MolQuest: A Benchmark for Agentic Evaluation of Abductive Reasoning in Chemical Structure Elucidation

Anonymous ACL submission

Abstract

Large language models (LLMs) show growing promise for scientific discovery, yet existing benchmarks rely on static, single-turn QA formats that cannot measure the iterative reasoning and strategic planning required in real research. We introduce **MolQuest**, an agent-based benchmark that reformulates molecular structure elucidation as a multi-turn interactive task grounded in authentic experimental data. Within a simulated laboratory, the model must actively plan experiments, acquire heterogeneous spectral evidence (NMR, MS, IR) on demand, and iteratively refine structural hypotheses under resource constraints—mirroring the abductive reasoning workflow of a practicing chemist. We evaluate 16 frontier LLMs and find that even the best model achieves only 51% accuracy, while most remain below 30%. Furthermore, the dynamic paradigm amplifies performance for some models yet degrades others, revealing strategic planning deficiencies invisible to static evaluations. MolQuest provides a reproducible, extensible framework that exposes the critical gap between pattern recognition and genuine scientific reasoning in current LLMs.

1 Introduction

Large Language Models (LLMs) are demonstrating significant value in scientific discovery, with their applications emerging as a key frontier in "AI for Science" research (Zheng et al., 2025). Recently, the successive releases of new models with strong reasoning capabilities—such as GPT-5.5 (OpenAI, 2026), Gemini 3-Pro (Google DeepMind, 2025a), and Qwen-3-Max (Yang et al., 2025)—have made it increasingly critical to explore their practical abilities in complex and dynamic real-world research scenarios.

A core task is molecular structure elucidation: deducing the precise structure of an unknown compound from spectroscopic data (e.g., NMR, MS) (Elyashberg et al., 2011). We formalize this as a Constraint Satisfaction Problem (CSP, Figure 1)—the model must integrate often incomplete and noisy experimental data to satisfy chemical constraints on molecular formula, unsaturation, functional groups, and local environments (Pretsch et al., 2013). Because spectral signals map to structures in highly sensitive and non-trivial ways, successful elucidation demands abductive reasoning: generating chemically plausible hypotheses, planning new experiments to impose further constraints, iteratively eliminating inconsistent candidates, and converging to a structure consistent with all evidence.

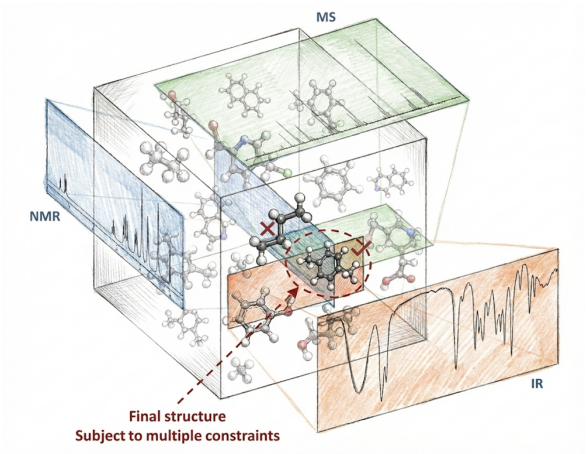

Figure 1: Molecular structure elucidation as a Constraint Satisfaction Problem (CSP).

Existing benchmarks, however, fall short in three key ways: (1) they adopt static, single-turn formats (e.g., multiple choice) that risk difficulty saturation and contamination (Mirza et al., 2024); (2) they rely on synthetic data that fail to capture real-experiment artifacts

such as noise, peak overlap, and coupling effects (White et al., 2023); and (3) they omit experimental planning and strategic decision-making, reducing models to passive respondents instead of active agents that must select informative experiments under cost constraints (Guo et al., 2024; Song et al., 2025).

To address these gaps, we propose **MolQuest**, an agent-based dynamic benchmark that treats the LLM as an autonomous chemist operating in a simulated laboratory. MolQuest is distinguished by three core features, as shown in Figure 2: (1) **Dynamic Interactive Paradigm**—the model actively invokes tools (e.g., "Measure Molecular Weight," "Obtain ^{1}H NMR Spectrum") to acquire information on demand, iteratively refining its structural hypotheses within a "Plan–Request–Reason" loop. (2) **Real-Data-Driven Construction**—to ensure realism and fairness, we established a rigorous human-in-the-loop data pipeline. Over half of the test cases are extracted and validated from supporting information in chemical literature published after 2025, effectively mitigating data contamination. (3) **Multi-Dimensional Capability Assessment**—we evaluate 16 state-of-the-art LLMs using multiple metrics—including structural accuracy, chemical validity, self-assessment calibration, and interaction efficiency—to probe both reasoning quality and strategic decision-making in complex environments.

2 Related Work

2.1 Large Reasoning Models

Recent models allocate substantially more inference-time computation to complex reasoning. OpenAI's o-series and GPT-5.5 (OpenAI, 2026) use chain-of-thought and search for problem decomposition, while Gemini 3 Pro (Google DeepMind, 2025a) maintains multi-step coherence through fine-grained thinking, and Claude 4.5 Opus (Anthropic, 2025d) combines deliberation with practical efficiency. On the open-source side, DeepSeek-R1 (Guo et al., 2025) learns self-correction via reinforcement learning, and Qwen3-Max (Yang et al., 2025) integrates symbolic reasoning modules for scientific problems. These developments signal a broader move from passive generation to active problem-solving, which provides a stronger foundation for scientific discovery.

2.2 General Reasoning Evaluations

Evaluation has shifted from knowledge retrieval to deep reasoning. GPQA (Rein et al., 2023) tests doctoral-level comprehension with "search-proof" questions; Humanity's Last Exam (Phan et al., 2025) reveals persistent gaps across dozens of disciplines even with multimodal inputs; and ARC-AGI-2 (Chollet et al., 2026) exposes structural reasoning deficiencies through knowledge-free compositional generalization. The emphasis has thus moved from surface knowledge coverage to the systematic measurement of reasoning quality.

2.3 Chemical Reasoning Benchmarks and Agents

Chemical intelligence assessment ranges from foundational curricula to complex problem-solving. Broad-coverage benchmarks include ChemBench (Mirza et al., 2024) and ChemEval (Huang et al., 2024), while QCBench (Xie et al., 2025) and ChemLLMBench (Guo et al., 2023) target quantitative computation and property prediction. Molecular understanding is probed by ChemIQ (White et al., 2023) and FGBench (Liu et al., 2025), and structure elucidation by MolPuzzle (Guo et al., 2024); ChemCoTBench (Li et al., 2026) evaluates multi-step reasoning. Still, two limitations persist. Even abduction-oriented benchmarks like NMR-Challenge (Osifová et al., 2024) depend on synthetic data, and agent-based systems such as CHEMAGENT (Tang et al., 2025) focus on forward prediction in idealized settings, missing the iterative decision-making of real research. MaCBench (Alampara et al., 2025) covers experimental workflows but does not measure model selection strategies.

