

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

Taolin Han^{1,2} Shuang Wu¹ Jinghang Wang¹
Yuhao Zhou^{1,3} Renquan Lv^{1,3} Bing Zhao¹ Hu Wei¹ Lin Qu¹

¹Alibaba Group, Hangzhou, China

²Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, China

³Zhejiang University, Hangzhou, China

hantaolin23@mails.ucas.ac.cn

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 while minimizing unnecessary tool calls—mirroring the abductive reasoning workflow of a practicing chemist. We evaluate 16 frontier LLM variants under two settings, yielding 32 model–setting configurations, and find that the best Agent-mode result reaches only 51.51% accuracy, while most remain substantially lower. 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 (Pichai et al., 2025), and Qwen3-Max (Qwen Team, 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., 2009). 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.

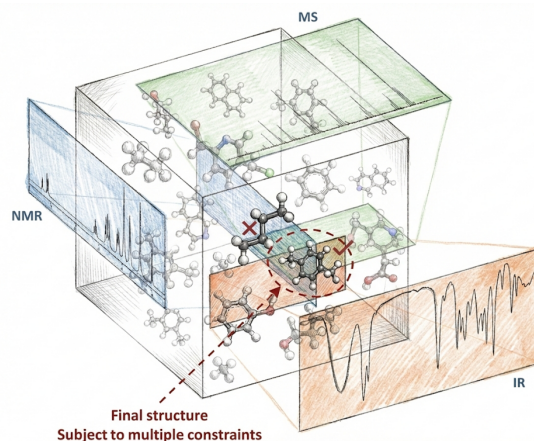


Figure 1: Molecular structure elucidation as a Constraint Satisfaction Problem (CSP).

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) many use curated or synthetic inputs rather than literature-reported experimental records, limiting their coverage of unresolved multiplets, missing evidence modalities, and inconsistent reporting (Mirza et al., 2024; Guo et al., 2024); and (3) they omit exper-

imental planning and strategic decision-making, reducing models to passive respondents instead of active agents that must select informative experiments efficiently (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 ^1H 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 LLM variants under Agent and Baseline settings 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. GPT-5.5 (OpenAI, 2026) supports configurable reasoning effort and tool use, Gemini 3 Pro (Pichai et al., 2025) targets complex multi-step and multimodal reasoning, and Claude Opus 4.5 (Anthropic, 2025b) is designed for coding and agentic workflows. Among open models, DeepSeek-R1 (Guo et al., 2025) develops reasoning capabilities through reinforcement learning, while Qwen3-Max (Qwen Team, 2025) scales model capacity and pretraining data. 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., 2025) 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 (Runcie et al., 2026) and FGBench (Liu et al., 2025), and structure elucidation by MolPuzzle (Guo et al., 2024); ChemCoTBench (Li et al., 2025) evaluates multi-step reasoning. Still, two limitations persist. Even abduction-oriented benchmarks like NMR-Challenge (Osifová et al., 2024) use fixed challenge sets, and agent-based systems such as CHEMAGENT (Tang et al., 2025) focus on library-augmented chemical reasoning rather than interactive experiment selection. 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 experiments, interpret literature-reported characterization records, and refine its hypotheses while its tool calls are recorded as an interaction-efficiency measure, mirroring a chemist’s iterative workflow.

Unlike traditional benchmarks that supply all evidence upfront, MolQuest deliberately withholds spectral data; the agent must acquire them on de-

