NMR Elucidation as an Agentic Search Problem, Not a Modeling Problem

Anonymous Authors[1]

Abstract

Structural elucidation from Nuclear Magnetic Resonance (NMR) data remains a fundamental bottleneck across chemistry, materials science, and biology. We demonstrate that an agentic AI system can perform this task at a level comparable to graduate-level chemistry students. Instead of training a model to directly map spectra to structures, we build a single autonomous agent, backed by a frozen LLM, that interacts with a curated environment with access to domain-specific processing tools, validation checks, tabulated chemical shifts, and instructions that outline the stepwise nature of a chemist's thinking process. On the Alberts dataset, our agent elucidates structures with a top-1 accuracy of 71%, comparable to the performance of graduate students at 66% top-1 accuracy. On the van Bramer and AstraZeneca datasets, our agent achieved 80% and 19% top-1 accuracy respectively, outperforming zero-shot end-to-end deep learning models which were trained on large datasets of simulated spectra. These results show that reframing NMR elucidation as an LLM-guided constrained search, rather than a modeling task, yields substantial gains and suggests a path toward multi-step orchestration frameworks that integrate a variety of tools, models, and domain knowledge to assist in automating spectroscopic analysis.

1. Introduction

Determining the structure of an unknown organic compound from routine spectra is a central task in analytical chemistry, underpinning drug discovery, natural product characterization, metabolomics, and quality control. However, the process is time consuming and requires years of experience. A chemist approaches the problem like a puzzle: given spectra and prior knowledge such as a molecular formula or a synthetic route, they derive constraints and reason over candidate molecules that satisfy them (Breitmayer, 2002).

The earliest attempt to automate structure elucidation was DENDRAL (Lindsay, 1980; Lederberg, 1987), one of the first expert systems designed to generate candidate chemical structures from raw Mass Spectrometry (MS) data. DENDRAL encoded a large knowledge base about MS, chemistry, and graph theory, which it leveraged to search for plausible structures and learn pruning rules that reduced the candidate space. The system's limitation was inherent to its design as all knowledge had to be hand-coded by domain experts as explicit rules, making the system brittle and difficult to generalize. The field subsequently shifted away from constraint-based search (Steinbeck, 2001; Funatsu & Sasaki, 1996; Munk, 1998; Howarth et al., 2020).

With the rise of machine learning, the dominant paradigm has become end-to-end prediction. A model is trained to directly map a spectrum to a molecular structure, typically represented as a SMILES string. This approach has seen substantial progress across all major spectroscopic modalities with models starting to achieve human level accuracy and revealing insights previously not possible (Jonas, 2019; Alberts et al., 2023; Hu et al., 2024; Stravs et al., 2022; Bohde et al., 2025; Wu et al., 2025; Alberts et al., 2025b; Priessner et al., 2026b; Alberts et al., 2025a; Wang et al., 2025; Jin et al., 2025). However, these direct approaches lack the interpretability and explicit reasoning of earlier expert systems.

The original puzzle-solving paradigm has received comparatively little attention in the modern machine learning era. Only recently have multimodal LLMs been evaluated on structure elucidation, framed explicitly as a puzzle-solving task requiring iterative hypothesis testing and integration of multifaceted spectroscopic data (Guo et al., 2024). Concurrent work on IR spectroscopy has explored multi-agent LLM frameworks that decompose the elucidation task into modular reasoning steps (Noh et al., 2025). Yet no existing work has applied a fully agentic, constraint-based system to structure elucidation from one-dimensional NMR spectra, one of the most accessible and routinely collected spectroscopic data in organic chemistry.