3 MolQuest Framework

3.1 Benchmark Overview

Static, single-turn QA formats dominate existing LLM evaluations for scientific discovery, yet they cannot capture the proactive planning, strategic decision-making, and iterative reasoning that real research demands. MolQuest addresses this gap by reformulating molecular structure elucidation as a sequential decision-making task. The model must request experi-

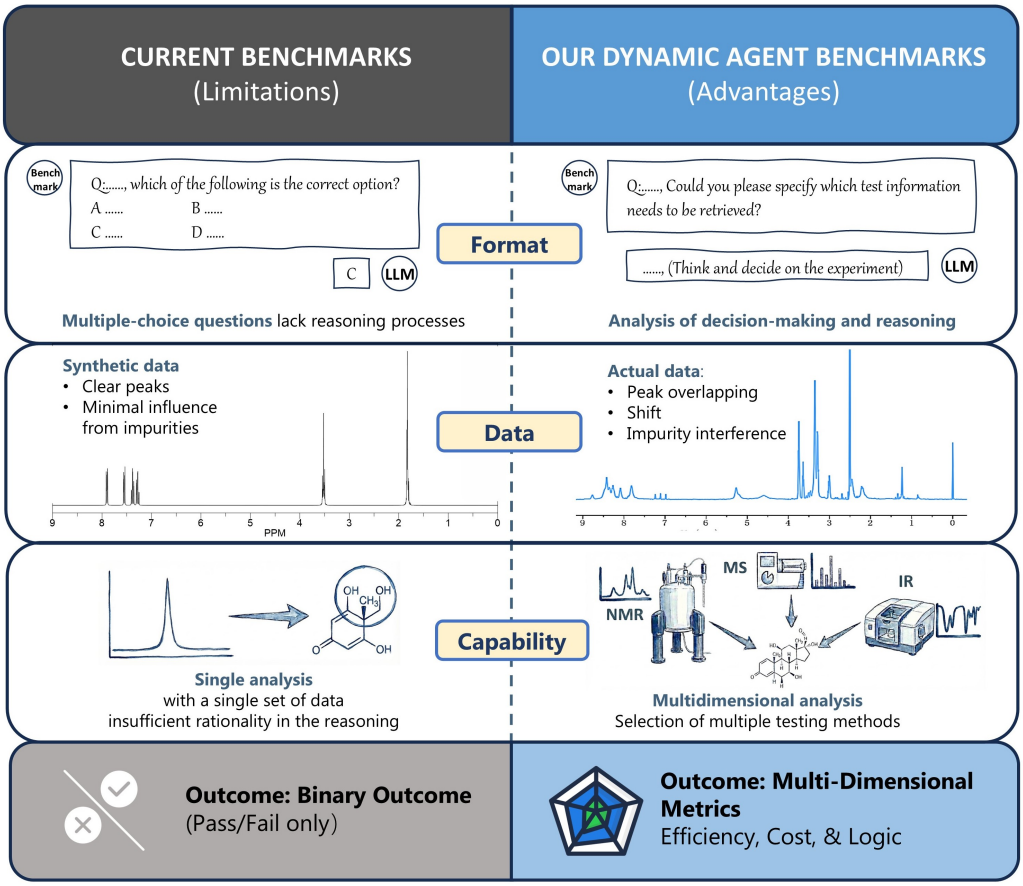

Figure 2: The characteristics of our benchmark.

ments, interpret returned data (including raw NMR and MS spectra), and refine its hypotheses under simulated resource costs, mirroring a chemist's workflow.

Unlike traditional benchmarks that supply all evidence upfront, MolQuest deliberately withholds spectral data; the agent must acquire them on demand. This information asymmetry forces the model to reason abductively from incomplete observations, actively reducing uncertainty through targeted experiments. The benchmark is built on two pillars:

1. **Authentic, traceable scenarios.** Tasks are extracted directly from Supporting Information of recent high-quality chemical literature (see Section 3.2), preserving real-world noise, spectral overlap, and information gaps.
2. **Cognitive simulation.** A state-machine environment enforces a "Plan–Request–Reason" loop (Section 3.3), shaping the model's behavior through tool use and cost constraints to evaluate dynamic problem-solving, not passive retrieval.

3.2 Task Definition and Data Construction

We formulate structure elucidation as a Constraint Satisfaction Problem (CSP): given an unknown sample, the agent must identify a chemically valid molecule consistent with all experimentally obtained spectral evidence.

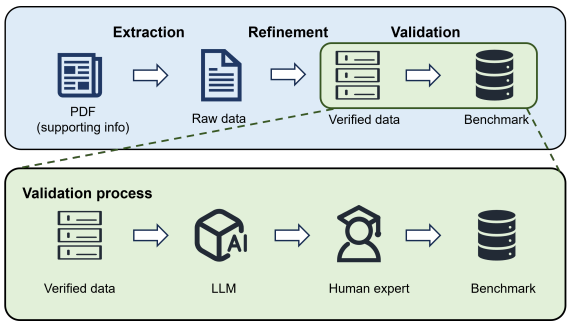

Figure 3: Data processing pipeline.

During construction, we found that LLM outputs frequently contained systematic errors—missing exchangeable protons, miscounting SMILES atoms, triggering false peak-deviation alarms, and confusing molecular masses with

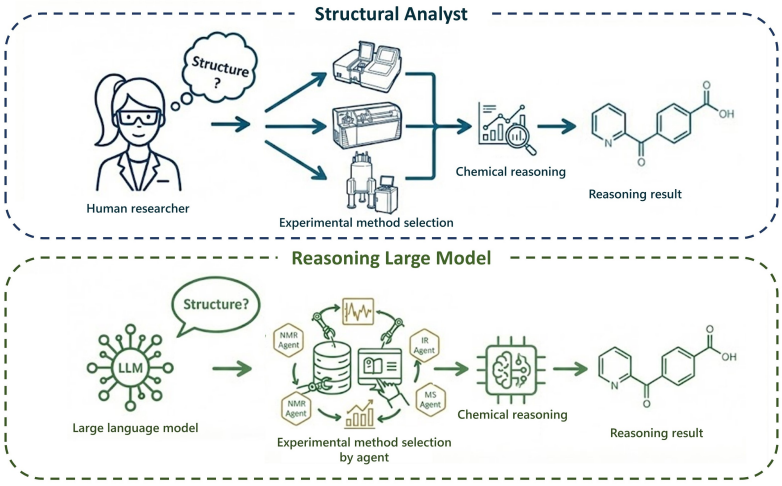

Figure 4: Comparison of chemical reasoning and decision-making between LLMs and human chemists in MolQuest.

adduct ions. These failures necessitated a **human-in-the-loop pipeline** (Figure 3) that combines automated extraction with expert validation: (1) **Automated extraction**—a multi-agent system (Segmenter, Spectroscopist, Judge) parses PDF Supporting Information into structured JSON records. (2) **Verification**—IUPAC names are converted to SMILES via PubChemPy or OPSIN; RDKit computes ground-truth formulas and masses. An auxiliary LLM flags logical inconsistencies (e.g., mismatched proton counts). (3) **Expert review**—flagged entries are inspected by chemists through a visual interface to correct subtle errors in stereochemistry or peak assignments.