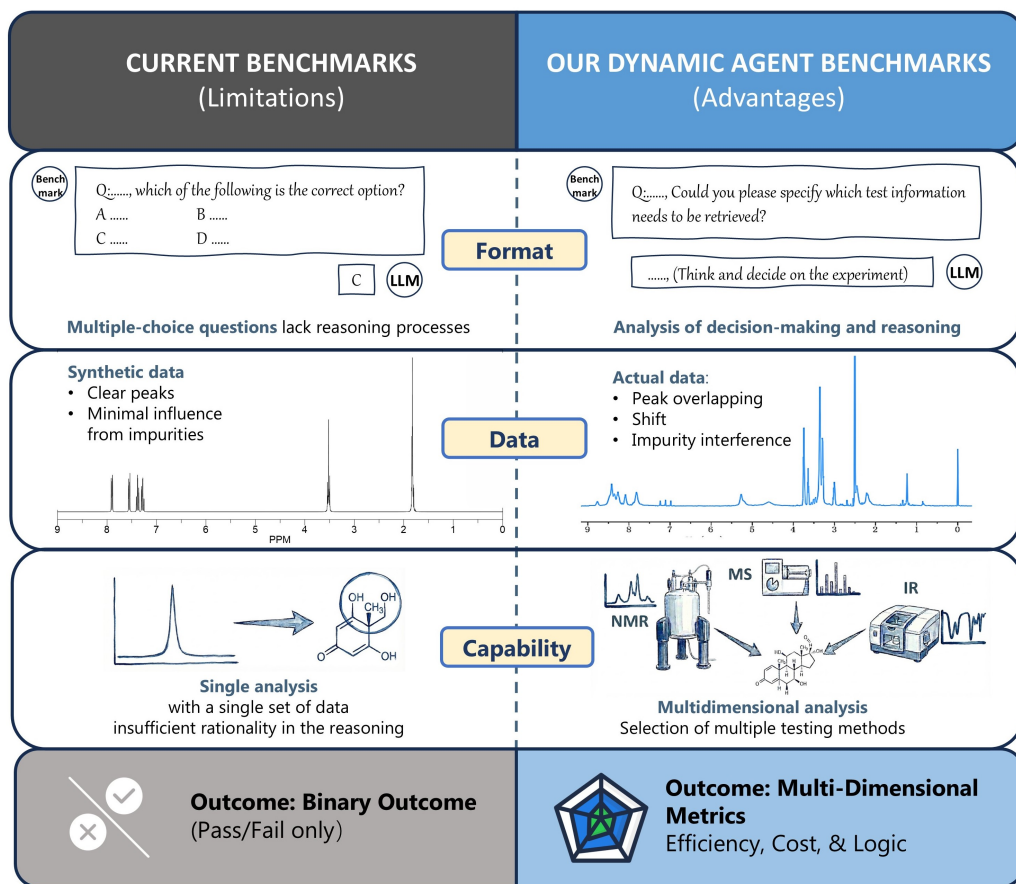


Figure 2: The characteristics of our benchmark.

mand. 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 reported peak overlap, incomplete descriptions, and missing evidence modalities.
- 2. Cognitive simulation.** A state-machine environment enforces a “Plan–Request–Reason” loop (Section 3.3), shaping the model’s behavior through iterative tool use and an explicit efficiency objective 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.

During construction, we found that LLM outputs frequently contained systematic errors—missing

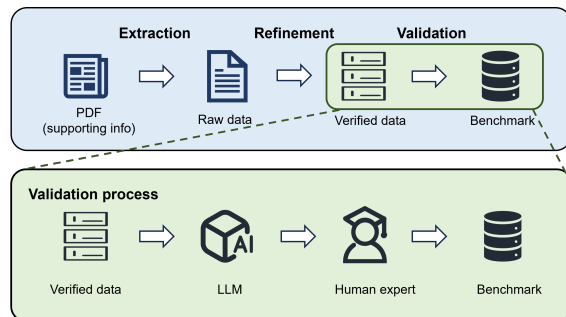


Figure 3: Data processing pipeline.

exchangeable protons, miscounting SMILES atoms, triggering false peak-deviation alarms, and confusing molecular masses with 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

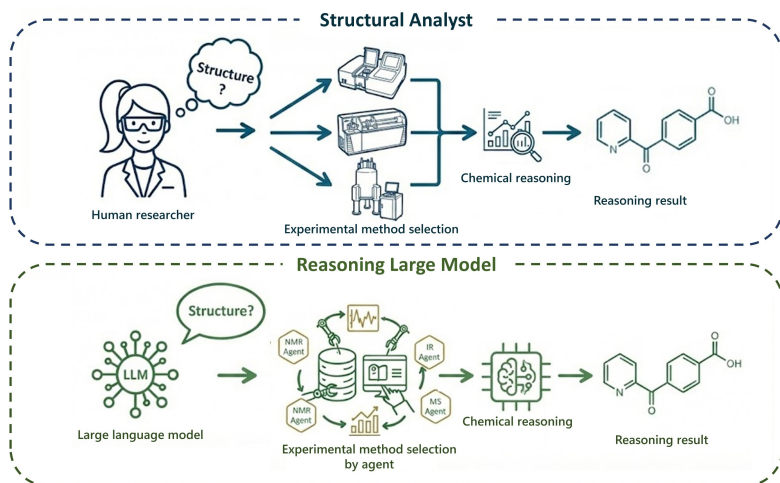


Figure 4: Comparison of chemical reasoning and decision-making between LLMs and human chemists in MolQuest.

counts). (3) **Expert review**—all 530 entries are reviewed by chemists through a visual interface, with automatically flagged records receiving targeted correction for subtle errors in stereochemistry, peak assignments, or molecular metadata.

Ground-truth fields, including compound names, SMILES, InChI/InChIKey, structure captions, and explicit textual identifiers, are excluded from model-visible inputs. Models receive only curated experimental fields and are never given the original Supporting Information documents. Here, “spectral data” denotes literature-reported characterization records, typically peak-list or descriptive text, rather than vendor instrument files, free-induction-decay data, or spectrum images. The records preserve reporting-level features such as unresolved multiplets, missing evidence modalities, and incomplete descriptions, but not instrument-level baseline noise.

This pipeline yielded 530 validated tasks spanning a broad molecular-weight range, providing a realistic foundation for evaluating dynamic scientific reasoning. All cases include molecular formula, molecular weight, and ^1H NMR records; 513 include ^{13}C NMR, 414 include HRMS, and 84 include IR records. Source journals and curation criteria are detailed in Appendix A.