In this paper, we revisit the puzzle solving approach: rather

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

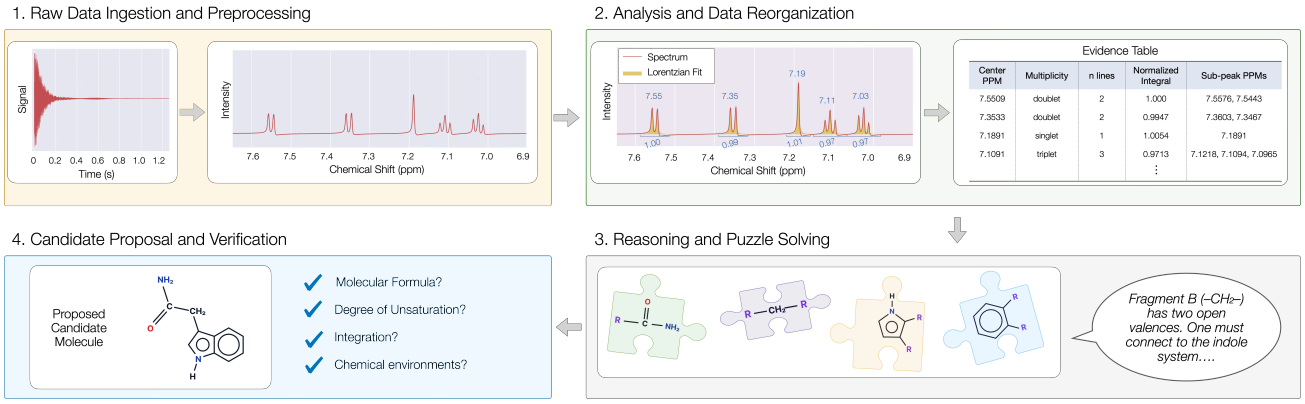

Figure 1. **Overview of NMR data pipeline for proposing molecular candidates**. The proposal pipeline is broken into four different phases: (1) Data Ingestion and processing - Ingesting the raw data and transforming it into human readable forms. (2) Analysis and data organization - Perform peak deconvolution and analysis of peaks, then generate data tables to use for reasoning about the data. (3) Reasoning - Reason on the extracted data, find correlations between peaks and what substructures may produce their signals. (4) Proposal and verification - Propose a molecule and verify that the molecule fits within the constraints of the data.

than learning a direct spectrum-to-structure mapping, we explore whether an agent can apply the puzzle-solving strategy on real world spectroscopic data. Our approach uses raw experimental NMR spectra together with the molecular formula, constructs structural constraints from the spectra, generates candidate molecules autonomously, and verifies them against the observed data without human intervention.

Our contributions are listed as follows:

- **Real experimental inputs.** We operate directly on raw NMR instrumentation files rather than tabulated peaks, reflecting the most realistic lab deployment setting.
- **Agentic framework and environment.** We evaluate modern agentic systems for NMR elucidation using only 1D NMR data and molecular formula as input. We introduce an agent-based harness for context management, structured reasoning, and preprocessing, together with preprocessing tools independent of commercial software.
- **Training-free approach.** Our approach does not involve training or the simulation of spectra datasets; instead it uses a frozen LLM as backbone and shifts the focus from assembling large datasets and leveraging compute to building an environment that distills the workflow of a human chemist.
- **Autonomous search.** The agent generates candidate molecules from scratch without prior seed structures and handles diverse solvent conditions.
- **Self-verification.** We designed validation tools that allow the agent to check candidate consistency against observed data without access to ground truth.

2. Methods

Our workflow starts with `FID` files (Free Induction Decay) which are preprocessed in a deterministic manner producing partially processed peak lists. For each task, the agent receives the same type of formats for inputs and outputs. **Inputs**: Partially processed peak lists. **Outputs**: SMILES strings of the top-10 candidates. The agent is able to adjust the partially processed peak lists as needed (remove noise, re-normalize, etc.) similar to how a human experimenter would during the reasoning process.

