating a huge gap between LLMs and humans in solving the molecule puzzles. It is  
 309 worth noting that all the constructed molecules are valid, even though only a small portion of them  
 310 (1.4%) exactly match the ground truth. Considering that the accuracy of the exact match is too strict,



311 we use FTS to analyze more about the chemical closeness of LLMs' answer to the ground truth. A  
312 MACCS FTS of 0.431 suggests that the generated molecules maintain a significant level of structural  
313 similarity. This indicates that even if the answers are not perfect replicas of the ground truth, they  
314 can still be chemically valid and potentially useful as structured hypotheses that could be relived by  
315 human scientists.

## 316 4.2 Success and Failure Analysis

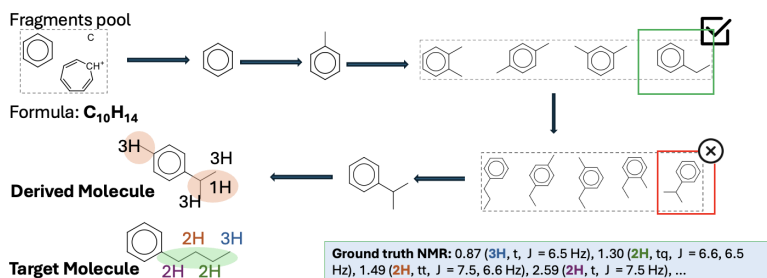


Figure 5: The target molecule contains four distinct non-aromatic hydrogen types, color-coded in the ground truth NMR. However, the model-derived molecule shows hydrogen counts of 3, 3, and 1, differing from the ground truth. The mismatch between the hydrogen types in the green section of the target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

317 The above analysis indicates that the most capable model, GPT-4o, performs **nearly perfectly**  
318 in Stage-1 of molecule understanding. However, its performance **drops** in Stage-2 for spectrum  
319 interpretation, and **worsens further in Stage-3** for molecule construction. We investigate in-depth  
320 how GPT-4o eventually fails on most of the puzzles after progressing through the tasks of these three  
321 stages. With the help of human evaluators, we gathered all the intermediate steps involved in solving a  
322 molecule puzzle and engaged them to scrutinize these steps. Fig. 5 presents case studies that illustrate  
323 the iterative steps involved in Stage-3, showcasing the most common errors made by GPT-4o: **the**  
324 **accumulation of errors in iterative steps, which can lead to catastrophic failures**. Note that  
325 this stage focuses on selecting the correct fragments and assembling them step by step to form the  
326 final molecular structure. We find that GPT-4o can initially succeed in picking the correct fragment  
327 when the structure is comparatively simple. However, as the process progresses, it does not select  
328 structures that satisfy all the requirements indicated by the NMR data. This difficulty arises because  
329 the reasoning requirements expand dramatically as more information and additional constraints need  
330 to be incorporated. More qualitative examples can be found in Appendix C.1.

## 331 5 Findings and Open Directions

332 Our evaluation has revealed specific limitations of state-of-the-art LLMs in automating molecular  
333 structure elucidation. We urge further collaborative efforts from the AI and chemistry communities to  
334 design more effective solutions, especially for the tasks in Stage 2 and Stage 3. Based on our findings,  
335 we next present the open directions for future research and development.

336 **Development of Specialized Multimodal LLMs Spectrum Interpretation in Stage 2.** As indi-  
337 cated in our results, the performance of LLMs notably declines beginning in Stage 2, where they  
338 struggle with the visual interpretation of  $^1H$  and  $^{13}C$  NMR spectra. This difficulty arises because  
339 NMR spectra feature sharp, unlabeled peaks with multiplicities that exhibit very small chemical shift  
340 differences, making them challenging for visual models to interpret. These multiplicities, however,  
341 contain crucial information about the chemical connectivity of molecular fragments. Similarly,  
342 closely spaced IR absorptions provide key insights for identifying functional groups. This presents a  
343 significant opportunity to develop specialized multimodal LLMs that can more effectively interpret  
344 these subtle and complex spectral details.

345 **Development of New Strategies for Leveraging LLMs in Chemical-Related Planning and**  
346 **Reasoning.** The failure analysis from Stage 3 has motivated us to explore more effective strategies  
347 for leveraging LLMs’ capabilities in planning and reasoning for fragment selection and assembly.  
348 Our first immediate approach was to employ the chain-of-thought technique [57], aiming to provide  
349 more structured reasoning and instructions for solving the molecular puzzle. However, despite  
350 implementing this method, the results were unsatisfactory, even performing worse than the zero-shot  
351 setting we initially reported in the paper. We plan to continue exploring this direction with different  
352 implementations and adjustments. A second approach involves utilizing LLMs as agents in a more  
353 dynamic and interactive manner. This strategy incorporates feedback loops, allowing the models  
354 to iteratively refine their responses based on new information or corrections. By doing so, we aim  
355 to mitigate the accumulation of errors in iterative steps and reduce the risk of catastrophic failures  
356 during the problem-solving process. In addition, we are investigating fine-tuning strategies to enhance  
357 the model’s ability to handle domain-specific tasks. This involves fine-tuning LLMs on curated  
358 chemical datasets that include detailed annotations of spectral data and molecular structures. The  
359 goal is to train the model to recognize subtle patterns and dependencies that are often missed in a  
360 general-purpose pre-trained model. By tailoring the model’s training to this domain, we expect to  
361 improve its reasoning and planning capabilities when interpreting complex spectra and assembling  
362 molecular fragments.

## 363 **6 Negative Societal Impacts**

364 Automating molecular elucidation using LLMs has significant benefits but also poses serious risks,  
365 especially regarding the creation of prohibited drugs. 1.)Facilitation of Illicit Drug Synthesis: LLMs  
366 could be used to design new synthetic drugs that evade current regulations, making it easier for illicit  
367 manufacturers to produce harmful substances. 2.)Lowering the Barrier to Entry: The technology  
368 could enable individuals with minimal expertise to create detailed molecular blueprints for prohibited  
369 drugs, increasing the potential for misuse. 3.) Regulatory Challenges: The rapid generation of novel  
370 compounds could overwhelm drug regulators, leading to delays in banning new synthetic drugs  
371 and complicating the control of harmful substances. 4.) Ethical and Legal Issues: Questions about  
372 responsibility and access to such powerful tools arise. Regulating who can use these technologies  
373 and for what purposes becomes crucial to prevent misuse.

## 374 **7 Broader Impact**

375 Our work has broad impacts across multiple dimensions. First, it offers valuable insights and  
376 recommendations for both AI researchers and chemists in academia and industry. These perspectives  
377 enhance the effective utilization of LLMs and guide future advancements in the field. Second,  
378 our approach to benchmarking and improving LLMs through real-world tasks like the MolPuzzle  
379 can also foster greater collaboration between computational scientists and chemists. By aligning  
380 AI technologies with traditional chemical research, these interdisciplinary efforts can accelerate  
381 the discovery of new materials, drugs, and chemical processes, potentially leading to significant  
382 advancements in healthcare and industry.

## 383 **8 Conclusion**

384 In this paper, we introduced MolPuzzle, a new benchmark challenge to advance our capabilities in  
385 molecular structure elucidation. We evaluated state-of-the-art LLMs on this task, revealing their  
386 strengths and limitations in handling complex chemical reasoning. Our analysis highlights significant  
387 performance gaps, particularly in spectrum interpretation and molecule construction. These findings  
388 not only suggest ways to improve LLM performance but also set the stage for transforming approaches  
389 to chemical research. MolPuzzle serves as a critical step toward harnessing the potential of LLMs  
390 in chemistry, fostering innovation and collaboration within the AI and chemistry communities to  
391 enhance scientific inquiry and application.