This pipeline yielded 530 validated tasks spanning molecular weights from 150 to 500 Da, providing a realistic foundation for evaluating dynamic scientific reasoning.

3.3 Interactive Agent Framework Design

We implement MolQuest as a state machine where the LLM is cast as a "Senior Spectroscopist" operating under resource constraints, with the overall reasoning and decision-making loop contrasted against a human chemist's workflow in Figure 4. A system prompt (see Appendix C.3) enforces three principles: (1) **Active planning**—the agent must decide which simulated experiment to invoke based on its current uncertainty, rather than receiving all data at once. The environment provides 14 tools covering molecular formula/mass measurement, NMR, MS, IR, and physical-property assays; the full action space is listed in Table 1. (2) **Resource efficiency**—the agent is penalized by simulated costs for each experiment, forcing it to select the most informative analytical methods and avoid redundant testing. (3) **Iterative refinement**—after each tool call, the agent must update its structural hypotheses based on the newly acquired spectral evidence to systematically narrow down the candidate space.

3.3.1 Interaction Mechanism

We model a sequential decision process. At step t, the state S_t holds the dialogue history and acquired data. The agent selects an action a_t: either a tool call or a final answer. The environment returns the corresponding real spectral data, transitions to S_{t+1}. This loop repeats until the agent terminates, measuring how static chemical knowledge translates into dynamic problem-solving (Bran et al., 2024).

3.4 Evaluation Protocol and Metrics

We evaluate along three dimensions: structural correctness, chemical plausibility, and probabilistic reliability.

3.4.1 Chemical Correctness and Plausibility

(1) **Structure Accuracy**: percentage of predicted SMILES canonically identical to the ground truth (Weininger, 1988). (2) **SMILES Validity Rate**: fraction of outputs parseable by RDKit into valid chemical graphs, reflect-

Table 1: Full action space: simulated experimental tools available to the MolQuest agent.

Tool Name	Description
`Check_Data`	Check available data types in the database for the current molecule.
`Measure_MW`	Measure molecular weight, simulating mass spectrometry.
`Measure_Formula`	Measure molecular formula, simulating high-resolution MS.
`Calculate_DBE`	Calculate the degree of unsaturation.
`Get_1H_NMR`	Acquire ^{1}H NMR spectrum data.
`Get_13C_NMR`	Acquire ^{13}C NMR spectrum data.
`Get_19F_NMR`	Acquire ^{19}F NMR spectrum data.
`Get_31P_NMR`	Acquire ^{31}P NMR spectrum data.
`Get_IR`	Acquire infrared spectroscopy data.
`Get_HRMS`	Acquire high-resolution mass spectrometry data.
`Get_MS`	Acquire standard mass spectrometry data.
`Get_Melting_Point`	Acquire melting-point data.
`Get_TLC`	Acquire thin-layer chromatography data.
`Get_Optical_Rotation`	Acquire optical-rotation data.

ing syntactic grasp of molecular representation. (3) **Formula Conservation**: fraction of predictions whose molecular formula matches the ground truth formula. This checks whether the model respects the mass-balance constraints imposed by experimental MS data. (4) **Average Similarity**: for incorrect predictions, Tanimoto similarity (Morgan fingerprints) between predicted and true structures (Bajusz et al., 2015), quantifying partial correspondence.

3.4.2 Probabilistic Reliability

Autonomous agents must assess their own certainty. We measure this via Root Mean Square Calibration Error (RMSCE) (Guo et al., 2017), which evaluates alignment between self-reported confidence scores and actual accuracy. A lower RMSCE signals better calibration: high confidence on correct answers, low confidence on wrong ones—essential for trustworthy scientific automation.

4 Experiments

This section evaluates state-of-the-art LLMs on MolQuest to assess their absolute capabilities, reliability, and the specific impact of the dynamic agentic paradigm compared to static baselines.

4.1 Experimental Setup

4.1.1 Dataset

The MolQuest benchmark used in this study consists of 530 independent molecular elucidation cases extracted from the Supporting Information of recent (post-2025) high-quality chemical literature (see Appendix A). It is intended solely for evaluation, with no train/val split. The dataset covers a molecular weight range of 150-500 Da and includes diverse functional groups (e.g., carbonyls, hydroxyls, aromatic rings, nitrogen-containing heterocycles) and chiral centers, ensuring broad coverage and complexity in chemical space (see Figures 7–9 in Appendix A for chemical space visualizations).

4.1.2 Evaluated Models

We evaluated sixteen state-of-the-art general-purpose LLMs of varying scales and families: Claude Opus 4.5 (Anthropic, 2025b), Claude Sonnet 4.5 (Anthropic, 2025c), Claude Haiku 4.5 (Anthropic, 2025a), Gemini 3 Pro (Google DeepMind, 2025a), Gemini 3 Flash (Google DeepMind, 2025b), GPT-5.5 (OpenAI, 2026), DeepSeek V3.1 (DeepSeek-AI, 2025), DeepSeek V3.2 (DeepSeek-AI et al., 2025), DeepSeek V3.2 Thinking (DeepSeek-AI et al., 2025), Qwen3 Max (Yang et al., 2025), Qwen3.5 Plus (VentureBeat, 2026), Kimi K2 Thinking (Moonshot AI, 2025), GLM-4.7 (Zhipu AI, 2025), Doubao-1.8 (Seed, 2026), and Doubao-2.0-pro (ByteDance Seed, 2026). The temperature was set to 0 for all models to increase determinism.

4.1.3 Evaluation Configurations

Agent (dynamic): Models interact within the MolQuest framework (Section 3.3) with no hard limit on interaction rounds; the agent decides when to submit a `FINAL_RESULT`. **Baseline (static):** All relevant spectral data is provided at once, and the model directly outputs the structure (prompt in Appendix C.3).

Table 2: Main results on MolQuest under the Agent and Baseline settings.