2.1. Data Preprocessing

Before the agent accesses the raw files, all raw FID and acquisition data are sanitized to remove any identifying metadata that could reveal the answer to the agent. The system can ingest raw NMR data collected using Bruker, Varian, and JEOL instruments. The agent is provided with an automated data processing pipeline whose exact output format of the peak lists depends on the type of NMR modality. For ^{1}H-NMR , the pipeline returns a list of the non-normalized peak positions, the normalized peak positions, the solvent peak position that was used for normalization, the integration ratios, and an initial attempt to normalize the peak integrations (the agent has autonomy to adjust the integrations or re-reference the spectrum to a different peak). For ^{13}C-NMR , the processing pipeline returns the non-normalized peak positions, the normalized peak positions, and the solvent peak position used for normalization. The steps of the NMR processing are further explained below in the following sections.

2.1.1. FID Preprocessing

The FID first must be ingested and preprocessed prior to conversion to the frequency domain. For all vendor data, this includes apodization using exponential multiplication line broadening to improve the signal-to-noise ratio, and zero-filling to increase frequency resolution. In the case of Bruker data, the digital filter is removed and any trailing bytes are also truncated to the nearest 4-byte boundary. Other than the trailing bytes truncation, all of these processing steps are performed using functionality provided by the `nmrglue` Python library (Helmus & Jaroniec, 2013).

2.1.2. Spectral Processing

A Fast Fourier Transform (FFT) is performed on the preprocessed FID to convert it from the time domain into the frequency domain, which is then converted to the ppm scale. This is then followed by phase and baseline correction. The phase correction first attempts to use the ACME autophase algorithm, as implemented in `nmrglue` (Chen et al., 2002). This phasing is then scored, and if it falls under a threshold score, a global grid search for the best phasing parameters is performed to maximize the phase score.

$$\text{phase score} = \frac{\sum_{i \in S,\ x_i > 0} x_i}{\sum_{i \in S,\ x_i < 0} |x_i|} \tag{1}$$

Where $x_i = \text{Re}(\phi(spectrum,\ p_0,\ p_1))$ is the real part of the spectrum phase (ϕ) being maximized after applying zero-order and first-order phase corrections, and S is the length of a signal mask, restricting the scoring to regions that actually contain peaks so noise doesn't distort the metric. The baseline correction is performed using an adaptive median-filter baseline correction. The solvent/TMS peak is then determined and the ppm scale is normalized to give the correct chemical shifts. Peak picking is then performed using the mean absolute deviation of the spectrum as a noise filter for SciPy's peak picking algorithm (Virtanen et al., 2020). Finally, for ^{1}H NMR, Lorentzian peaks are fitted to each peak following:

$$L(x) = \frac{A\gamma^2}{(x - x_0)^2 + \gamma^2} \tag{2}$$

where $L(x)$ is the Lorentzian peak, x_0 is the position, A is the amplitude, and γ is the half width at half maximum (HWHM). These peaks can then be used by the agent to begin building evidence tables for elucidation of the molecular species.

2.2. Agentic System

The agent is a single LLM driver built upon OpenAI AgentsSDK with access to a registry of curated tools but without coding capabilities and without browsing access. The prompt for the agent is assembled at runtime using a set of knowledge documents which provide all of the data input/output formats, tool documentation, data tables, etc. These knowledge documents provide no examples of NMR elucidation problems, only general chemistry knowledge about NMR elucidation and a description of how to approach structural elucidation. Crucially, the agent is given no seed structures and no access to the wider internet to ensure that any generated structures are due to its own reasoning, or information stored within the model. The agent only has access to the raw NMR spectra and the molecular formula from which it is tasked to predict the correct structure.

Once the data preprocessing is complete, the agent is given a dictionary called the `EvidenceTable`, which contains lists of peaks which are very roughly filtered, to remove clear baseline noise, and imprecisely processed (ppm, integration, etc.) The agent is allowed to reprocess the peaks if it feels it needs to, and adjust the table accordingly. From this table, the agent first generates an `EnvironmentTable` where it tries to classify different functional groups and substructures from the spectra. The agent then uses the `EnvironmentTable` to assemble different possible structures for proposal, using its internal knowledge to reason about the peak splitting, J-couplings, etc. searching for correlated peaks to justify the structure proposals.