## 392 Acknowledgments and Disclosure of Funding

393 This work was supported by the National Science Foundation (CHE-2202693) through the NSF  
394 Center for Computer-Assisted Synthesis (C-CAS).

## 395 References

- 396 [1] Venkat Venkatasubramanian. The promise of artificial intelligence in chemical engineering: Is  
397 it here, finally? *AIChE Journal*, 65(2):466–478, 2019.
- 398 [2] Zachary J Baum, Xiang Yu, Philippe Y Ayala, Yanan Zhao, Steven P Watkins, and Qiongqiong  
399 Zhou. Artificial intelligence in chemistry: current trends and future directions. *Journal of*  
400 *Chemical Information and Modeling*, 61(7):3197–3212, 2021.
- 401 [3] Alexandre Blanco-Gonzalez, Alfonso Cabezon, Alejandro Seco-Gonzalez, Daniel Conde-  
402 Torres, Paula Antelo-Riveiro, Angel Pineiro, and Rebeca Garcia-Fandino. The role of ai in drug  
403 discovery: challenges, opportunities, and strategies. *Pharmaceuticals*, 16(6):891, 2023.
- 404 [4] Xuan-Quy Dao, Ngoc-Bich Le, Bac-Bien Ngo, and Xuan-Dung Phan. Llms’ capabilities at the  
405 high school level in chemistry: Cases of chatgpt and microsoft bing ai chat. 2023.
- 406 [5] Suryanarayanan Balaji, Rishikesh Magar, Yayati Jadhav, et al. GPT-MolBERTa: GPT Molecular  
407 Features Language Model for molecular property prediction. *arXiv preprint arXiv:2310.03030*,  
408 2023.
- 409 [6] Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xi-  
410 angliang Zhang, et al. What can large language models do in chemistry? a comprehensive  
411 benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–  
412 59688, 2023.
- 413 [7] Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe  
414 Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint*  
415 *arXiv:2304.05376*, 2023.
- 416 [8] Alan M Rosan. Organic structures from spectra, (field, ld; sternhell, s.; kalman, jr), 2002.
- 417 [9] Roman M Balabin, Ekaterina I Lomakina, and Ravilya Z Safieva. Neural network (ANN)  
418 approach to biodiesel analysis: analysis of biodiesel density, kinematic viscosity, methanol and  
419 water contents using near infrared (NIR) spectroscopy. *Fuel*, 90(5):2007–2015, 2011.
- 420 [10] Liu Cao, Mustafa Guler, Azat Tagirdzhanov, Yi-Yuan Lee, Alexey Gurevich, and Hosein  
421 Mohimani. Moldiscovery: Learning mass spectrometry fragmentation of small molecules.  
422 *Nature communications*, 12(1):3718, 2021.
- 423 [11] Xi Xue, Hanyu Sun, Minjian Yang, Xue Liu, Hai-Yu Hu, Yafeng Deng, and Xiaojian Wang.  
424 Advances in the application of artificial intelligence-based spectral data interpretation: A  
425 perspective. *Analytical Chemistry*, 95(37):13733–13745, 2023.
- 426 [12] Marvin Alberts, Teodoro Laino, and Alain C Vaucher. Leveraging infrared spectroscopy for  
427 automated structure elucidation. 2023.
- 428 [13] Jorge Navaza and Pedro Saludjian. [33] amore: An automated molecular replacement program  
429 package. In *Methods in enzymology*, volume 276, pages 581–594. Elsevier, 1997.
- 430 [14] Paul D Adams, Pavel V Afonine, Gábor Bunkóczi, Vincent B Chen, Nathaniel Echols, Jeffrey J  
431 Headd, Li-Wei Hung, Swati Jain, Gary J Kapral, Ralf W Grosse Kunstleve, et al. The phenix  
432 software for automated determination of macromolecular structures. *Methods*, 55(1):94–106,  
433 2011.

- 434 [15] Peter H Zwart, Pavel V Afonine, Ralf W Grosse-Kunstleve, Li-Wei Hung, Thomas R Ioerger,  
435 Airlie J McCoy, Erik McKee, Nigel W Moriarty, Randy J Read, James C Sacchettini, et al.  
436 *Automated structure solution with the PHENIX suite*. Springer, 2008.
- 437 [16] Patrick C Fricker, Marcus Gastreich, and Matthias Rarey. Automated drawing of structural  
438 molecular formulas under constraints. *Journal of chemical information and computer sciences*,  
439 44(3):1065–1078, 2004.
- 440 [17] Gábor Bunkóczi, Nathaniel Echols, Airlie J McCoy, Robert D Oeffner, Paul D Adams, and  
441 Randy J Read. Phaser. mrange: automated molecular replacement. *Acta Crystallographica*  
442 *Section D: Biological Crystallography*, 69(11):2276–2286, 2013.
- 443 [18] Lily A Chylek, Leonard A Harris, Chang-Shung Tung, James R Faeder, Carlos F Lopez, and  
444 William S Hlavacek. Rule-based modeling: a computational approach for studying biomolecular  
445 site dynamics in cell signaling systems. *Wiley Interdisciplinary Reviews: Systems Biology and*  
446 *Medicine*, 6(1):13–36, 2014.
- 447 [19] Andre Lavanchy, Tomas Varkony, Dennis H Smith, Neil AB Gray, William C White, Ray-  
448 mond E Carhart, Bruce G Buchanan, and Carl Djerassi. Rule-based mass spectrum prediction  
449 and ranking: Applications to structure elucidation of novel marine sterols. *Organic Mass*  
450 *Spectrometry*, 15(7):355–366, 1980.
- 451 [20] Christoph Steinbeck. Seneca: A platform-independent, distributed, and parallel system for  
452 computer-assisted structure elucidation in organic chemistry. *Journal of chemical information*  
453 *and computer sciences*, 41(6):1500–1507, 2001.
- 454 [21] Mikhail Elyashberg. Identification and structure elucidation by nmr spectroscopy. *TrAC Trends*  
455 *in Analytical Chemistry*, 69:88–97, 2015.
- 456 [22] Stefan Kuhn, Björn Egert, Steffen Neumann, and Christoph Steinbeck. Building blocks for  
457 automated elucidation of metabolites: Machine learning methods for nmr prediction. *BMC*  
458 *bioinformatics*, 9:1–19, 2008.
- 459 [23] Mikhail Elyashberg and Dimitris Argyropoulos. Computer assisted structure elucidation (case):  
460 current and future perspectives. *Magnetic Resonance in Chemistry*, 59(7):669–690, 2021.
- 461 [24] Michael A Skinnider, Fei Wang, Daniel Pasin, Russell Greiner, Leonard J Foster, Petur W Dals-  
462 gaard, and David S Wishart. A deep generative model enables automated structure elucidation  
463 of novel psychoactive substances. *Nature Machine Intelligence*, 3(11):973–984, 2021.
- 464 [25] Ivan M Novitskiy and Andrei G Kutateladze. Du8ml: Machine learning-augmented density  
465 functional theory nuclear magnetic resonance computations for high-throughput in silico solu-  
466 tion structure validation and revision of complex alkaloids. *The Journal of Organic Chemistry*,  
467 87(7):4818–4828, 2022.
- 468 [26] Maribel O Marcarino, Maria M Zanardi, Soledad Cicetti, and Ariel M Sarotti. Nmr calcula-  
469 tions with quantum methods: development of new tools for structural elucidation and beyond.  
470 *Accounts of Chemical Research*, 53(9):1922–1932, 2020.
- 471 [27] Marvin Alberts, Federico Zipoli, and Alain C Vaucher. Learning the Language of NMR:  
472 Structure Elucidation from NMR spectra using Transformer Models. 2023.
- 473 [28] Fei Wang, Jaanus Liigand, Siyang Tian, David Arndt, Russell Greiner, and David S Wishart.  
474 Cfm-id 4.0: more accurate esi-ms/ms spectral prediction and compound identification. *Analyti-  
475 cal chemistry*, 93(34):11692–11700, 2021.
- 476 [29] Fei Wang, Dana Allen, Siyang Tian, Eponine Oler, Vasuk Gautam, Russell Greiner, Thomas O  
477 Metz, and David S Wishart. Cfm-id 4.0—a web server for accurate ms-based metabolite  
478 identification. *Nucleic acids research*, 50(W1):W165–W174, 2022.