3.3 Interactive Agent Framework Design

We implement MolQuest as a state machine where the LLM is cast as a “Senior Spectroscopist” operating under a tool-use efficiency objective, 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 prompt instructs the agent to minimize unnecessary experiments, and every tool call is counted as one unit of interaction cost for analysis. We do not assign heterogeneous monetary or laboratory-time prices to different tools. (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 stored characterization record and 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

Tool Name	Description
Check_Data	Return a list of evidence types available for the current molecule.
Measure_MW	Return molecular weight as a numeric value in Da.
Measure_Formula	Return the curated molecular formula as text.
Calculate_DBE	Return the calculated degree of unsaturation.
Get_1H_NMR	Return reported peak-list text: shifts, multiplicities, couplings, and integrations when available.
Get_13C_NMR	Return reported ^{13}C chemical-shift text and acquisition metadata when available.
Get_19F_NMR	Return reported ^{19}F NMR description text when available.
Get_31P_NMR	Return reported ^{31}P NMR description text when available.
Get_IR	Return reported IR absorption-peak text when available.
Get_HRMS	Return reported m/z , adduct, and calculated/-found exact-mass text.
Get_MS	Return reported m/z and fragment-peak description text.
Get_Melting_Point	Return reported melting-point value or range.
Get_TLC	Return reported R_f and eluent description text.
Get_Optical_Rotation	Return reported rotation, solvent, and concentration text.

Table 1: Full action space and the literature-reported data format returned by each MolQuest tool.

truth (Weininger, 1988). Both predictions and references are first canonicalized with RDKit, so alternative valid SMILES strings representing the same molecular graph are treated as equivalent. We use stereochemistry-aware canonical accuracy as the primary, conservative metric. As a complementary robustness measure, we also report stereochemistry-agnostic accuracy, computed after removing atom and bond stereochemical annotations before canonicalization and comparison. (2) **SMILES Validity Rate**: fraction of outputs parseable by RDKit into valid chemical graphs, reflecting syntactic grasp of molecular representation. (3) **Formula Conservation**: among RDKit-valid predictions, the fraction whose molecular formula matches the ground-truth formula. Reporting va-

lidity separately avoids counting a syntax failure twice while retaining a direct measure of mass-balance consistency. (4) **Average Similarity**: mean Tanimoto similarity (Morgan fingerprints) between predicted and true structures over all RDKit-valid predictions, including exact matches (Bajusz et al., 2015).

3.4.2 Probabilistic Reliability

Autonomous agents must assess their own certainty. Following the standard confidence-binning framework (Guo et al., 2017), we use a 10-bin root-mean-square calibration error (RMSCE) adapted to molecular prediction. It measures alignment between self-reported confidence and Tanimoto similarity over all 530 outputs; invalid SMILES receive similarity zero. A lower RMSCE indicates that reported confidence more closely tracks end-to-end structural fidelity—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 broad molecular-weight range 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). Although modest relative to automatically generated benchmarks, this scale reflects an explicit quality–quantity tradeoff: every case requires a validated ground-truth structure, aligned experimental evidence, expert review, and a query interface that excludes answer-bearing identifiers. We therefore treat these cases as a substantial expert-validated evaluation set for structure elucidation, while avoiding broad model-family claims beyond the tested configurations.

Model	Accuracy $\uparrow$		Validity $\uparrow$		Similarity $\uparrow$		Calibration $\downarrow$		Formula Cons. $\uparrow$	
	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 [†]	25.47	28.49	97.74	97.17	58.84	62.05	24.30	15.43	69.11	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	52.48	29.87	16.01
DeepSeek V3.2 [†]	16.60	20.38	71.13	56.79	53.55	64.12	27.21	23.07	64.46	81.73
Doubao-1.8	17.17	16.60	73.77	70.75	53.76	53.91	48.03	47.87	64.71	60.27
Doubao-2.0 Pro	24.91	24.72	92.45	87.92	57.44	58.06	36.94	36.52	70.20	71.89
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 [†]	7.92	16.04	65.47	68.87	39.82	50.72	48.40	43.90	25.36	50.14
GPT-5.5 High [‡]	42.64	52.64	95.66	96.23	73.53	78.54	14.67	9.98	95.86	97.65
Kimi K2 [†]	11.32	20.57	73.02	71.89	47.17	60.20	38.16	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 [†]	17.55	18.30	81.51	72.08	49.90	54.06	44.99	37.69	59.49	66.49
Qwen3.5 Plus [†]	19.25	28.49	91.32	91.13	52.36	60.56	34.64	35.36	63.02	72.88

[†] Models with dedicated "Thinking" mode enabled.

[‡] Reasoning via parameter: extended thinking (Claude) / reasoning effort (GPT).

Formula conservation is conditioned on RDKit-valid predictions; all other denominators are stated in Section 3.4.

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

4.1.2 Evaluated Models

We evaluated 16 state-of-the-art general-purpose LLM variants 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 Thinking (Pichai et al., 2025), Gemini 3 Flash (Doshi, 2025), GPT-5.5 (OpenAI, 2026), DeepSeek V3.1 (DeepSeek-AI, 2025), DeepSeek V3.2, DeepSeek V3.2 Thinking (DeepSeek-AI et al., 2025), Qwen3 Max, Qwen3 Max Thinking (Qwen Team, 2025; Yang et al., 2025), Qwen3.5 Plus (Qwen Team, 2026), Kimi K2 Thinking (Moonshot AI, 2025), GLM-4.7 (Zhipu AI, 2025), Doubao-1.8 (ByteDance Seed, 2026b), and Doubao-2.0-pro (ByteDance Seed, 2026a). Each variant was assessed under both Agent and Baseline settings, yielding 32 model-setting configurations. 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).