2.3. Verification Rules

In order for the agent to be able to determine a stopping point for the proposed structure, it is given a set of tools to verify that the molecule fits the data. It must be emphasized that this is *not* verifying that the prediction matches the ground truth, but rather that the prediction fits the data constraints, ensuring that the model is predicting a structure that aligns with the data it is provided.

Molecular Formula: It checks that the formula fo the candidate proposed by the agent matches the input formula exactly. This is a hard constraint that the agent must satisfy.

Degree of Unsaturation (DoU): This quantity is calculated symbolically and counts the number of rings and double/triple bonds present in a molecule. The rule checks that the integer value of the DoU stored in the evidence table against the value computed from the molecular formula input. It ensures that when constructing the evidence table no drift in the DoU value occurs.

Integration Values (^{1}H-NMR): It checks that the observed ^{1}H integrals account for (approximately) every proton the formula asks for. The tolerance is ± 1 protons to account for noise, as well as a tolerance given for the *possible* number

of exchangeable protons given by the sum of number of nitrogen, oxygen, and sulfur atoms in the molecular formula.

Chemical Environment Counts: The last constraint we check the chemical environments which compares the number of distinct ^{1}H and ^{13}C environments in the evidence table against the number of expected environments based on the topological symmetry of the proposed molecule which is calculated using `RDKit`.

3. Experiments

3.1. Evaluation Datasets

All datasets used for evaluation are experimental, and we use the raw instrument files as input without any human preprocessing or intervention. We deliberately exclude peak-picked tabulated reports extracted from chemistry papers. Because these reports are human generated, they vary by practitioner and introduce biases that do not reflect real-world inference conditions.

Chemistry Education (Van Bramer & Bastin, 2023). Consists of 247 small organic molecular structures paired with their ^{1}H-NMR and ^{13}C-NMR NMR spectra `mnova` files that have been curated for educational purposes. It contains generally clean spectra, with clean peak separation for low complexity structures.

Alberts Dataset (Alberts et al., 2025a). As proposed, this dataset originally consists of 16 small organic molecules. One of the samples was found to be degraded and we removed this sample from our evaluation. As such, we evaluate on 15 molecules instead of 16, and re-adjusted the accuracies reported in (Alberts et al., 2025a) to reflect this reduction in the dataset. For each molecule ^{1}H-NMR , ^{13}C-NMR and IR spectra were measured. Human experts analyzed the spectra for each molecule, providing a strong human-performance baseline. As one of the molecules in the dataset is degraded and does not match the measured spectra, we exclude this molecule from our evaluation and scale the benchmark results accordingly.

AstraZeneca Dataset (Priessner et al., 2025; Rowlands et al., 2025). This dataset consists of 34 molecules for which the ^{1}H-NMR and ^{13}C-NMR were measured, among others. The molecules are chemically diverse, drug-like structures, selected to assess model performance under realistic laboratory conditions, also accounting for real-world factors such as impurities and spectral noise.

3.2. Evaluation Settings

Benchmarking NMR Elucidation methods under identical conditions is challenging. There are several aspects that limit direct comparison. First, methods use different combinations of spectroscopic modalities and different input formats for training and inference. Second, because of this, preprocessing of samples is not consistent across methods; not all methods can be evaluated on any dataset or even be evaluated on the full dataset, as some samples can be invalid for certain methods. Finally, reported accuracy depends on how SMILES stereochemistry handled. If removed completely, accuracy can be higher than it should be, because enantiomers and diastereomers may be counted as correct. However, depending on the modalities used, it might not be physically possible to differentiate between enantiomers, which over-penalizes a method on a task it cannot resolve.