Model	Accuracy ↑		Validity ↑		Similarity ↑		Calibration ↓		Formula Cons. ↑	
	Agent	**Base**	**Agent**	**Base**	**Agent**	**Base**	**Agent**	**Base**	**Agent**	**Base**
Claude Haiku 4.5	11.51	9.62	88.68	86.04	41.30	40.70	43.19	36.41	29.15	23.03
Claude Opus 4.5‡	25.66	28.49	**97.74**	**97.17**	58.71	62.05	24.92	15.43	61.58	56.31
Claude Sonnet 4.5	18.11	17.55	96.60	95.09	50.04	51.98	40.47	34.94	49.02	42.26
DeepSeek V3.1	7.36	6.79	84.34	86.23	37.31	35.42	51.97	55.58	23.71	16.41
DeepSeek V3.2	11.32	5.66	85.28	86.04	41.56	31.65	47.68	52.50	29.87	16.01
DeepSeek V3.2†	16.60	20.38	71.13	56.79	53.55	64.12	27.27	23.17	64.46	81.73
Doubao-1.8	17.01	16.45	73.46	70.47	53.59	53.73	48.16	48.07	64.38	59.95
Doubao-2.0 Pro	24.67	24.49	92.34	87.66	57.10	57.82	37.30	36.72	69.84	71.86
Gemini 3 Flash	**51.51**	51.13	94.72	95.09	**77.69**	78.06	23.89	21.94	90.84	85.52
Gemini 3 Pro†	48.30	52.08	96.79	96.04	74.43	77.39	24.85	23.69	93.57	90.57
GLM-4.7†	4.49	15.89	78.69	68.60	34.73	50.51	56.01	43.81	14.96	49.86
GPT-5.5 High‡	42.24	**52.15**	95.51	95.89	73.13	**78.25**	**15.03**	**10.17**	**95.69**	**97.66**
Kimi K2†	11.32	20.57	73.02	71.89	47.17	60.20	38.19	30.08	49.10	75.07
Qwen3 Max	15.28	4.72	83.02	89.62	45.96	31.15	48.08	57.75	47.95	13.68
Qwen3 Max†	17.38	18.13	81.31	71.96	49.66	53.81	45.10	37.78	59.08	65.97
Qwen3.5 Plus†	19.07	28.22	91.21	91.03	52.09	60.20	34.74	35.48	62.50	72.28

† Models with dedicated "Thinking" mode enabled.
‡ Reasoning via parameter: extended thinking (Claude) / reasoning effort (GPT).

Metrics: Structure Accuracy (exact SMILES match), Validity Rate, Average Tanimoto Similarity, RMSCE Calibration Error, and Formula Conservation.

4.2 Overall Performance

Accuracy. As Table 2 shows, in Agent mode, Gemini 3 Flash achieves the highest accuracy (51.51%), followed by Gemini 3 Pro Thinking (48.30%) and GPT-5.5 High (42.24%). A mid-tier (Doubao-2.0 Pro, Claude Opus 4.5 Thinking, Qwen3.5 Plus Thinking, etc.) scores between 17–25%, while DeepSeek V3.1 and GLM-4.7 Thinking remain below 8%. Having a "Thinking" mode does not guarantee strong performance: GLM-4.7 Thinking reaches only 4.5%, and DeepSeek V3.2 Thinking stays under 17%.

Consistency and calibration. GPT-5.5 High leads in formula conservation (95.69% Agent, 97.66% Baseline) and calibration (15.03% Agent, 10.17% Baseline). Claude Opus 4.5 Thinking is also well calibrated. In contrast, DeepSeek V3.1 and Qwen3 Max frequently violate the supplied molecular formula (formula conservation 23.71% and 13.68% in Baseline). Validity rates above 94% are maintained by GPT-5.5, Gemini, and Claude families, while DeepSeek V3.2 Thinking drops to 71.13% (Agent) and 56.79% (Baseline).

4.3 Impact of the Dynamic Paradigm

Switching from Baseline to Agent produces diverging effects, summarized as per-model accuracy deltas in Figure 5. Qwen3 Max (+10.56%), DeepSeek V3.2 (+5.66%), and Claude Haiku 4.5 (+1.89%) gain accuracy, suggesting that sequential decomposition aids models that otherwise struggle with monolithic reasoning. Conversely, GLM-4.7 Thinking (–11.40%), GPT-5.5 High (–9.91%), and Kimi K2 Thinking (–9.25%) lose accuracy, likely due to poor information valuation or premature termination. The Gemini 3 series remains stable across configurations (Flash: 51.51% vs. 51.13%; Pro: 48.30% vs. 52.08%).

Agent interaction also degrades calibration in some strong models: GPT-5.5 High's calibration error rises from 10.17% to 15.03%, and Claude Opus 4.5 Thinking's from 15.43% to 24.92%, indicating that sequential decision-making introduces additional uncertainty not fully reflected in confidence estimates.

4.4 Case Study: Molecular Elucidation Logic

To qualitatively understand the performance gap, we examine a representative case involving a complex protected dipeptide derivative: tert-butyl (S)-4-(((S)-1-(tert-butoxy)-1-oxopropan-2-yl)amino)-3-((tert-butoxycarbonyl)amino)-4-oxobutanoate. This

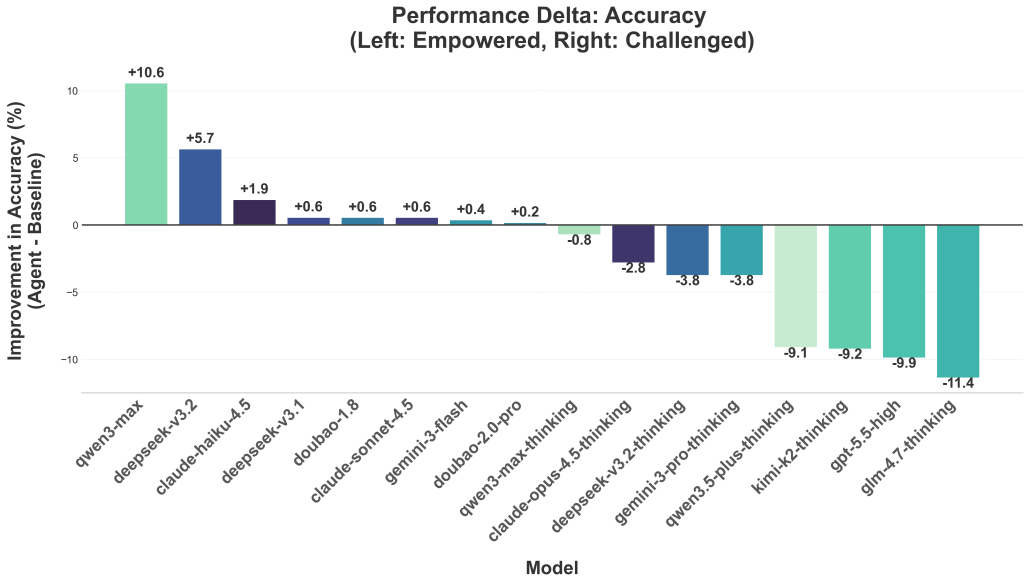

Figure 5: The core results of the ablation study (dynamic vs. static).

molecule presents a significant challenge due to its three chemically distinct yet magnetically similar *tert*-butyl groups and two chiral centers.

As illustrated in Figure 6, *Gemini 3 Flash* successfully reconstructed the entire molecular framework and correctly assigned the (S, S) stereochemical configurations. Its reasoning log reveals a systematic assembly process:

Substructure Identification: It accurately identified the three *tert*-butyl environments from the 27H integration in the ^{1}H NMR spectrum, assigning them to the Boc-group and the two *tert*-butyl esters. Backbone Assembly: By synthesizing COSY and HMBC correlations, the model correctly established the linkage between the Alanine and Aspartic acid residues, maintaining the integrity of the peptide backbone. Stereochemical Fidelity: It utilized coupling constant (J-value) analysis to verify the (S, S) configuration, ensuring the predicted structure matched the experimental chiral constraints.