- 479 [30] Peter Lasch, Max Diem, Wolfgang Hänsch, and Dieter Naumann. Artificial neural networks  
480 as supervised techniques for ft-ir microspectroscopic imaging. *Journal of Chemometrics: A*  
481 *Journal of the Chemometrics Society*, 20(5):209–220, 2006.
- 482 [31] Jens Behrmann, Christian Etmann, Tobias Boskamp, Rita Casadonte, Jörg Kriegsmann, and Pe-  
483 ter Maaß. Deep learning for tumor classification in imaging mass spectrometry. *Bioinformatics*,  
484 34(7):1215–1223, 2018.
- 485 [32] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng  
486 Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose  
487 vision-language models with instruction tuning. *Advances in Neural Information Processing*  
488 *Systems*, 36, 2024.
- 489 [33] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal,  
490 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel  
491 Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M.  
492 Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz  
493 Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec  
494 Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- 495 [34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances*  
496 *in neural information processing systems*, 36, 2024.
- 497 [35] Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao,  
498 Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. Scemqa: A scientific college entrance level  
499 multimodal question answering benchmark. *arXiv preprint arXiv:2402.05138*, 2024.
- 500 [36] Yujun Zhou, Jingdong Yang, Kehan Guo, Pin-Yu Chen, Tian Gao, Werner Geyer, Nuno Moniz,  
501 Nitesh V Chawla, and Xiangliang Zhang. Labsafety bench: Benchmarking llms on safety issues  
502 in scientific labs. *arXiv preprint arXiv:2410.14182*, 2024.
- 503 [37] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens,  
504 Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal  
505 understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.
- 506 [38] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao  
507 Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical  
508 reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.
- 509 [39] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang,  
510 Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-  
511 tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*,  
512 36, 2024.
- 513 [40] Zhichun Guo, Kehan Guo, Bozhao Nan, Yijun Tian, Roshni G Iyer, Yihong Ma, Olaf Wiest,  
514 Xiangliang Zhang, Wei Wang, Chuxu Zhang, et al. Graph-based molecular representation  
515 learning. *arXiv preprint arXiv:2207.04869*, 2022.
- 516 [41] G. A. Landrum. Rdkit: Open-source cheminformatics software. <http://www.rdkit.org>, 2020.
- 517 [42] OpenAI. Hello GPT-4o. <https://openai.com/index/hello-gpt-4o/>, 2023.
- 518 [43] OpenAI. GPT-3.5-Turbo: Enhancements and Applications. [https://openai.com/models/](https://openai.com/models/gpt-3.5-turbo)  
519 [gpt-3.5-turbo](https://openai.com/models/gpt-3.5-turbo), 2023.
- 520 [44] Anthropic. Introducing the Claude-3 Family. [https://www.anthropic.com/news/](https://www.anthropic.com/news/claude-3-family)  
521 [claude-3-family](https://www.anthropic.com/news/claude-3-family), 2023.
- 522 [45] Google. Introducing gemini: our largest and most capable ai model, 2023.

- 523 [46] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis  
524 Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language  
525 model for science. *arXiv preprint arXiv:2211.09085*, 2022.
- 526 [47] Meta. Introducing Meta Llama 3. <https://llama.meta.com/llama3/>, 2023.
- 527 [48] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng,  
528 Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot  
529 impressing gpt-4 with 90%\* chatgpt quality, march 2023. URL [https://lmsys.org/blog/2023-03-](https://lmsys.org/blog/2023-03-30-vicuna)  
530 [30-vicuna](https://lmsys.org/blog/2023-03-30-vicuna), 3(5), 2023.
- 531 [49] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh  
532 Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile  
533 Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- 534 [50] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual  
535 instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- 536 [51] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge,  
537 Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- 538 [52] Taffee T Tanimoto. Elementary mathematical theory of classification and prediction. *Journal of*  
539 *Biomedical Science and Engineering*, 1958.
- 540 [53] David Ratcliff, John W.; Metzener. Pattern matching: The gestalt approach, 1988.
- 541 [54] Debadutta Dash, Rahul Thapa, Juan M Banda, Akshay Swaminathan, Morgan Cheatham,  
542 Mehr Kashyap, Nikesh Kotecha, Jonathan H Chen, Saurabh Gombar, Lance Downing, et al.  
543 Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare  
544 delivery. *arXiv preprint arXiv:2304.13714*, 2023.
- 545 [55] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, and Heng Ji. Translation between molecules  
546 and natural language. *arXiv preprint arXiv:2204.11817*, 2022.
- 547 [56] Frederic P Miller, Agnes F Vandome, and John McBrewster. Levenshtein distance: Information  
548 theory, computer science, string (computer science), string metric, damerau? Levenshtein  
549 distance, spell checker, hamming distance, 2009.
- 550 [57] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le,  
551 Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models.  
552 *Advances in neural information processing systems*, 35:24824–24837, 2022.
- 553 [58] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li,  
554 Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2023 update. *Nucleic acids*  
555 *research*, 51(D1):D1373–D1380, 2023.
- 556 [59] Kevin Maik Jablonka, Luc Patiny, and Berend Smit. Making molecules vibrate: Interactive web  
557 environment for the teaching of infrared spectroscopy, 2022.
- 558 [60] nmrdb.org. Predict and simulate nmr spectra. <https://www.nmrdb.org/>, 2024. Accessed:  
559 2024-10-26.

## 560 **A MolPuzzle Benchmark Details**

561 This section complements Section 3 with a fine-grained summary of the dataset collection, results  
562 validation, and evaluation procedure, along with a fuller characterization of the task instances and the  
563 corresponding prompts.

### 564 **A.1 Data Collection**

565 The initial molecules were selected by referencing the textbook *Organic Structures from Spectra, 4th*  
566 *Edition*, available as an online PDF on ResearchGate. We chose 234 molecules based on spectrum  
567 tasks involving IR, MS,  $^1\text{H-NMR}$ , and  $^{13}\text{C-NMR}$  to reflect a difficulty level suitable for graduate  
568 students[8].

569 To address copyright concerns, we excluded molecules with publicly available mass spectrometry  
570 (MS) spectra in open-source databases from our study. The remaining spectra were sourced from  
571 public resources, notably the PubChem database[58]. For additional spectra that were not available,  
572 we used simulation methods[59][29] and provided a Jupyter notebook to generate these data, ensuring  
573 high-quality spectra for analysis. Our final dataset comprised 200 molecules.