Specialized Models. We benchmark against different types of deep learning models, multi-modal, language models, such as Alberts et al. (2025a)'s model and the MultiModalSpectralTransformer (MMST) (Priessner et al., 2026a). Alberts et al. (2025a)'s model is trained on the molecular formula, ^{1}H-NMR , ^{13}C-NMR , and IR spectra to predict SMILES (Weininger, 1988) as output. This model is pretrained on simulated data before being finetuned on experimental data. On the other hand, MMST is trained exclusively on simulated data and follows the same input output format as Alberts et al. (2025a)'s model. We also include a diffusion-based model ChefNMR (Xiong et al., 2025) trained on the molecular formula, ^{1}H-NMR , and ^{13}C-NMR which predict the 3D atomic positions of a molecule.

Agent Baselines. We instantiate a coding agent, Claude Code (`claude-opus-4.7`) and Codex (`gpt-5.5`), in an isolated Docker container with access to a Python interpreter but without internet browsing or our domain-specific tools. It receives the same inputs as our system, the raw spectral data (molecular formula, ^{1}H-NMR and ^{13}C-NMR `fid` files), together with a comprehensive prompt that specified the task and the reasoning steps, similar to our agent's prompt. This setup allows for comparison between the LLM's out of the box capabilities and the environment that we constructed for the task of NMR elucidation. We sandboxed both agents, baseline and ours, to avoid a leakage of the answer. Due to resource constraints, we were able to perform three independent runs for Codex for all the datasets, but only one for Claude Code, thus we report mean and standard deviation for Codex and the mean of a single full run for Claude Code.

Metrics We report top-k accuracy based on exact matches between the canonical SMILES strings of the candidates proposed by the agent and the ground truth. We retain all stereochemical information in our SMILES, so an incorrect stereochemical prediction results *always* in a zero hit. This is in contrast with some of the models we benchmark against, such as Xiong et al. (2025), where stereochemistry

Table 1. **Benchmark results on experimental datasets for NMR elucidation.** We evaluate molecular structure prediction from experimental spectra and molecular formula on three datasets: Chemistry Education (van Bramer), Alberts, and AstraZeneca. We report top-k accuracy and Tanimoto similarity to the ground-truth structure (higher is better). Column N indicates the total number of samples used by a given method for metrics calculations. Results for prior work and human evaluations are extracted from the original sources. See notes below for differences in evaluation protocols.