Metrics: stereochemistry-aware and

stereochemistry-agnostic Structure Accuracy, 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 (42.64%). A broad mid-tier (Doubao-2.0 Pro, Claude Opus 4.5, Qwen3.5 Plus Thinking, etc.) remains well below the leaders. Having a "Thinking" mode does not guarantee strong performance: GLM-4.7 Thinking reaches 7.92%, and DeepSeek V3.2 Thinking reaches 16.60%.

Stereochemistry robustness. Removing stereochemical annotations increases accuracy by an average of 1.73 percentage points in Agent mode and 1.47 points in Baseline mode across the 16 model variants. For example, Agent accuracy rises from 51.51% to 54.34% for Gemini 3 Flash and from 7.36% to 9.62% for DeepSeek V3.1. The modest gains across the performance range leave the overall ranking pattern and our main conclusions unchanged. Full stereochemistry-agnostic results are reported in Appendix D.

Consistency and calibration. GPT-5.5 High leads in formula conservation (95.86% Agent, 97.65% Baseline) and calibration (14.67% Agent,

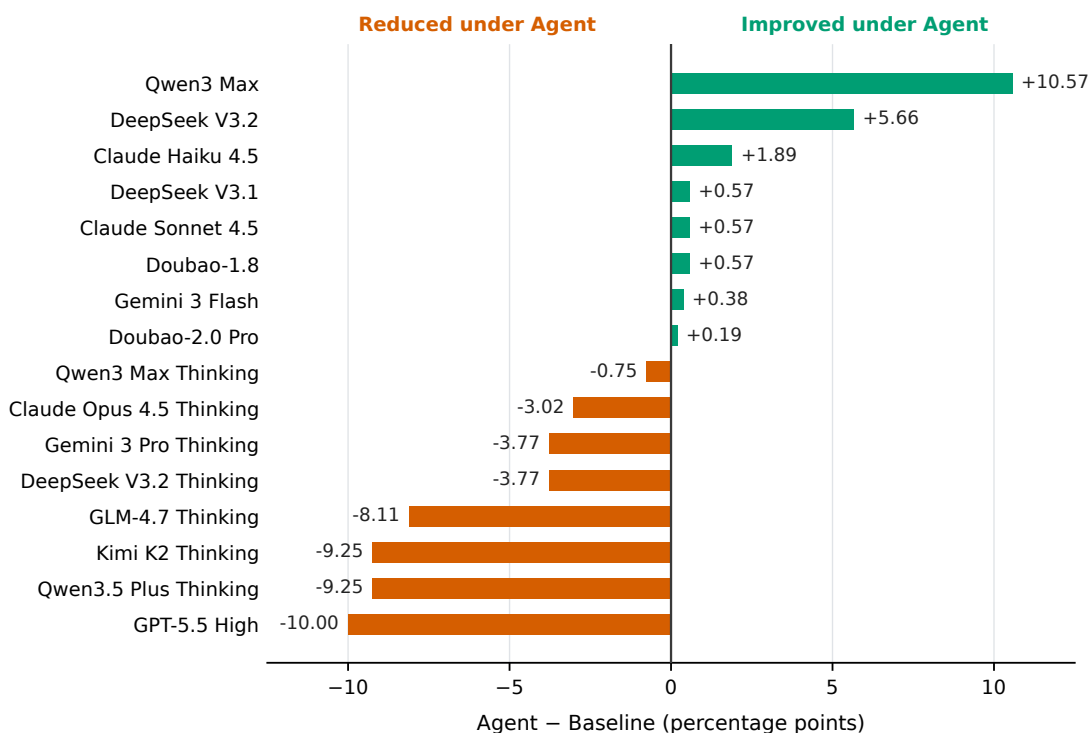


Figure 5: Per-model accuracy change from Baseline to Agent evaluation. Positive values indicate improvement under Agent evaluation, whereas negative values indicate reduced accuracy.

9.98% Baseline). Claude Opus 4.5 is also well calibrated. In contrast, DeepSeek V3.1 and Qwen3 Max frequently violate the supplied molecular formula (formula conservation 23.71% in Agent and 13.68% in Baseline). GPT-5.5, the Gemini 3 configurations, and Claude Opus/Sonnet maintain high validity rates, 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 (+10.57 points), DeepSeek V3.2 (+5.66 points), and Claude Haiku 4.5 (+1.89 points) gain accuracy, suggesting that sequential decomposition aids models that otherwise struggle with monolithic reasoning. Conversely, GLM-4.7 Thinking (-8.11 points), GPT-5.5 High (-10.00 points), and Kimi K2 Thinking (-9.25 points) lose accuracy, likely due to poor information valuation or premature termination. The Gemini 3 series changes less markedly (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 9.98% to 14.67%, and Claude Opus 4.5’s from 15.43% to 24.30%, indicating that se-

quential decision-making introduces additional uncertainty not fully reflected in confidence estimates.