574 Given the challenges associated with NMR spectrum images, some spectra were obtained from  
575 simulated data in text format for  $^1\text{H-NMR}$  and  $^{13}\text{C-NMR}$ . This approach ensured clarity and accuracy  
576 in the evaluation of molecular structures.

577 To assess the multiple-stage abilities of LLMs, we designed a unique question-and-answer evaluation.  
578 This framework tested the LLMs' capabilities in interpreting and integrating data from different types  
579 of spectra, simulating real-world challenges. Details of this evaluation framework are provided in the  
580 next section.

### 581 **A.2 Template design**

582 Each template was crafted to target specific skills within molecular understanding. For instance,  
583 saturation identification challenges the models' ability to discern the degree of saturation in a molecule,  
584 which is crucial for understanding its chemical reactivity and stability. Aromatic ring identification  
585 tests the models' ability to recognize benzene-like structures, which are fundamental in organic  
586 chemistry due to their common occurrence and unique properties. Saturation degree calculation  
587 pushes the models to apply quantitative analysis, requiring not just recognition but also computation  
588 based on molecular structures.

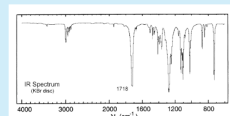
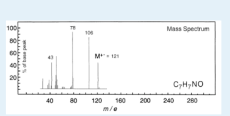
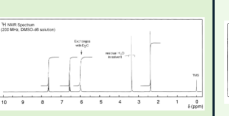
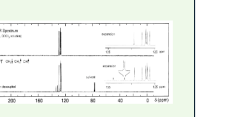
589 By diving deeper into the rationale behind each template and the kind of chemical knowledge they  
590 are designed to test, we can better appreciate how these tasks simulate real-world applications in  
591 chemistry. This approach not only tests the models' basic recognition abilities but also their capacity  
592 to perform complex reasoning and apply theoretical knowledge practically. The template examples  
593 are in A.3.

594 **A.3 Stage1 QA Samples**

Table 3: QA samples for the molecule understanding task

Task	Prompt
Saturation Identification	Question: Could the molecule with the formula C <sub>8</sub> H <sub>10</sub> O potentially be Saturated? Answer: No Model response: No.
Aromatic Ring Identification	Question: Could the molecule with the formula C <sub>8</sub> H <sub>10</sub> O have aromatic rings? Answer: Yes Model response: Yes.
Functional Group Identification	Question: Could the molecule with the formula C <sub>6</sub> H <sub>14</sub> O <sub>2</sub> potentially contain a Amine group, given the Degree of Unsaturation is 0.0? Answer: No Model response: No, the molecule doesn't contain Amine group
Saturation Degree Calculation	Question: Calculate the Degree of Unsaturation of the molecule with the formula C <sub>8</sub> H <sub>10</sub> O? Answer: 4.0 Model response: 2

595 **A.4 Stage2 QA Samples**

IR Interpretation	MASS Interpretation	H-NMR Interpretation	C-NMR Interpretation
 <p><b>Question:</b> Does the IR spectrum contains broad absorption peak of N-H stretching around 3200-3600 cm<sup>-1</sup>?</p>	 <p><b>Question:</b> Examine the MASS spectrum to determine if the molecule could potentially contain specific fragments: Ether.</p>	 <p><b>Question:</b> Examine the H-NMR spectrum to determine if the molecule could potentially contain specific functional groups: Phenol?</p>	 <p><b>Question:</b> Examine the C-NMR spectrum to determine if the molecule could potentially contain specific fragments: Ester.</p>
Answer: No Model response: No	Answer: No Model response: Yes	Answer: No Model response: No	Answer: No Model response: Yes

596 **A.5 Stage3 QA Samples**

Table 4: QA samples for the molecule construction task

Task	Prompt
H-NMR Elucidation	Question: Calculate the number of different types of hydrogen atoms present in the molecule, based on the given H-NMR: 4.51-4.61 (4H, 4.56 (s), 4.56 (s)), 7.06-7.32 (10H, 7.13 (dddd, J = 7.9, 7.7, 1.8, 0.6 Hz), 7.13 (dddd, J = 7.9, 7.7, 1.8, 0.6 Hz), 7.25 (dddd, J = 7.9, 1.5, 1.3, 0.6 Hz), 7.25 (dddd, J = 7.9, 1.5, 1.3, 0.6 Hz), 7.26 (tt, J = 7.7, 1.5 Hz), 7.26 (tt, J = 7.7, 1.5 Hz)) Answer: 4 Model response: 3.
C-NMR Elucidation	Question: Analyze the given C-NMR data and determine the number of different types of carbon atoms present in the molecule, based on given C-NMR: 39.3 (1C, s), 63.4 (1C, s), 127.8 (1C, s), 128.4 (2C, s), 128.8 (2C, s), 134.2 (1C, s). Only output the number. Answer: 6 Model response: 8

597 **B Evaluation Experiments**598 **B.1 Experimental Setting**

599 During our testing phase, we selected 100 questions and employed two distinct prompting strategies  
600 with the large language model (LLM). Initially, the LLM was tasked with directly answering the



601 questions. In a subsequent approach, the same queries were presented, but the model was prompted to  
602 execute a chain-of-thought reasoning process before responding. Each question in our dataset begins  
603 with a comprehensive description of the chemical context, along with specified answer formats and  
604 detailed guiding rules. To ensure a balanced representation of each task category, for tasks in Stage 1,  
605 the distribution ratio for Saturation Identification (SI), Functional Group Identification (FI), Aromatic  
606 Ring Identification (AI), and Saturation Degree Calculation (SC) is set at 2:3:3:2. In Stage 2, we  
607 have randomly selected 100 questions from each category of the spectrum. For Stage 3, we randomly  
608 selected 100 questions focused on H-NMR and C-NMR analyses.

609 We carried out this evaluation over three rounds, analyzing responses using both accuracy and the  
610 F1 score for tasks involving Saturation Identification (SI), Functional Group Identification (FI), and  
611 Aromatic Ring Identification (AI). For Saturation Degree Calculation (SDC), which yields numerical  
612 results, we assessed accuracy by comparing the count of correct matches to the ground truth data.  
613 The detailed results are reported in Table A.3. To ensure that all results are presented in a way that  
614 facilitates direct comparison, only those using similar evaluation metrics(AI, FI, AI) are included  
615 in the main table. For the SI, AI, and FI tasks, we use the F1 score and Accuracy to evaluate their  
616 performance since they are classification tasks. For the SDC task, the answer is a numerical number,  
617 so we only use the accuracy score to measure the performance of the LLMs. This approach helps to  
618 keep the evaluation coherent and focused on comparable data points.

## 619 **B.2 Human Evaluation**

620 To evaluate the performance of large language models (LLMs) on specialized tasks against expert  
621 humans, we recruited six graduate students from chemistry department to solve the MolPuzzle  
622 benchmark. These students, having recently completed a graduate-level course in Molecular Structural  
623 Elucidation, represented a highly skilled group of human participants.

624 For the experiment, we randomly selected six questions from the MolPuzzle dataset for each stage of  
625 the study. These questions were presented to the students in different formats according to the stage:  
626 In Stages 1 and 2, the questions were simple Yes/No or required short answers. In Stage 3, to align  
627 with the conventional methods chemists use to express chemical structures, students were asked to  
628 upload images of their hand-drawn structures instead of using SMILES strings. These images were  
629 manually compared to the ground truth to calculate scores.