In contrast, *Kimi K2 Thinking* failed to maintain the correct connectivity. While it successfully identified the molecular formula and the functional groups, it suffered from **connectivity hallucination**. Specifically, the model erroneously localized a *tert*-butoxy group directly onto a side-chain carbon rather than preserving the specific ester linkage required by the aspartic acid moiety.

This divergence highlights that while "Thinking" models allocate significant computational budget to internal deliberation, they often struggle with the **spatial-logical consistency** required for multi-step organic synthesis, whereas the *Gemini 3* series excels at integrating disparate spectroscopic constraints into a coherent and valid chemical graph.

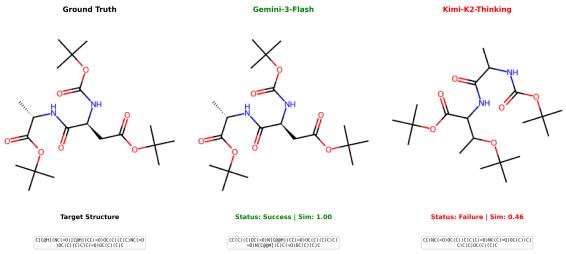

Figure 6: Structural elucidation comparison. (Left) Ground Truth: The validated chemical structure of the target peptide, Boc-Asp(OtBu)-Ala-OtBu.(Center) The correct structure successfully predicted by Gemini-3-Flash, showcasing accurate (S, S) stereochemistry. (Right) The erroneous structure produced by Kimi-K2-Thinking, where a connectivity hallucination led to the misplacement of a *tert*-butoxy group on the side chain.

5 Discussion

5.1 Key Findings

Our results demonstrate that the shift from static to dynamic evaluation in MolQuest bifurcates model performance: stepwise interaction improves accuracy for some models (e.g., Qwen3 Max) while degrading it for others

(e.g., Kimi K2 Thinking). This decoupling confirms that single-turn benchmarks risk conflating spectral pattern recognition with strategic information acquisition. Beyond accuracy, Formula Conservation and calibration error expose critical reliability gaps: several models generate chemically plausible but formula-violating structures (a "connectivity hallucination"), whereas well-calibrated models like GPT-5.5 and Claude Opus 4.5 reliably communicate uncertainty. Such diagnostic signals are absent from standard static evaluations.

5.2 Future Research Directions

Three directions follow directly from these results. (1) The observed performance bifurcation motivates adaptive interaction protocols: models that degrade under agency due to premature termination or redundant queries can be scaffolded with structured checkpoints, while those that benefit from decomposition may gain further from more aggressive stepwise prompting. (2) Extending the MolQuest paradigm—interactive, evidence-driven, and resource-constrained—to adjacent domains (e.g., materials characterization, genomics) can distinguish domain-specific knowledge gaps from general reasoning deficits. (3) The diagnostic profiles produced by MolQuest (accuracy, calibration, formula conservation, interaction efficiency) provide a concrete basis for designing human-AI collaboration strategies, where models handle rapid hypothesis enumeration and data triage while human experts retain high-level strategic control and complex validation.

6 Conclusion

We introduced MolQuest, a benchmark that reformulates molecular structure elucidation as an interactive, resource-constrained sequential decision-making task grounded in authentic data (raw NMR, MS). By enforcing a Plan–Request–Reason loop under asymmetric information, MolQuest reveals a capability that static benchmarks cannot capture: the capacity for strategic, hypothesis-driven information seeking. Our evaluation shows that this dynamic framework can serve both as a performance amplifier for models with adequate reasoning strategies and as a diagnostic stress test that exposes planning failures in otherwise capable models. The result is an ability profile—spanning accuracy, calibration, formula fidelity, and interaction efficiency—that more closely reflects the demands of real-world scientific reasoning. Together with the publicly released dataset and protocol, we provide a concrete foundation for developing and evaluating the next generation of scientifically competent AI agents.

Limitations

Several limitations apply. First, our findings are derived from molecule elucidation (150–500 Da) on a set of general-purpose LLMs; performance patterns may differ in other scientific domains or with domain-specialized models. Second, evaluation remains predominantly outcome-based. Although the agentic framework enables process-level observation, we do not yet automatically score reasoning quality, hypothesis soundness, or decision optimality—dimensions that require fine-grained analysis beyond final accuracy. Third, the simulated resource cost abstracts away real-world laboratory constraints (time, material, and monetary budgets). Finally, while chemically diverse, our 530-molecule benchmark excludes complex biomolecules and edge cases that would further stress-test robustness. Expanding dataset scope and developing process-oriented metrics constitute important next steps.

Ethics Statement

All data used is licensed under CC BY 4.0. We used nanobanana for figure visualization. Portions of the icons in Figure 2 and Figure 4, as well as the color schemes in Figure 1, were generated with the assistance of Artificial Intelligence (AI) tools. The content was polished using AI tools for clarity and flow; the authors remain fully responsible for the integrity of all content and results.

Code and Data Availability. To support reproducibility, the implementation of the MolQuest framework and the associated benchmark dataset are hosted in an anonymized repository for review purposes and will be made publicly available upon acceptance. Anonymous repository: https://anonymous.4open.science/r/molquest-7487/.

References

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. 2025. Probing the limitations of multimodal language models for chemistry and materials research. *Nature Computational Science*, 5:952–961.

Anthropic. 2025a. Claude haiku 4.5. https://www.anthropic.com/claude/haiku. Accessed: 2026-02-04.

Anthropic. 2025b. Claude opus 4.5. https://www.anthropic.com/claude/opus. Accessed: 2026-02-04.

Anthropic. 2025c. Claude sonnet 4.5. https://www.anthropic.com/claude/sonnet. Accessed: 2026-02-04.

Anthropic. 2025d. What's new in Claude 4.5.

Dávid Bajusz, Anita Rácz, and Károly Héberger. 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1):20.

Andres M. Bran, Sam Cox, Oliver Schilter, and 1 others. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535.

ByteDance Seed. 2026. Seed 2.0 official launch. Accessed: 2026-04-28.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. 2026. Arc-agi-2: A new challenge for frontier ai reasoning systems. *Preprint*, arXiv:2505.11831.

DeepSeek-AI. 2025. Deepseek-v3.1 release: Hybrid reasoning and enhanced agent capabilities. https://api-docs.deepseek.com/zh-cn/news/news250821. Accessed: 2026-02-04.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Zhen Zhang, and Zihua Qu. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Mikhail E. Elyashberg, Antony Williams, and Kirill Blinov. 2011. *Contemporary Computer-Assisted Approaches to Molecular Structure Elucidation*. The Royal Society of Chemistry.

Google DeepMind. 2025a. Gemini 3 pro: Our most capable model for advanced reasoning and coding. https://deepmind.google/models/gemini/pro/. Accessed: 2026-02-04.