4.4 Process-Level Analysis

Because MolQuest’s agentic setting is intended to measure information-seeking behavior, we retrospectively analyze the stored run records and tool-call trajectories on the same 530 cases. At the run level, we summarize LLM and planner calls, token use, duration, and call counts for correct and incorrect episodes. The results show that more interaction does not automatically improve performance: DeepSeek V3.1 averages 8.29 LLM calls but reaches only 7.36% accuracy, whereas Gemini 3 Flash averages 6.54 calls and reaches 51.51%. Full run-level statistics for all 16 Agent-mode model variants are provided in Appendix D.

For models with complete stored trajectories, we additionally compute average reasoning rounds and tool calls, repeated-call and unavailable-evidence-call rates, and accuracy points per average tool call (Acc./Call). Table 5 in Appendix D reports the complete results. DeepSeek V3.1 makes the most calls but has the lowest Acc./Call; Claude Sonnet 4.5 almost never requests unavailable evidence yet remains inaccurate; and Gemini 3 Flash is most efficient despite making some unavailable-evidence

calls. These diagnostics characterize interaction efficiency and evidence selection, but do not directly score the semantic quality of intermediate hypotheses or establish globally optimal tool sequences.

4.5 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 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* reconstructed the molecular framework and produced a final SMILES matching the stored (*S*, *S*) reference. Its reasoning log reveals a systematic assembly process:

Substructure Identification: It accurately identified the three *tert*-butyl environments from the 27H integration in the ^1H NMR spectrum, assigning them to the Boc-group and the two *tert*-butyl esters. **Backbone Assembly:** By jointly interpreting the reported $^1\text{H}/^{13}\text{C}$ chemical shifts, integrations, multiplicities, molecular formula, and HRMS record, the model established the linkage between the Alanine and Aspartic acid residues and maintained the peptide backbone. **Reference Agreement:** Its final SMILES also matched the stored (*S*, *S*) stereochemical annotation. This exact match should not be interpreted as independent proof that the available achiral measurements uniquely establish the absolute configuration.

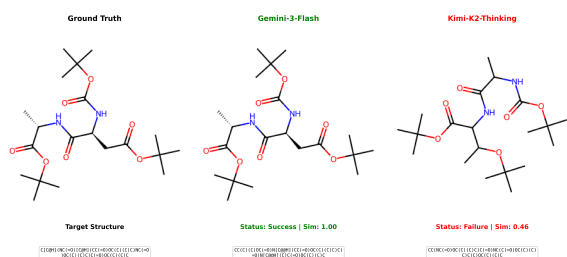


Figure 6: Structural elucidation comparison. (Left) Ground Truth: The validated chemical structure of the target peptide, Boc-Asp(OtBu)-Ala-OtBu. (Center) The structure predicted by Gemini-3-Flash, including agreement with the stored (*S*, *S*) annotation. (Right) The erroneous structure produced by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

In contrast, *Kimi K2 Thinking* failed to main-

tain 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.

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 tool-using—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 sequential decision-making task grounded in literature-reported NMR, MS, and related characterization records. 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 small-molecule elucidation on a set of general-purpose LLMs; performance patterns may differ in other scientific domains or with domain-specialized models. Moreover, MolQuest evaluates recovery of the literature-reported reference structure, but the exposed characterization records may be consistent with multiple constitutional or stereochemical candidates in some cases; the stereochemistry-agnostic metric partially addresses, but does not eliminate, this ambiguity. Second, although we report run-level and trace-level behavioral diagnostics, these proxies do not directly evaluate the semantic quality of intermediate hypotheses, causal evidence use, or the global optimality of experiment selection. Because the dynamic and static settings also differ in evidence ordering, context organization, and interaction structure, their performance gap should be interpreted as an effect of the broader interactive protocol rather than a causal estimate of planning ability alone. Third, our unit tool-call measure is only a

proxy for efficiency and does not model heterogeneous real-world laboratory costs in time, material, or money. 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 ambiguity-aware, 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 publicly available at <https://github.com/jiaodul307/molquest>. The repository includes structured data records derived from scientific literature.

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](#). Accessed: 2026-08-21.
- Anthropic. 2025b. [Claude Opus 4.5](#). Accessed: 2026-08-21.
- Anthropic. 2025c. [Claude Sonnet 4.5](#). Accessed: 2026-08-21.
- 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, et al. 2024. [Augmenting large language models with chemistry tools](#). *Nature Machine Intelligence*, 6(5):525–535.
- ByteDance Seed. 2026a. [Seed 2.0 official launch](#). Accessed: 2026-04-28.