		Top–1		Top–2		Top–5		
	METHOD	Acc.% ↑	Tan. ↑	Acc.% ↑	Tan. ↑	Acc.% ↑	Tan. ↑	N
van Bramer (Van Bramer & Bastin, 2023)								
Model	Alberts et. al. (zero shot)†	50.30	N/A	64.00	N/A	67.10	N/A	171
	Alberts et. al. (finetuned 5x CV)†	69.10±11.20	N/A	91.50±5.60	N/A	N/A	N/A	
	Alberts et. al. (finetuned 5x CV + 33k unpaired exp. samples)†	**96.20±7.50**	N/A	**98.80±2.50**	N/A	N/A	N/A	
	ChefNMR L (zero shot)†,*	56.00	∼0.68	∼65	∼0.75	∼70	∼0.80	238
Agent	Ours (`gpt-5.4`)	80.42	0.88	86.67	0.58	90.00	0.38	236
	Ours (`kimi-k2.6`)	80.87	0.90	84.35	0.56	90.00	0.39	
	Ours (`qwen3.5-122b`)	55.23	0.70	67.78	0.50	76.15	0.40	
	Baseline (`Codex 5.5`)	60.70±2.50	0.72±0.02	67.60±1.20	0.78±0.01	78.70±1.20	0.88±0.01	236
	Baseline (`CC Opus-4.7`)	18.30	0.23	24.40	0.15	26.80	0.10	
Alberts (Alberts et al., 2025a)								
	Grad Students al.†	66.67±3.06	N/A	82.21±3.06	N/A	84.45±3.95	N/A	15
Model	Alberts et. al.†	69.33±3.06	N/A	77.34±3.06	N/A	81.33±3.95	N/A	15
Agent	Ours (`gpt-5.4`)	66.67±5.44	0.76±0.04	71.11±3.14	0.60±0.01	71.11±3.14	0.43±0.01	15
	Ours (`kimi-k2.6`)	**71.11±6.29**	**0.80±0.04**	**75.56±3.14**	**0.61±0.02**	**77.78±3.14**	**0.43±0.02**	
	Ours (`qwen3.5-122b`)	37.78±8.31	0.56±0.06	37.78±12.57	0.45±0.05	44.44±13.70	0.36±0.03	
	Baseline (`Codex 5.5`)	31.10±10.20	0.47±0.09	43.80±9.30	0.52±0.11	48.90±7.70	0.60±0.06	15
	Baseline (`CC Opus-4.7`)	26.70	0.39	26.70	0.32	26.70	0.25	
AstraZeneca (Priessner et al., 2026b)								
	Alberts et. al.†	14.70±4.20	N/A	18.80±2.40	N/A	22.90±2.20	N/A	34
Model	MMST (base model)†	0.00	N/A	0.00	N/A	3.00	N/A	34
	MMST (trained on analogues)†	12.00	N/A	38.00	N/A	44.00	N/A	
	MMST (trained on test)†	31.00	N/A	56.00	N/A	81.00	N/A	
Agent	Ours (`gpt-5.4`)	7.84±1.39	0.38±0.01	14.71±0.00	0.35±0.00	16.67±1.39	0.31±0.01	34
	Ours (`kimi-k2.6`)	**19.38±2.51**	**0.66±0.07**	**24.66±4.79**	**0.57±0.06**	**27.85±6.83**	**0.51±0.05**	
	Ours (`qwen3.5-122b`)	2.13±1.52	0.22±0.01	3.11±2.41	0.21±0.01	3.11±2.41	0.20±0.00	
	Baseline (`Codex 5.5`)	2.90±2.90	0.28±0.02	6.40±7.00	0.31±0.02	10.00±7.40	0.34±0.02	34
	Baseline (`CC Opus-4.7`)	11.80	0.33	11.80	0.28	11.80	0.26	

† Reported performance from source. * Reports accuracy removing stereochemical features.

is removed from the SMILES strings. The top-k Tanimoto similarity is calculated using Morgan fingerprints (radius 2, 2048 bits) between the candidates and the ground truth. Per-dataset metrics are averaged across all tasks and reported as mean and standard deviation over independent runs.

4. Results

4.1. Benchmarks

The results on the Alberts, AstraZeneca, and Chemistry Education datasets are shown in Table 1. These benchmarks cover complementary settings: expert-evaluation, drug-like molecules measured under realistic laboratory conditions, and cleaner educational spectra. We compare our agent with specialized NMR elucidation models, human experts where available, and general purpose coding-agent baselines. Because these methods differ in inputs, preprocessing requirements and metric reporting, the results should be interpreted together with the caveats described in Section 3.2. For the Chemistry Education evaluations, ChefNMR reports evaluations on 238/247 tasks due to inconsistencies or wrongly paired data, our Agent was able to complete 236/247 tasks, and Alberts et al. (2025a) evaluates on 171/247 in order to have paired 1D NMR and IR samples. For the Alberts and