630 We also imposed self-regulated time constraints to mirror the challenging nature of molecular  
631 structural elucidation. Beyond individual stage evaluations, we presented each participant with a  
632 complete molecule puzzle, consisting of a formula and four spectral images. The students were tasked  
633 with solving these puzzles within a 20-minute time frame. Impressively, all participants successfully  
634 submitted their solutions within the allotted period.

635 Our study included a component where human evaluators were involved to assess the performance  
636 of the AI models. To ensure the protection and ethical treatment of all participants, we conducted a  
637 thorough risk assessment. Potential risks identified included privacy concerns and the mental strain  
638 of repetitive tasks. Mitigation strategies, such as ensuring anonymity and providing breaks, were  
639 implemented to protect our evaluators.

640 The study was submitted for review and received approval from our Institutional Review Board (IRB).  
641 The IRB approval number is [insert approval number], which verifies that our protocols met all ethical  
642 guidelines for research involving human subjects. Throughout the project, we adhered strictly to  
643 these protocols to ensure ongoing compliance with ethical standards.

## 644 **B.3 Stage1**

645 Molecule understanding requires comprehensive analysis and interpretation of molecular structures,  
646 with a focus on chemical properties and spectroscopic data. In our study, we created a dataset of  
647 234 molecules and developed eight distinct question templates across four categories: **Saturation**  
648 **Identification(SI), Functional Group Identification(FI), Aromatic Ring Identification(AI), and**

649 **Saturation Degree Calculation(SC)**. These templates assess the ability to identify substructures,  
 650 compute saturation levels, and infer structural presence, incorporating concepts in the chemistry  
 651 reasoning process. Each question also necessitates a deep understanding of molecular bonding,  
 652 stereochemistry, and functional group identification. Responses were generated using the RDKit  
 653 library, ensuring precise and reliable answers grounded in established chemical informatics.

Table 3: The accuracy( $\uparrow$ ), F1 score( $\uparrow$ ) in 4 different molecule understanding categories, the best LLMs are in bold font.

Model	CoT	SI		AI		FI		SC
		F1	Acc	F1	Acc	F1	Acc	Acc
GPT-4o	-	<b>1±0.0</b>	<b>1±0.0</b>	0.943±0.016	0.944±0.015	0.934±0.005	0.966±0.0	0.667±0.003
GPT-4o	✓	<b>1±0.0</b>	<b>1±0.0</b>	<b>0.911±0.031</b>	<b>0.911±0.031</b>	0.689±0.025	0.766±0.027	0.816±0.062
GPT-3.5	-	0.451±0.025	0.825±0.075	0.816±0.017	0.816±0.075	0.826±0.075	0.683±0.016	0.5±0.099
GPT-3.5	✓	0.448±0.026	0.816±0.008	0.798±0.025	0.800±0.027	0.526±0.053	0.622±0.031	0.533±0.131
Claude-3-opus	-	0.361±0.009	0.556±0.023	<b>0.988±0.015</b>	0.988±0.015	<b>0.934±0.001</b>	0.966±0.001	<b>0.856±0.016</b>
Claude-3	✓	0.760±0.189	0.903±0.046	0.878±0.025	0.867±0.001	0.547±0.112	0.843±0.081	0.900±0.025
Gemini-pro	-	0.285±0.020	0.399±0.040	0.775±0.093	0.788±0.083	0.646±0.052	0.748±0.051	0.200±0.004
Gemini-pro	✓	0.391±0.045	0.651±0.108	0.685±0.088	0.688±0.087	0.562±0.018	0.629±0.023	0.283±0.062
LLama3	-	0.367±0.018	0.583±0.047	0.490±0.030	0.533±0.027	0.472±0.133	0.588±0.0	0.0±0.0
LLama3	✓	0.473±0.011	0.899±0.040	0.384±0.026	0.533±0.0	0.570±0.035	0.799±0.047	0.017±0.001
Vicuna-13b	-	0.031±0.022	0.033±0.025	0.500±0.087	0.522±0.083	0.308±0.038	0.311±0.041	0.0±0.0
Vicuna-13b	✓	0.380±0.023	0.616±0.062	0.342±0.006	0.522±0.157	0.516±0.080	0.855±0.016	0.0±0.0
Mistral-7b	-	0.221±0.014	0.283±0.025	0.384±0.005	0.500±0.0	0.319±0.014	0.322±0.157	0.0±0.0
Mistral-7b	✓	0.433±0.007	0.766±0.023	0.342±0.006	0.522±0.016	0.601±0.102	0.877±0.031	0.0±0.0

## 654 B.4 Stage2

655 The Spectrum interpretation tasks mainly measure the capability of LLMs in analyzing images  
 656 related to identifying key substructures indicated by the spectrum plot. In this study, we utilize  
 657 four distinct types of spectral images: nuclear magnetic resonance (NMR), infrared spectroscopy  
 658 (IR), mass spectrometry, and others. Each type of data offers insights into various aspects of the  
 659 molecular structure. We’ve created specific question templates for each spectrum, targeting peak  
 660 and substructure identification factors. These templates are designed manually and emphasize the  
 661 intricate connection between the spikes or troughs in the figures and the structures of the molecules.  
 662 Responses were generated using the RDKit library to ensure correctness.

663 The findings from Stage 2 are presented in Table 4. We exclusively focus on the zero-shot learning  
 664 outcomes, as our observations indicate that implementing chain-of-thought prompting leads to a  
 665 deterioration in model performance. To address this limitation, we offer qualitative insights in C.1.

Table 4: The accuracy( $\uparrow$ ), F1 score( $\uparrow$ ) for IR, MASS spectrum, H-NMR, and C-NMR interpretation tasks. "-" means the results are not interoperable

Model	Stage-2 Tasks							
	IR Interpretation		MASS Interpretation		H-NMR Interpretation		C-NMR Interpretation	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc
GPT-4o	<b>0.656±0.052</b>	<b>0.713±0.06</b>	<b>0.609±0.042</b>	<b>0.767±0.042</b>	<b>0.618±0.026</b>	<b>0.864±0.007</b>	<b>0.639±0.107</b>	<b>0.892±0.049</b>
Claude-3-opus	0.440±0.006	0.476±0.055	0.398±0.032	0.466±0.019	0.572±0.190	0.842±0.017	0.554±0.075	0.716±0.042
Gemini-3-pro-vision	0.194±0.002	0.119±0.016	0.116±0.036	0.124±0.038	0.545±0.048	0.851±0.062	0.492±0.016	0.619±0.044
LLava1.5-8b	0.256±0.026	0.414±0.044	0.101±0.021	0.104±0.26	0.118±0.008	0.186±0.011	0.254±0.015	0.472±0.023
Qwen-VL-Chat	0.243±0.027	0.392±0.043	0.125±0.006	0.116±0.021	0.255±0.007	0.611±0.031	-	-
InstructBLIP-7b	0.239±0.020	0.263±0.014	0.101±0.021	0.104±0.26	-	-	0.044±0.006	0.064±0.023
InstructBLIP-13b	0.239±0.020	0.263±0.014	0.101±0.021	0.104±0.26	-	-	0.047±0.014	0.067±0.025

## 666 B.5 Stage-3

667 Constructing a molecule involves a detailed analysis of NMR data, which is critical for understanding  
 668 its structure. H-NMR data are essential as they provide information about the hydrogen environments  
 669 within the molecule, including the number and types of hydrogen atoms (such as aliphatic or

670 aromatic), as well as their connectivity. Conversely, C-NMR data offer in-depth insights into the  
671 carbon framework, illustrating the distribution and linkage of carbon atoms within the molecule.  
672 In our study, to evaluate the ability of large language models (LLMs) to interpret NMR data, we  
673 generated 1,171 question-and-answer (QA) pairs. These pairs focus on key NMR interpretation tasks,  
674 such as counting hydrogen atom types and identifying substructures, which are critical for accurate  
675 analysis.