- ByteDance Seed. 2026b. [Seed1.8 model card: Towards generalized real-world agency](#). *Preprint*, arXiv:2603.20633.
- Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. 2025. [ARC-AGI-2: A new challenge for frontier AI reasoning systems](#). *Preprint*, arXiv:2505.11831.
- DeepSeek-AI. 2025. [DeepSeek-V3.1 release](#). Accessed: 2026-08-21.
- DeepSeek-AI et al. 2025. [DeepSeek-V3.2: Pushing the frontier of open large language models](#). *Preprint*, arXiv:2512.02556.
- Tulsee Doshi. 2025. [Gemini 3 Flash: Frontier intelligence built for speed](#). Accessed: 2026-08-21.
- Mikhail E. Elyashberg, Antony Williams, and Kirill Blinov. 2011. *Contemporary Computer-Assisted Approaches to Molecular Structure Elucidation*. The Royal Society of Chemistry.
- 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, Haowei Zhang, et al. 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. 2025. [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, Nawaf Alampara, et al. 2024. [Are large language models superhuman chemists?](#) *arXiv preprint arXiv:2404.01475*.
- Moonshot AI. 2025. [Introducing Kimi K2 Thinking](#). Accessed: 2026-08-21.
- OpenAI. 2026. [GPT-5.5 model](#). Accessed: 2026-08-21.
- Zuzana Osifová, Ondřej Socha, and Martin Dračinský. 2024. [NMR-Challenge.com: Exploring the most common mistakes in NMR assignments](#). *Journal of Chemical Education*, 101(6):2561–2569.
- Long Phan, Alice Gatti, et al. 2025. [Humanity’s last exam](#). *Preprint*, arXiv:2501.14249.
- Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. 2025. [A new era of intelligence with Gemini 3](#). Accessed: 2026-08-21.
- Ernö Pretsch, Philippe Bühlmann, and Martin Badertscher. 2009. *Structure Determination of Organic Compounds: Tables of Spectral Data*, 4th edition. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Qwen Team. 2025. [Qwen3-Max: Just scale it](#). Accessed: 2026-08-21.
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#). Accessed: 2026-08-21.
- 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.
- Nicholas T. Runcie, Charlotte M. Deane, and Fergus Imrie. 2026. [Assessing the chemical intelligence of large language models](#). *Journal of Chemical Information and Modeling*, 66(1):216–227.
- Zhangde Song, Jieyu Lu, Yuanqi Du, et al. 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 Cohen, and Mark Gerstein. 2025. [ChemAgent: Self-updating library in large language models improves chemical reasoning](#). *Preprint*, arXiv:2501.06590.
- 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.

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

D. Additional Robustness and Process Analyses

- D.1 Stereochemistry-Agnostic Accuracy
- D.2 Run-Level Process Diagnostics
- D.3 Trace-Level Process Diagnostics

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 the 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 an independent ^1H NMR record; ^{13}C NMR and other characterization evidence are retained when reported in the source Supporting Information.

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 radius-2, 2048-bit Morgan fingerprints. We use 3D t-SNE with a fixed random seed for visualization and the same fingerprints for the pairwise Tanimoto analyses.

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.

Chemical-space embedding

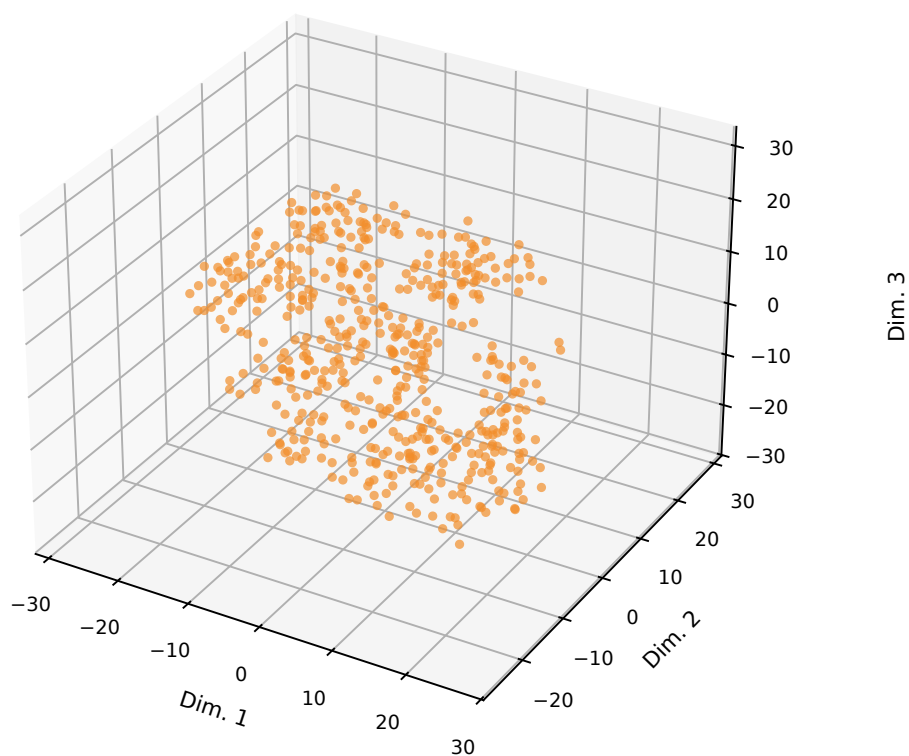


Figure 7: Chemical-space visualization of the benchmark molecular set using Morgan fingerprints and three-dimensional t-SNE (random seed 42).