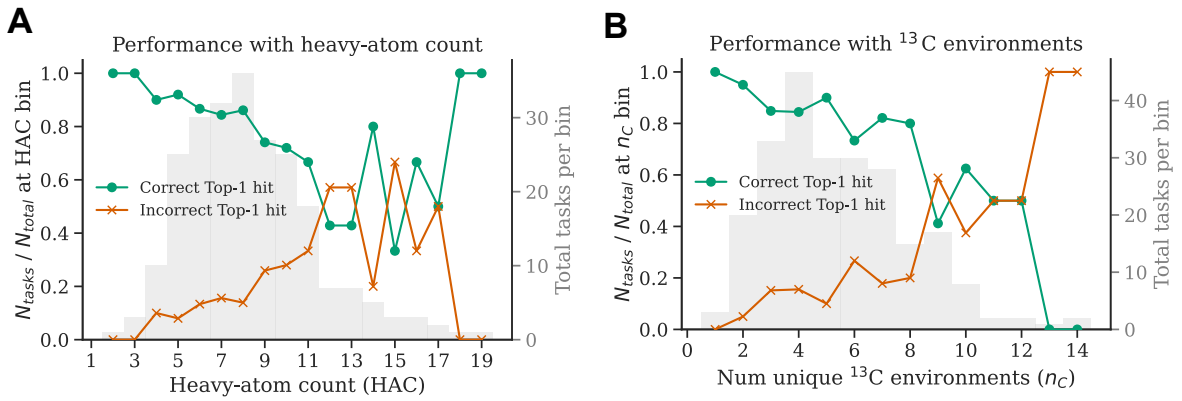

Figure 2. **Stratified evaluations by selected molecular properties.** Performance on the Chemistry Education dataset for `kimi-k2.6` stratified by (A) Heavy-atom count and (B) Number of unique ^{13}C environments. Green lines show the fraction of tasks within a bin where the agent was correct, orange lines show the fraction where the agent was wrong. The faint gray histogram shows the total number of tasks per bin. Performance tends to decrease with the number of unique carbon environments, but so does the absolute number of tasks, making the extraction of conclusions more difficult.

AstraZeneca datasets, all models are evaluated in the totality of the samples.

Across all datasets, our agent performs competitively with, and sometimes outperforms, models that have been trained on large amounts of simulated spectra. Regarding agentic baselines, the fact that a general purpose coding-agent consistently underperforms our agent indicates that the performance gains of our agent not only come from the underlying LLM, but from the task-specific environment that was provided. On the Alberts dataset, our agent has comparable performance to a graduate-level chemistry student. Irregardless of the method, performance degrades when testing on the more complex AstraZeneca dataset.

4.2. Stratified evaluations

In this section, we characterize stratified performance across different axes that make NMR elucidation more complex. For instance, it is common for models to degrade in performance with molecular weight, as shown in Alberts et al. (2024a) or when exposed to real-world molecules of interest, as shown in Alberts et al. (2024b); Xiong et al. (2025). Additionally, as the number of unique chemical environments increases in the molecule, the complexity of the resulting spectrum increases and parts of the spectrum may become more difficult to interpret. One clear example of this is multiplet resolvability as peak overlap becomes more prevalent (Aguilar et al., 2010; Halabalaki et al., 2014). Figure 2 shows the performance of the agent on the van Bramer dataset changes with heavy-atom count and the number of unique ^{13}C environments, which can be regarded as a proxy for the molecular complexity and spectral complexity respectively. Unsurprisingly, as the complexity increases, the task of elucidating the structure becomes more difficult for the agent. For higher heavy atom counts and unique ^{13}C environments the trends become relatively noisy. This can be attributed to the small number of molecules and spectra that fall in this range and is clearly an area which must be investigated in future studies.

4.3. Case studies

Figure 3 shows representative examples of successes and failures of our agent. We observe that the agent is able to differentiate between isomers, correctly interpreting the subtle differences between the spectra even in cases where isomerization would not lead to large changes in the spectra. Three failure modes are prominent: near-misses, structural failures and stereochemistry-related failures. After analyzing the reasoning traces of the agent, we observed that structural failures are relatively varied. For instance, the model will often narrow in on a certain region of the chemical space, and struggles to re-consider earlier assumptions. When assigning spectral peaks as artifacts, noise, solvent or signal is more difficult, the agent can struggle to find a molecular candidate that satisfies the chemical formula, while also having accurate integrations and number of chemical environments, entering into a long reasoning loop that can derail. These failures can be connected to the robustness of the Agent's predictions, in the case of `kimi-k2.6` we observe that its predictions on the Alberts dataset are more robust than in the AstraZeneca dataset. For the Alberts dataset, the agent is correct 3/3 times in approximately 67% of the tasks and has mixed predictions for 13% of the tasks (correct 1/3 or 2/3 times). In contrast, the Agent's correct predictions are significantly less robust for AstraZeneca tasks, with 13% of the tasks being correctly predicted 3/3 times, with mixed predictions amounting to 16%. See more examples in Supplementary Materials.