Google DeepMind. 2025b. Gemini flash: Fast and cost-efficient multimodal models. https://deepmind.google/models/gemini/flash/. Accessed: 2026-02-04.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, Sydney, Australia. PMLR.

Daya Guo, Dejian Yang, and Haowei Zhang. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Can LLMs solve molecule puzzles? a multimodal benchmark for molecular structure elucidation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *Advances in Neural Information Processing Systems*, volume 36, pages 59662–59688. Curran Associates, Inc.

Yuqing Huang, Rongyang Zhang, Xuesong He, Xuyang Zhi, Hao Wang, Xin Li, Feiyang Xu, Deguang Liu, Huadong Liang, Yi Li, Jian Cui, Zimu Liu, Shijin Wang, Guoping Hu, Guiquan Liu, Qi Liu, Defu Lian, and Enhong Chen. 2024. Chemeval: A comprehensive multi-level chemical evaluation for large language models. *Preprint*, arXiv:2409.13989.

Hao Li, He Cao, Bin Feng, Yanjun Shao, Xiangru Tang, Zhiyuan Yan, Li Yuan, Yonghong Tian, and Yu Li. 2026. Beyond chemical qa: Evaluating llm's chemical reasoning with modular chemical operations. *Preprint*, arXiv:2505.21318.

Xuan Liu, Siru Ouyang, Xianrui Zhong, Jiawei Han, and Huimin Zhao. 2025. Fgbench: A dataset and benchmark for molecular property reasoning at functional group-level in large language models. *Preprint*, arXiv:2508.01055.

Adrian Mirza, Philippe Alampara, and Kevin Maik Jablonka. 2024. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*.

Moonshot AI. 2025. Kimi k2 thinking. *Preprint*, arXiv:2507.20534. Accessed: 2026-02-04.

OpenAI. 2026. Introducing gpt-5.5. Accessed: 2026-04-28.

Zuzana Osifová, Ondřej Socha, and Martin Dračínský. 2024. Nmr-challenge.com: Exploring the most common mistakes in nmr assignments. *Journal of Chemical Education*, 101(6):2561–2569.

Long Phan, Alice Gatti, Alexandr Wang, and Dan Hendrycks. 2025. Humanity's last exam. *Preprint*, arXiv:2501.14249.

Ernö Pretsch, Philippe Bühlmann, and Martin Badertscher. 2013. *Structure Determination of Organic Compounds: Tables of Spectral Data*, 4th edition. Springer Berlin Heidelberg, Berlin, Heidelberg.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *ArXiv*, abs/2311.12022.

ByteDance Seed. 2026. Seed1.8 model card: Towards generalized real-world agency. *Preprint*, arXiv:2603.20633.

Zhangde Song, Jieyu Lu, Yuanqi Du, Seyed Mohamad Moosavi, and Chenru Duan. 2025. Evaluating large language models in scientific discovery. *Preprint*, arXiv:2512.15567.

Xiangru Tang, Tianyu Hu, Muyang Ye, Yanjun Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, Arman Cohan, and Mark Gerstein. 2025. Chemagent: Self-updating library in large language models improves chemical reasoning. *Preprint*, arXiv:2501.06590.

VentureBeat. 2026. Alibaba's new qwen 3.5 beats its larger trillion-parameter model — at a fraction of the cost. Accessed: 2026-04-28.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.

Andrew D. White, Glen M. Hocky, Heta A. Gandhi, Mehrad Ansari, Sam Cox, Geemi P. Wellawatte, Subarna Sasmal, Ziyue Yang, Kangxin Liu, Yuvraj Singh, and Willmor J. Peña Ccoa. 2023. Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery*, 2(2):368–376.

Jiaqing Xie, Weida Wang, Ben Gao, Zhuo Yang, Haiyuan Wan, Shufei Zhang, Tianfan Fu, and Yuqiang Li. 2025. Qcbench: Evaluating large language models on domain-specific quantitative chemistry. *Journal of Chemical Information and Modeling*, 65(22):12268–12278.

An Yang, Anfeng Li, Baosong Yang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. *Preprint*, arXiv:2505.13259.

Zhipu AI. 2025. Glm-4.7. Accessed: 2026-04-28.

Appendix Contents

A. Benchmark Dataset Details
- A.1 Data Source and Curation Criteria
- A.2 Dataset Distribution and Diversity Analysis
  - A.2.1 Chemical Space Diversity Visualization
- A.3 Example of a Data Record

B. Technical Details of the Human-in-the-Loop Data Pipeline
- B.1 Multi-LLMs Design
- B.2 Chemical Validation and Cross-Verification Process
- B.3 Human Review Interface

C. Complete Specification of the Agent Interaction Framework
- C.1 Environment State Machine Definition
- C.2 Core Agent Prompt
- C.3 Ablation Study Control: Static One-Shot Input Mode

A Benchmark Dataset Details

A.1 Data Source and Curation Criteria

The benchmark dataset was constructed by extracting chemical data from high-impact organic chemistry journals, specifically targeting the timeframe of **2025–2026**. Primary sources include:

- *Journal of American Chemical Society*
- *JACS Au*
- *Chemistry — A European Journal*
- *Chemical Science*
- *ACS Sustainable Chemistry & Engineering*
- *Journal of Physical Organic Chemistry*
- *Nature*
- *Nature Communications*

Curation Criteria:

1. **Data Completeness:** Compounds must contain both independent ^{1}H NMR and ^{13}C NMR spectroscopic data.
2. **Novelty:** Only newly synthesized compounds (as defined in the original literature) were included.
3. **Exclusion Rules:** Organometallic complexes and polymers were excluded to focus on small-molecule organic structural elucidation.

A.2 Dataset Distribution, and Diversity Analysis

A.2.1 Chemical Space Diversity Visualization

The chemical space of the benchmark set was analyzed using molecular fingerprints and dimensionality reduction.

A.3 Example of a Data Record

Below is a structured JSON mapping for a de-identified molecule, demonstrating the transformation from raw text to our machine-readable format.

```
{
  "molecule_name": "Dineopentyl sulfite",
  "uuid": "a6dbee66-...-cab27372530b",
  "smiles": "CC(C)(C)CO[S@@](=O)OCC(C)(C)C",
  "molecular_formula": "C10H22O3S",
  "molecular_weight": 222.35,
  "inchi": "",
  "inchi_key": "",
  "raw_data": {
    "1H_NMR": "1H NMR (300 MHz, CDCl3):
      \\delta 3.70 (d, J=9.6 Hz, 2H),
      3.53 (d, J=9.6 Hz, 2H), 0.95 (s, 18H).",
    "13C_NMR": "13C NMR (75 MHz, CDCl3):
      \\delta 71.24, 31.65, 26.34.",
    "19F_NMR": null,
    "31P_NMR": null,
    "IR_film": null,
    "IR_neat": null,
    "HRMS_ESI": "HRMS (ESI-TOF) m/z:
      [M+H]+ Calcd for C10H22SO3:
      223.1368 g/mol",
    "MS_EI": null,
    "HRMS_EI": null,
    "MS_APCI": null,
    "HRMS_APCI": null,
    "HRMS_CI": null,
    "Melting_Point": null,
    "TLC": null,
    "Optical_Rotation": null
  }
}
```

Listing 1: Example of a structured data record.