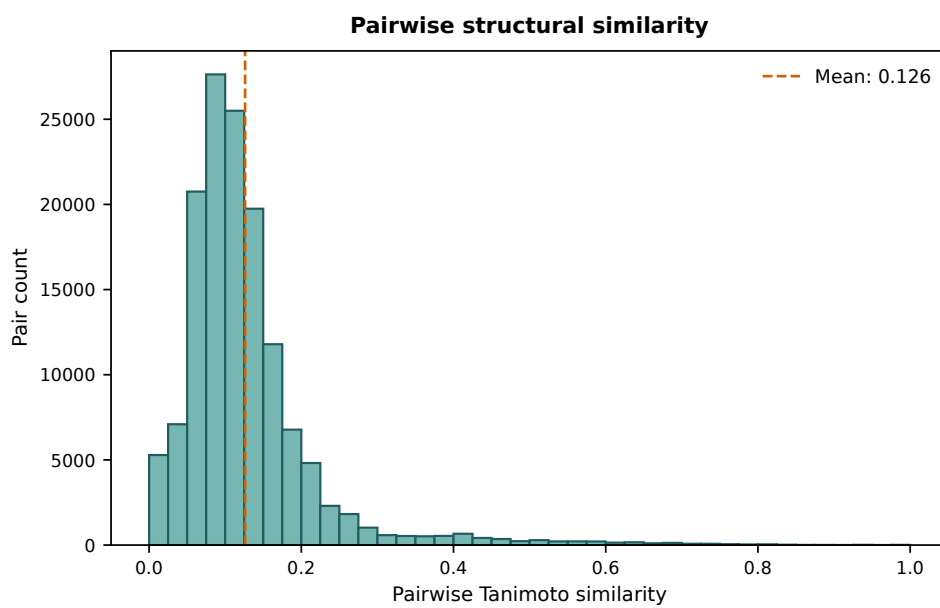


Figure 8: Distribution of pairwise Tanimoto similarities for the benchmark molecular set.

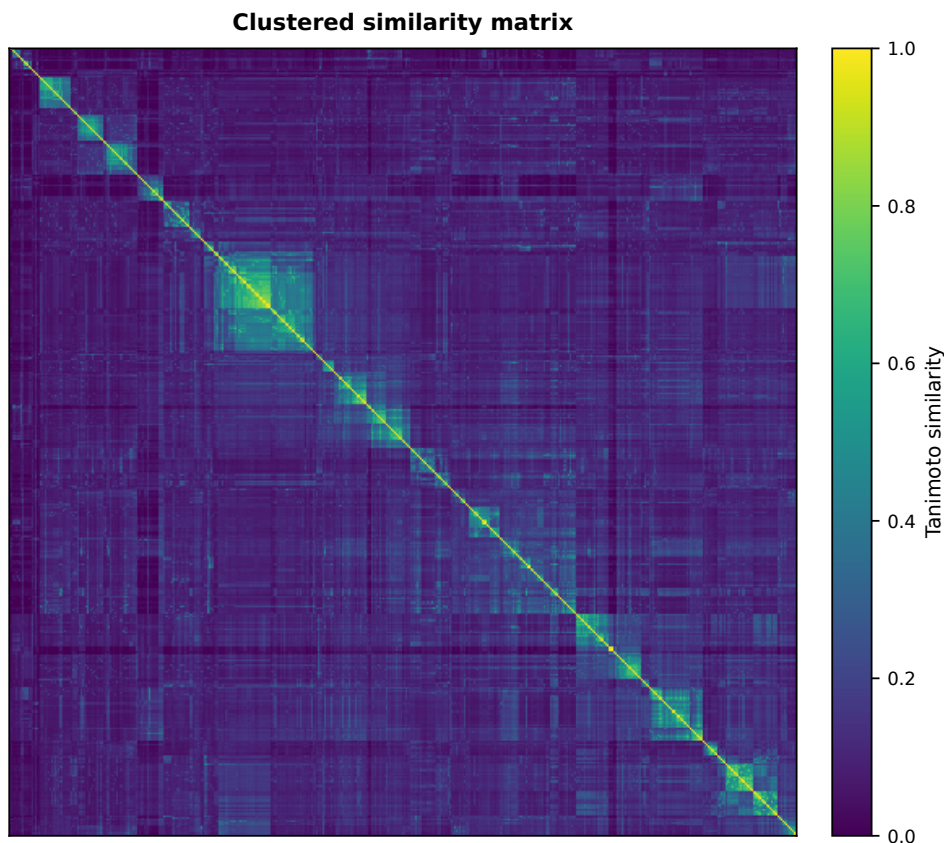


Figure 9: Hierarchically clustered pairwise Tanimoto-similarity matrix for the benchmark molecular set.

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

B.1 Multi-LLMs design

The pipeline uses three core agents:

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

- $s_t = \{messages, known_data\}$ represents the interaction state at time t ; separately, c_t records the cumulative number of tool calls as a unit interaction-cost proxy.
- $T(s_t, a_t) \rightarrow s_{t+1}$ 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}.
```

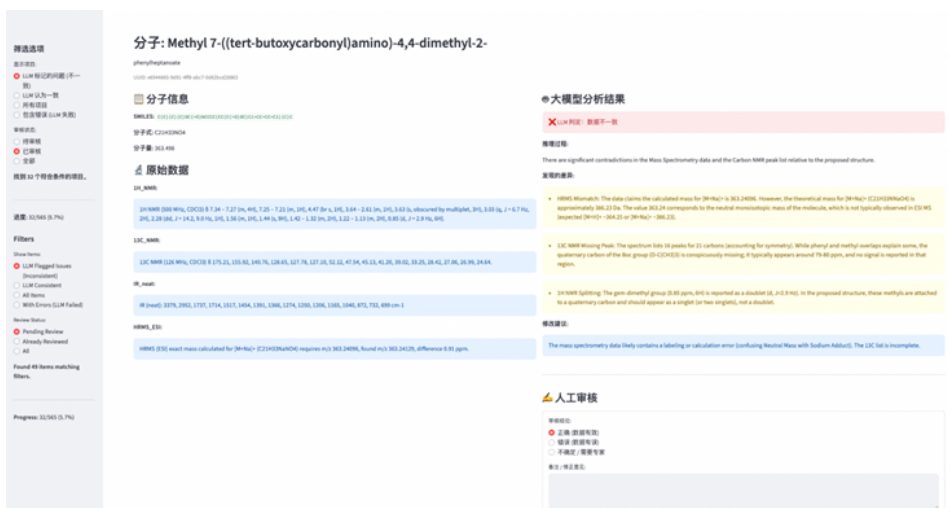


Figure 10: Screenshot of the human review interface showing the “Red Flag” system for flagged data points.

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.
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 experimental data available to the agent 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)

D Additional Robustness and Process Analyses

D.1 Stereochemistry-Agnostic Accuracy

Table 3 complements the stereochemistry-aware accuracy in Table 2. It is computed from the stored final predictions on the same 530 cases, without rerunning any model. Stereochemical annotations are removed from both the prediction and reference before RDKit canonicalization and exact comparison.

Model	Agent	Baseline
Claude Haiku 4.5	12.08	10.38
Claude Opus 4.5 Thinking	27.74	30.19
Claude Sonnet 4.5	20.19	19.81
DeepSeek V3.1	9.62	7.17
DeepSeek V3.2	12.64	5.85
DeepSeek V3.2 Thinking	17.92	21.70
Doubao-1.8	18.68	17.74
Doubao-2.0 Pro	26.04	25.47
Gemini 3 Flash	54.34	54.53
Gemini 3 Pro Thinking	50.38	54.34
GLM-4.7 Thinking	7.92	16.79
GPT-5.5 High	45.85	56.60
Kimi K2 Thinking	13.21	21.89
Qwen3 Max	16.79	5.28
Qwen3 Max Thinking	19.43	19.06
Qwen3.5 Plus Thinking	21.13	30.57

Table 3: Stereochemistry-agnostic accuracy (%) on the common 530 cases.

D.2 Run-Level Process Diagnostics

Table 4 summarizes interaction length for all 16 Agent-mode model variants. LLM calls include planner and final-answer calls where present. Correct/incorrect columns condition the average LLM-call count on the final stereochemistry-aware outcome.

D.3 Trace-Level Process Diagnostics

Table 5 reports the complete trace analysis for configurations with stored tool-call trajectories. Acc./Call divides accuracy percentage points by average tool calls; repeated and unavailable rates use all tool calls as the denominator.

Model	Avg. LLM	Avg. Planner	LLM (Correct)	LLM (Wrong)
Claude Haiku 4.5	5.39	5.39	5.03	5.44
Claude Opus 4.5 Thinking	5.96	4.96	5.36	6.17
Claude Sonnet 4.5	6.53	5.53	6.18	6.61
DeepSeek V3.1	8.29	7.29	7.38	8.37
DeepSeek V3.2	7.76	6.76	7.07	7.85
DeepSeek V3.2 Thinking	5.70	4.70	5.56	5.73
Doubao-1.8	5.50	4.50	4.75	5.66
Doubao-2.0 Pro	5.62	4.62	4.68	5.93
Gemini 3 Flash	6.54	5.54	6.11	7.00
Gemini 3 Pro Thinking	6.72	5.72	6.00	7.38
GLM-4.7 Thinking	7.14	6.14	6.36	7.20
GPT-5.5 High	5.43	4.43	4.80	5.90
Kimi K2 Thinking	5.84	4.84	5.90	5.83
Qwen3 Max	8.04	7.04	7.41	8.15
Qwen3 Max Thinking	5.62	5.62	6.13	5.52
Qwen3.5 Plus Thinking	6.75	5.75	6.53	6.80

Table 4: Run-level process diagnostics on the common 530 cases.

Model	Acc.	Rounds	Calls	Acc./Call	Repeated (%)	Unavailable (%)
Claude Haiku 4.5	11.51	3.21	6.71	1.71	0.03	0.37
Claude Sonnet 4.5	18.11	4.53	6.99	2.59	0.00	0.03
DeepSeek V3.1	7.36	6.29	7.24	1.02	0.39	0.05
DeepSeek V3.2	11.32	3.33	6.00	1.89	0.28	1.04
DeepSeek V3.2 Thinking	16.60	3.70	5.19	3.20	0.29	0.40
Gemini 3 Flash	51.51	3.90	6.42	8.03	0.03	8.12
Kimi K2 Thinking	11.32	2.36	4.76	2.38	4.08	7.61
Qwen3 Max	15.28	4.63	6.39	2.39	0.03	3.45

Table 5: Trace-level process diagnostics on the common 530 cases.