676 Despite observing moderate accuracy from the LLMs in Stage 2 of our testing, we enhanced the  
677 quality of the QA pairs in Stage 3 by providing the LLMs with verified NMR data, generated by using  
678 nmrdB[60]. This approach ensures that the data used is reliable and helps maintain the integrity of  
679 our results. The findings from Stage 2 are presented in Table. We exclusively focus on the zero-shot  
680 learning outcomes, as our observations indicate that implementing chain-of-thought prompting leads  
681 to a deterioration in model performance. To address this limitation, we offer qualitative insights in

Table 5: The F1 score( $\uparrow$ ) for H-NMR, and C-NMR Structure Elucidation

Method	H-NMR Elucidation	C-NMR Elucidation
GPT-4o	<b>0.524±0.021</b>	<b>0.506±0.037</b>
Claude-3-opus	0.395±0.008	0.313±0.029
Gemini-pro	0.333±0.012	0.308±0.031
Llama3	0.341±0.015	0.352±0.017
Vicuna-13b	0.181±0.013	0.244±0.001
Mistral-7b	0.131±0.032	0.122±0.027

---

**Algorithm 1** Fragment-Based Molecule Assembly Algorithm

---

**Input:** Fragment pool (SMILES strings), NMR description, Original molecular formula, Original unsaturation degree**Output:** Assembled molecule that satisfies molecular formula and NMR data

---

```

1: Initialize:
2:   Set iteration count  $k \leftarrow 0$ 
3:   Set remaining formula  $\leftarrow$  Original molecular formula
4:   Set remaining unsaturation  $\leftarrow$  Original unsaturation degree

5: 1. Initial Fragment Selection:
6: Prompt LLM with fragment pool and NMR description to select an initial fragment
7: Extract and store the selected fragment

8: 2. Chemical Formula and Unsaturation Check:
9: Convert selected fragment to its chemical formula and unsaturation degree
10: Update remaining formula and unsaturation by subtraction

11: while remaining formula has multiple main atoms and  $k < 5$  do
12:   Increment iteration count  $k \leftarrow k + 1$ 

13:   3. Iterative Fragment Assembly:
14:   Prompt LLM to select additional fragments considering remaining formula and unsaturation
15:   Concatenate selected fragments to form a potential molecule

16:   4. Molecule Validation and NMR Matching:
17:   Validate the new molecule using RDKit for connectivity
18:   if multiple valid molecules exist then
19:     LLM ranks molecules based on NMR match
20:     Select the molecule that best matches the NMR data
21:   end if

22:   5. Subsequent Assembly and Adjustment:
23:   After successful connection, update remaining formula and remaining unsaturation
24: end while

25: 6. Termination Conditions:
26: if no valid fragments can be selected or remaining formula is fully satisfied or  $k \geq 5$  then
27:   Terminate the assembly process
28: end if

29: 7. Final Output:
30: Record the final assembled molecule and intermediate stages
31: if final molecule fits original molecular formula and NMR data then
32:   Return valid solution
33: else
34:   Return no valid solution found
35: end if

```

---

## 683 C.1 Qualitative Results

684 In this section, we present several examples using GPT-4’s chain-of-thought (CoT) reasoning to  
685 facilitate a clearer understanding of the results. We have enlisted two Ph.D. candidates from the  
686 chemistry department to evaluate these CoT outcomes. The analysis uses color coding to indicate the

687 accuracy of the generated text: green signifies correct responses, red indicates incorrect ones, and  
688 yellow denotes responses that are partially correct.