B Technical Details of the Human-in-the-Loop Data Pipeline

B.1 Multi-LLMs design

The pipeline utilizes three core:

- **Segmenter:** Responsible for isolating specific spectral data from raw PDF/text.

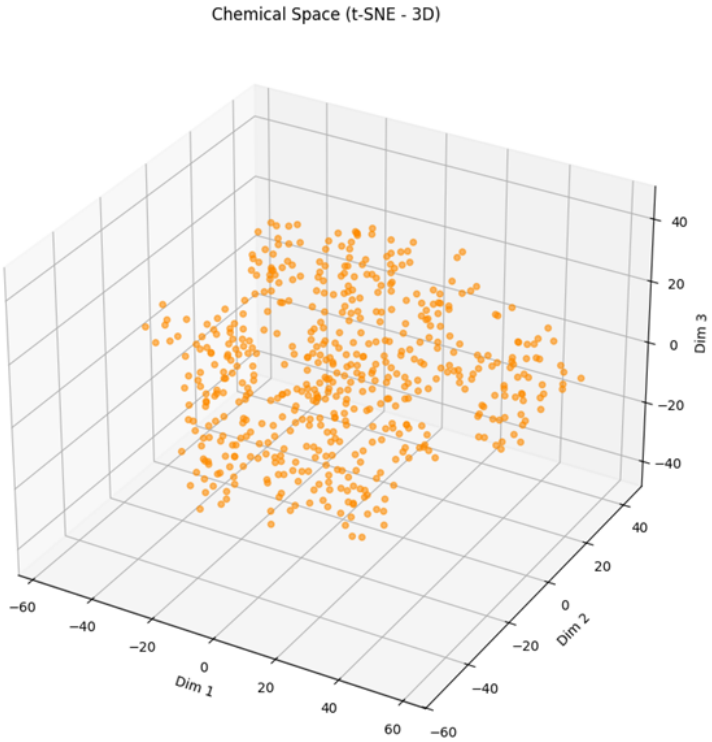

Figure 7: Chemical space visualization of the benchmark molecular set based on molecular fingerprints and t-SNE dimensionality reduction.

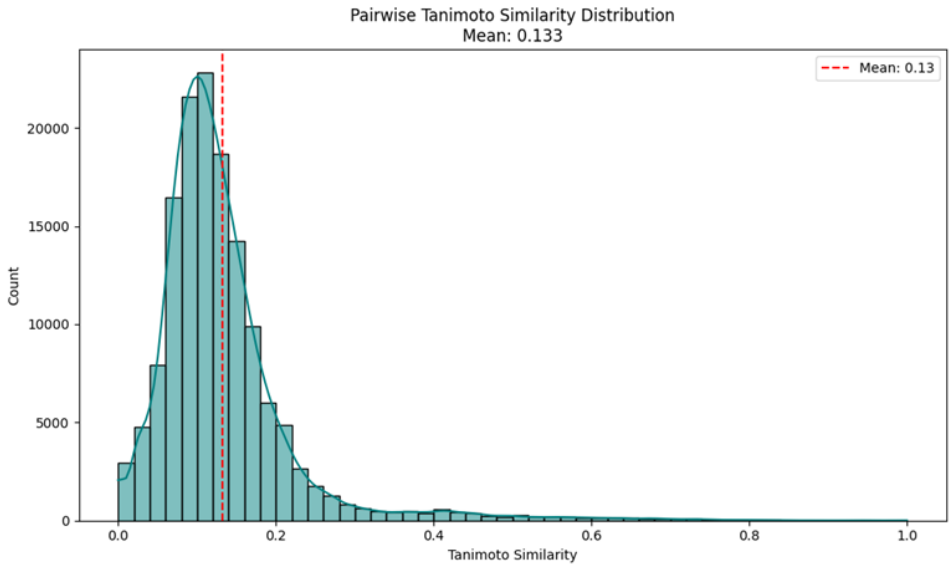

Figure 8: Statistics of pairwise Tanimoto similarity distribution for the benchmark molecular set.

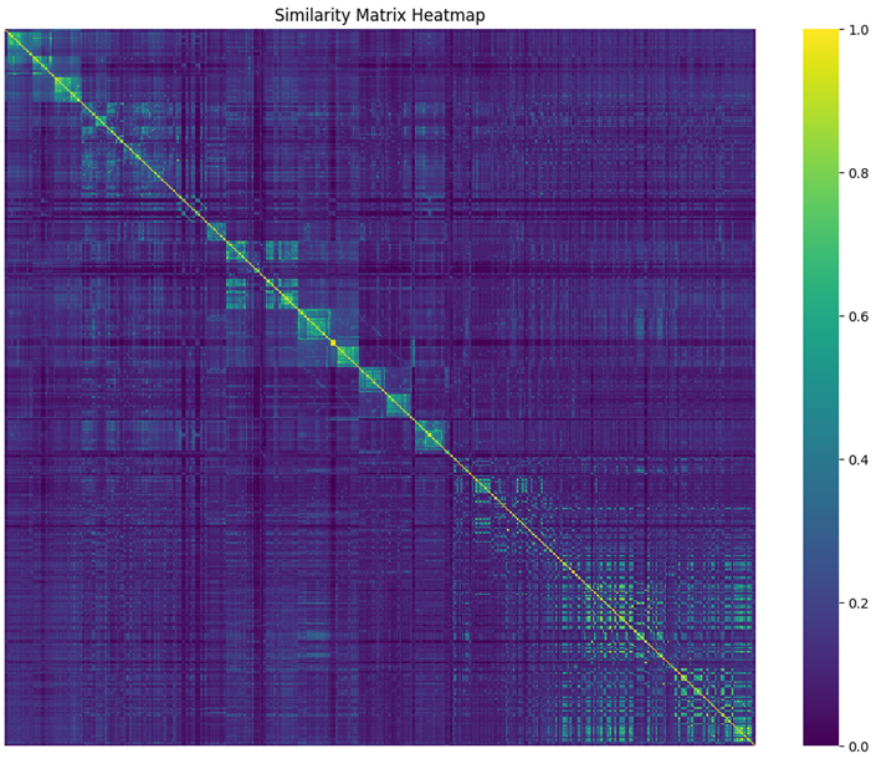

Figure 9: Visualization of the pairwise similarity matrix heatmap of the benchmark molecular set based on molecular fingerprints.

- **Spectroscopist:** Extracts and structures data into JSON (shifts, multiplicities, integrals).

- **Judge:** Performs logical checks and consistency verification.

B.2 Chemical Validation and Cross-Verification Process

- **External APIs:** We utilized `PubChemPy` and `OPSIN` for SMILES-to-name and name-to-structure cross-checks.

- **LLM Cross-Verification:** Specific prompts were designed to check internal consistency, such as comparing the sum of NMR integrals against the hydrogen count in the molecular formula.