### 689 C.1.1 Stage 2 examples

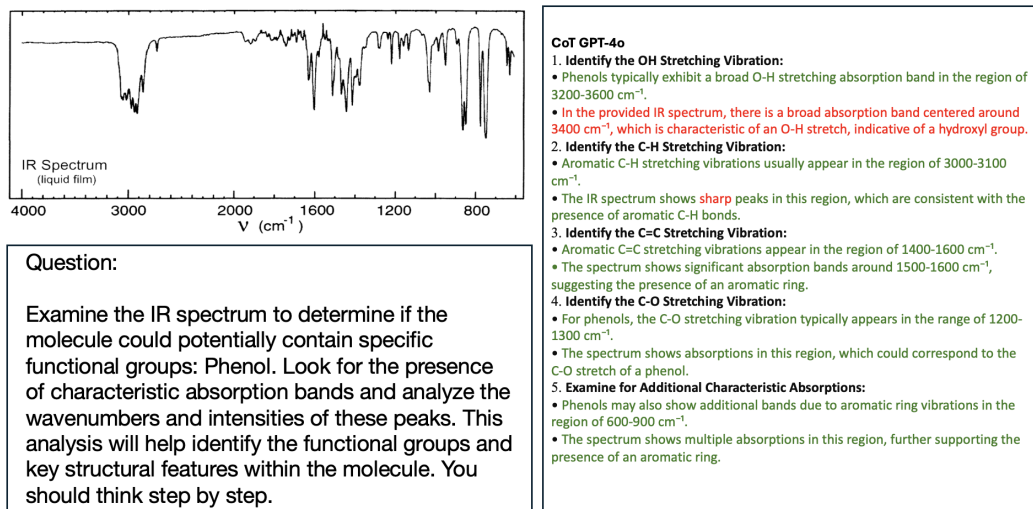


Figure 6: Human annotated IR spectrum interpretation

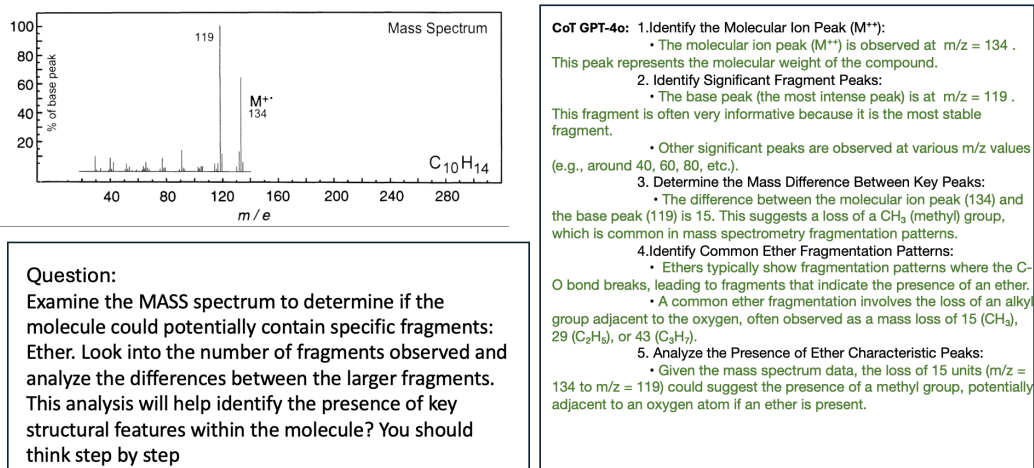


Figure 7: Human annotated MASS spectrum interpretation

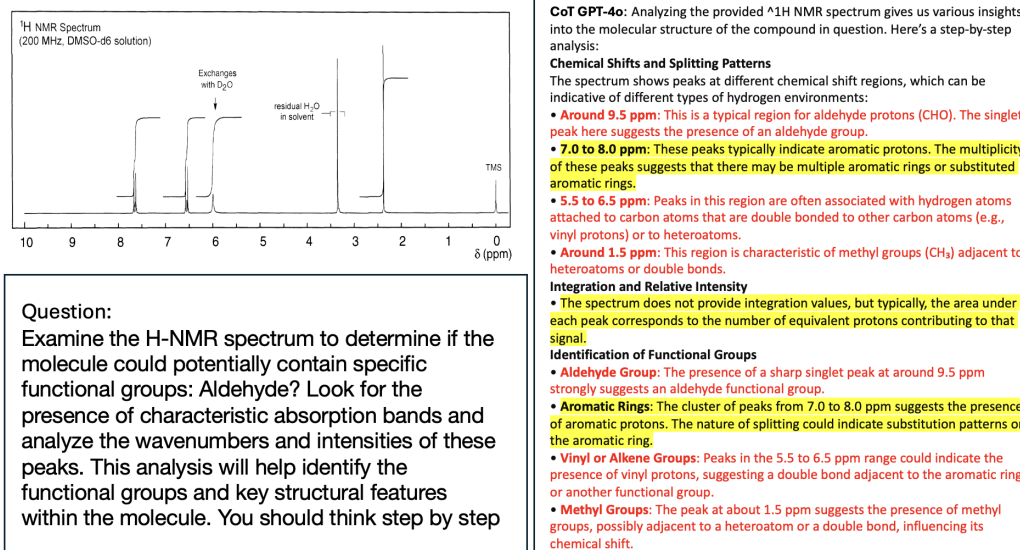


Figure 8: Human annotated H-NMR spectrum interpretation

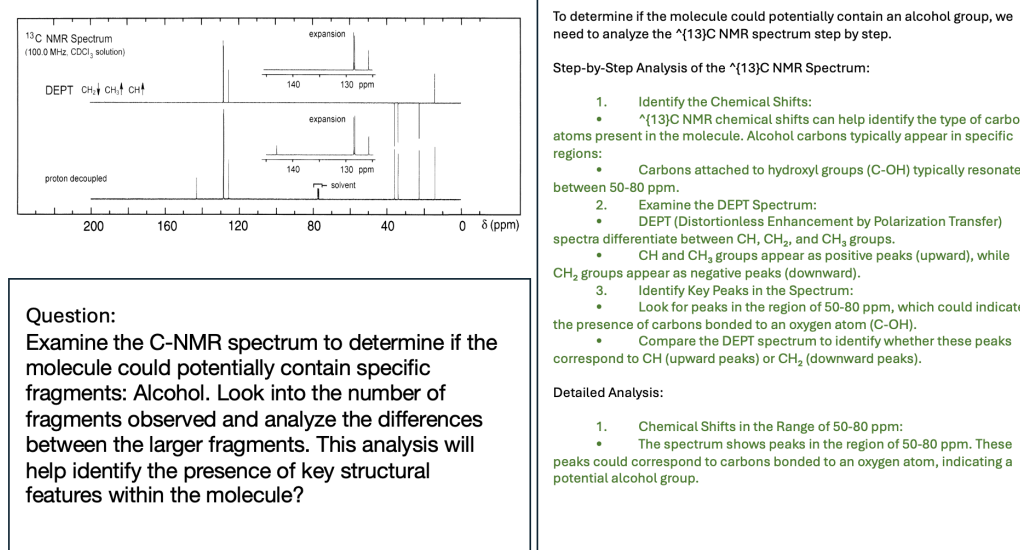


Figure 9: Human annotated C-NMR spectrum interpretation

As an expert organic chemist, your task is to analyze and determine the potential structures that can be derived from a given NMR data. Analyze the given C<sub>13</sub>C-NMR data and determine the number of different types of carbon atoms present in the molecule based on given C<sub>13</sub>C-NMR: [ $\delta$  39.3 (1C, s), 63.4 (1C, s), 127.8 (1C, s), 128.4 (2C, s), 128.8 (2C, s), 134.2 (1C, s)]. You should think step-by-step.

**CoT GPT-4o****1. Identify the Chemical Shifts and Corresponding Carbon Types:**

- $\delta$  39.3 (1C, s): This chemical shift is typically associated with aliphatic carbons, such as those in a methyl or methylene group adjacent to electronegative atoms.
- $\delta$  63.4 (1C, s): This chemical shift is usually indicative of a carbon atom bonded to an oxygen atom, such as in an alcohol or ether group.
- $\delta$  127.8 (1C, s): This shift falls within the region typical for aromatic carbons.
- $\delta$  128.4 (2C, s): These shifts are also within the aromatic region and suggest two equivalent aromatic carbon atoms.
- $\delta$  128.8 (2C, s): Similarly, these shifts suggest two more equivalent aromatic carbon atoms.
- $\delta$  134.2 (1C, s): This shift is also indicative of an aromatic carbon, possibly a substituted one or part of a double bond adjacent to an aromatic ring.

**2. Count the Number of Distinct Carbon Environments:**

- Aliphatic Carbon: 1 carbon at  $\delta$  39.3
- Carbon-Oxygen (Alcohol/Ether) Carbon: 1 carbon at  $\delta$  63.4
- Aromatic Carbons: 1 carbon at  $\delta$  127.8, 2 equivalent carbons at  $\delta$  128.4, 2 equivalent carbons at  $\delta$  128.8, 1 carbon at  $\delta$  134.2

**3. Summarize the Types of Carbon Atoms:**

- Aliphatic carbon (single): 1 type (1 carbon)
- Carbon bonded to oxygen: 1 type (1 carbon)
- Aromatic carbons:
  - 1 type at  $\delta$  127.8 (1 carbon)
  - 1 type at  $\delta$  128.4 (2 equivalent carbons)
  - 1 type at  $\delta$  128.8 (2 equivalent carbons)
  - 1 type at  $\delta$  134.2 (1 carbon)

Figure 10: Human annotated H-NMR Elucidation

As an expert organic chemist, your task is to analyze and determine the potential structures that can be derived from a given NMR data. Generate the number of different types of hydrogen atoms present in the molecule based on given H-NMR: [ $\delta$  2.23 (3H, s), 3.79 (3H, s), 7.04 (2H, ddd, J = 8.3, 1.3, 0.4 Hz), 7.97 (2H, ddd, J = 8.3, 1.6, 0.4 Hz)]. You should think step by step.

**CoT GPT-4o****1. Identify the Chemical Shifts and Corresponding Hydrogen Types:**

- $\delta$  2.23 (3H, s): This singlet indicates 3 equivalent hydrogen atoms. These are likely part of a methyl group (CH<sub>3</sub>) that is not split by neighboring hydrogens.
- $\delta$  3.79 (3H, s): This singlet also indicates 3 equivalent hydrogen atoms, suggesting another methyl group (CH<sub>3</sub>), possibly adjacent to an electronegative atom like oxygen (e.g., methoxy group -OCH<sub>3</sub>).
- $\delta$  7.04 (2H, ddd, J = 8.3, 1.3, 0.4 Hz): This complex splitting pattern (doublet of doublet of doublets) indicates 2 equivalent hydrogen atoms in an aromatic environment, with coupling constants suggesting interactions with neighboring hydrogens.
- $\delta$  7.97 (2H, ddd, J = 8.3, 1.6, 0.4 Hz): Another complex splitting pattern, also indicating 2 equivalent hydrogen atoms in an aromatic environment, with similar coupling constants indicating interactions with neighboring hydrogens.