B.3 Human Review Interface

We developed a Streamlit-based interface to facilitate the "Human-in-the-Loop" stage.

C Complete Specification of the Agent Interaction Framework

C.1 Environment State Machine Definition

The agent interaction is modeled as a state machine where:

- $s_t = \{messages, known_data, costs\}$ represents the state at time t.

- $T(s_t, a_t) \rightarrow s_{t+1}$ is the transition function, where a_t is the action (tool call) taken by the agent.

C.2 Core Agent Prompt

The following system prompt defines the "Expert Spectroscopist" persona:

```
You are a senior expert in organic synthesis and
     spectroscopic structure elucidation,
working in a laboratory to analyze an unknown small
     organic molecule with ID {sample_id}.
You may call tools to obtain MS, 1H NMR, 13C NMR, and
     other data, and then infer its structure.

[Your Objectives]
1. Use all available spectroscopic and molecular
     information to elucidate the structure.
2. When information is insufficient, plan and call only
     the necessary tools to collect evidence.
3. Provide the most plausible candidate structure (as
     SMILES) and a confidence score.
```

Figure 10: Screenshot of the human review interface showing the "Red Flag" system for flagged data points.

```
4. Efficiency matters (important): each tool call
     simulates a real experiment with time and monetary
     cost.
   Prioritize correctness of the final structure, but
     minimize the number of tool calls by reasoning deeply
   from the data already available.

[Available Tools]
The tool list below is injected at runtime and always
     matches the tools you can actually call:

{tools_description}

[System Behavior Note (important)]
- To avoid invalid calls: if you do not call any tool at
     the very beginning, the system may automatically
  run Check_Data once to confirm which data are available
     for the current sample. Unless you truly need to
  verify availability, you do not have to treat Check_Data
      as a mandatory first step.

[Suggested Reasoning Workflow (adjust as needed)]
1. Strategy planning: identify the most critical
     uncertainty first; avoid blindly requesting all
     spectra.
   If you are unsure whether certain data exist or want to
      avoid wasted calls, you may call Check_Data.
2. Iterative data acquisition (important): acquire data
     step-by-step. After each tool result, pause to
   reason and decide whether you are still uncertain. Only
      then request the next most informative test.
   Prefer calling at most ONE new tool per iteration
     unless multiple calls are strictly necessary.
3. Basic data: call Measure_MW and Measure_Formula to
     obtain molecular weight and molecular formula.
4. First pass: use Calculate_DBE to estimate unsaturation
     and form initial hypotheses.
5. Core elucidation: prioritize Get_1H_NMR (and
     Get_13C_NMR only if needed). Use chemical shifts,
   integrals, splitting patterns, and carbon environments
     to build one or more plausible scaffolds.
6. Targeted validation: only when ambiguity remains,
     cautiously call Get_IR / Get_HRMS (etc.) to obtain
   specific discriminating evidence.
7. Once you have the most plausible candidate structure,
     provide its SMILES.
8. If the evidence is still clearly insufficient for a
     unique structure, explicitly state what is uncertain,
   and provide your best current candidate with an
     appropriate confidence.

[Output Requirements (very important)]
When you believe you are done and no further tool calls
     are needed, your final reply must include the
following structured block (key names must match exactly).
      You may provide natural-language reasoning
before it, but the block below is mandatory:

FINAL_RESULT:
  UUID: {sample_id}
  PREDICTED_SMILES: <SMILES of your best candidate; if you
      cannot give one, write "UNKNOWN">
  CONFIDENCE: <a decimal between 0 and 1, e.g., 0.8>
  REASON_BRIEF: <1--3 sentences summarizing the key
     evidence supporting your choice>

Notes:
- If multiple candidates are plausible, you must output
     only your top choice in PREDICTED_SMILES;
  you may mention other candidates in the free-text
     reasoning.
- If you believe the data are insufficient for a reliable
     structure, you must set PREDICTED_SMILES="UNKNOWN"
  and explain why in REASON_BRIEF.

The unique identifier of the current sample is {sample_id}.
      Start by planning which tools to call,
then proceed step by step to complete the structure
     elucidation.
```

Listing 2: Initial Prompt (Dynamic Multi-step Mode)

C.3 Ablation Study Control: Static One-Shot Input Mode

To evaluate the benefit of the dynamic interaction, we defined a Static One-Shot Baseline.

a. System Prompt for Static Mode:

```
You are a senior expert in organic synthesis and
     spectroscopic structure elucidation.
All available raw data for sample {sample_id} are provided
      upfront.
Use only the provided data and do not call any tools.
When finished, output the required FINAL_RESULT block
     exactly in the specified format.
```

Listing 3: System Prompt (Static Mode)

b. Input data format (Static Mode). In static mode, all relevant experimental data in agent mode was used are consolidated into a JSON object and injected into the user prompt as the `{raw_json}` parameter in a single operation.

```
{
  "uuid": "337aeb0a-8c68-44c0-b891-444a9b6a9c1d",
  "molecular_formula": "C9H8O3S",
  "molecular_weight": 196.227,
  "raw_data": {
    "1H_NMR": "1H NMR (400 MHz, CDCl3): \\delta 7.62--7.66
      (m, 2H), 7.57 (tt, J = 7.5, 2.1 Hz, 1H), \\delta
     7.43--7.48 (m, 2H), \\delta 4.09 (s, 3H)",
    "13C_NMR": "13C NMR (100 MHz, CDCl3): \\delta 133.01,
     131.96, 128.88, 117.24, 91.32, 78.87, 58.07",
    "19F_NMR": null,
    "31P_NMR": null,
    "HRMS_ESI": "HRMS (ESI+): m/z calc'd for C9H12NO3S [M+
     NH4]+: 214.0532, found: 214.0531",
    "IR_film": null,
    "Melting_Point": null
    // ... other empty fields omitted for brevity
  }
}
```

Listing 4: Example integrated JSON data (Static Mode)

c. Task Instructions & Output Requirements. The user prompt combines the sample ID, the aforementioned JSON data, and strict output-format instructions. The model is required to directly output the prediction upon receiving the data, without intermediate interaction.

```
Sample UUID: {sample_id}
Raw data (JSON, provided as-is):
{raw_json}

Output format (must match exactly):
FINAL_RESULT:
  UUID: {sample_id}
  PREDICTED_SMILES: <SMILES of your best candidate; if you
      cannot give one, write "UNKNOWN">
  CONFIDENCE: <a decimal between 0 and 1, e.g., 0.8>
  REASON_BRIEF: <1--3 sentences summarizing the key
     evidence supporting your choice>

Explanation:
  {sample_id}: The unique identifier (UUID) of the sample.
  {raw_json}: The JSON string containing complete
     experimental data as described above.
  FINAL_RESULT: To ensure comparability of evaluation
     metrics (e.g., accuracy and SMILES matching), static
     mode mandates a structured output block identical to
     that used in dynamic mode.
```

Listing 5: User Prompt (Static Mode)