**2. Determine the Number of Different Types of Hydrogen Environments:**

- Aliphatic Hydrogens:
  - 3 equivalent hydrogens at  $\delta$  2.23 (likely CH<sub>3</sub> group)
  - 3 equivalent hydrogens at  $\delta$  3.79 (likely CH<sub>3</sub> group, possibly part of a methoxy group)
- Aromatic Hydrogens:
  - 2 equivalent hydrogens at  $\delta$  7.04
  - 2 equivalent hydrogens at  $\delta$  7.97
- The molecule has a total of 4 distinct types of hydrogen atoms:
  - Methyl group hydrogens:
    - 3 equivalent hydrogens at  $\delta$  2.23 (CH<sub>3</sub>)
    - 3 equivalent hydrogens at  $\delta$  3.79 (likely part of a methoxy group, -OCH<sub>3</sub>)
  - Aromatic hydrogens:
    - 2 equivalent hydrogens at  $\delta$  7.04
    - 2 equivalent hydrogens at  $\delta$  7.97

Figure 11: Human annotated C-NMR Elucidation

691 Based on the observation in spectrum interpretation, it is safe to say that the GPT-4o model has a  
 692 basic grasp of the concept and the task at hand, but their deficiency mainly resides in obtaining fully  
 693 correct information from the spectrum images and also they lack in-depth understanding. As for the  
 694 problems in Stage 3, although the GPT-4o outputs are mostly correct, they are too general which  
 695 shows the poor understanding of GPT-4o models in solving this intricate task. The stage 3 task can  
 696 be roughly broken down into 3 subtasks: obtaining the correct information from the spectrum image,  
 697 deducing the correct structural information from the spectral information, and finally translating this  
 698 structural information into a correct molecular structure. GPT-4o models seem to perform well in the  
 699 second subtask, and moderately for simple structures in the third subtask but seem to be especially  
 700 struggling with the first subtask in the case of NMR spectra. This indicates the gap in current LLMs  
 701 in fully interpreting data therefore more advanced models and approaches should be developed to  
 702 tackle the problem.

703 **C.1.3 Complex Molecules**

704 In addition to presenting molecules extracted from textbooks, we also demonstrate how the large  
705 language model (LLM) handles complex molecular structures. As illustrated in Figure 12, complex  
706 molecules typically have a larger pool of fragments. This expansion results in a greater number  
707 of valid elucidation paths, complicating the selection process for an appropriate starting point.  
708 Successfully navigating this enlarged pool necessitates an in-depth understanding of each fragment's  
709 properties and the associated, more intricate NMR data. In this context, LLMs may struggle because  
710 they often lack the nuanced chemical intuition and detailed analytical capabilities that human experts  
711 possess. Such limitations can lead to inaccuracies in interpreting complex interactions within NMR  
712 spectra, making LLMs less reliable.

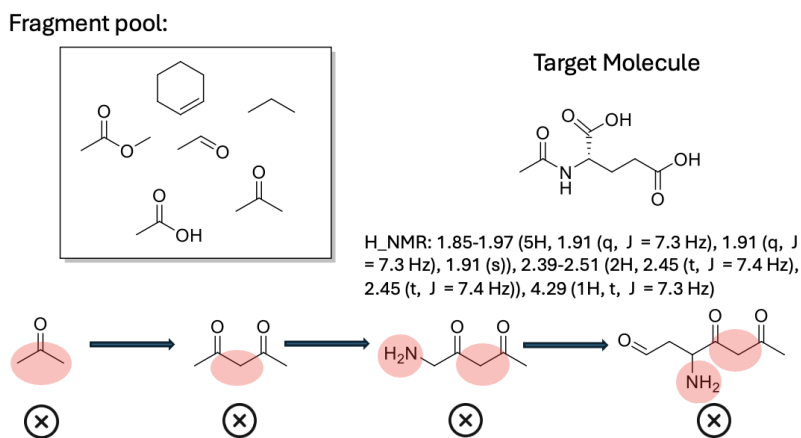


Figure 12: Complex molecule Structure Elucidation



## 713 **D Compute Resources**

714 For the execution of various models in our experiments, distinct compute resources were utilized  
715 based on the model’s accessibility and computational requirements. Specifically, for models like  
716 Claude 3, GPT, and Gemini, we employed API calls to facilitate their operation, leveraging the  
717 existing infrastructure provided by their respective platforms. This approach allowed us to access  
718 these models without the need for local computational resources, thereby streamlining the process.  
719 Conversely, for all other open-sourced models employed in our study, we conducted the experiments  
720 locally using an NVIDIA A100 GPU. This high-performance computing unit was chosen due to its  
721 advanced capabilities in handling extensive computations and large model requirements efficiently.

## 722 Checklist

- 723 1. For all authors...
- 724 (a) Do the main claims made in the abstract and introduction accurately reflect the paper's  
725 contributions and scope? [Yes]
- 726 (b) Did you describe the limitations of your work? [Yes], see Section 7
- 727 (c) Did you discuss any potential negative societal impacts of your work? [Yes], we have  
728 discussed the broader impact in session 7
- 729 (d) Have you read the ethics review guidelines and ensured that your paper conforms to  
730 them? [Yes]
- 731 2. If you are including theoretical results...
- 732 (a) Did you state the full set of assumptions of all theoretical results? [No]
- 733 (b) Did you include complete proofs of all theoretical results? [N/A]
- 734 3. If you ran experiments (e.g. for benchmarks)...
- 735 (a) Did you include the code, data, and instructions needed to reproduce the main experi-  
736 mental results (either in the supplemental material or as a URL)? [Yes], the code is  
737 available at <https://github.com/KehanGuo2/MolPuzzle>.
- 738 (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they  
739 were chosen)? [Yes]
- 740 (c) Did you report error bars (e.g., with respect to the random seed after running experi-  
741 ments multiple times)? [Yes], we report the standard deviation for our result.
- 742 (d) Did you include the total amount of computing and the type of resources used (e.g.,  
743 type of GPUs, internal cluster, or cloud provider)? [Yes], the total GPU usage is  
744 reported in Appendix D.
- 745 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
- 746 (a) If your work uses existing assets, did you cite the creators? [Yes]
- 747 (b) Did you mention the license of the assets? [Yes]
- 748 (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
- 749 (d) Did you discuss whether and how consent was obtained from people whose data you're  
750 using/curating? [Yes]
- 751 (e) Did you discuss whether the data you are using/curating contains personally identifiable  
752 information or offensive content? [Yes]
- 753 5. If you used crowdsourcing or conducted research with human subjects...
- 754 (a) Did you include the full text of instructions given to participants and screenshots, if  
755 applicable? [Yes], see Appendix section B.2
- 756 (b) Did you describe any potential participant risks, with links to Institutional Review  
757 Board (IRB) approvals, if applicable? [Yes], see Appendix section B.2.
- 758 (c) Did you include the estimated hourly wage paid to participants and the total amount  
759 spent on participant compensation? [Yes], see Appendix section B.2.