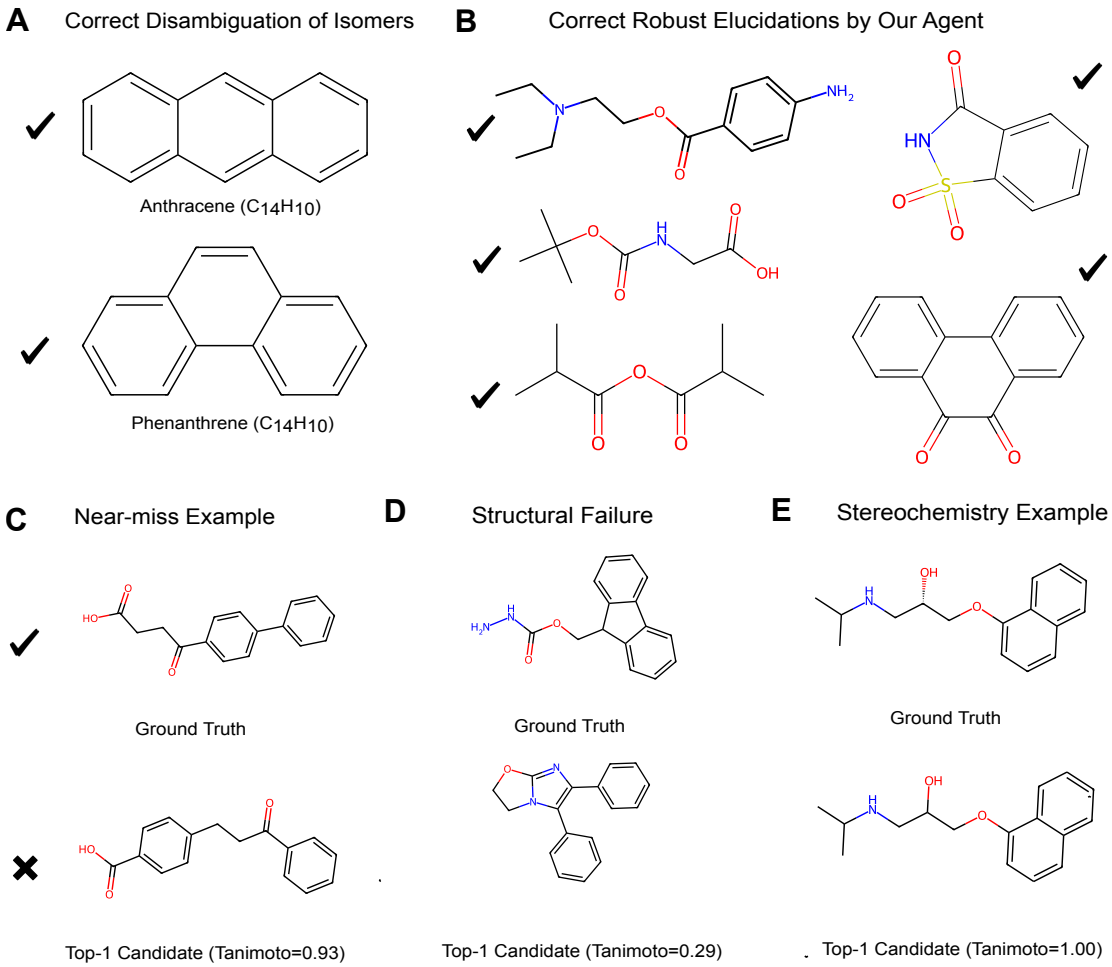

Figure 3. **Case studies of top-1 predictions by our agent.** All examples are predictions from `kimi-k2.6`. Top: examples that were predicted robustly, with the agent returning the same answer in all or nearly all queries. (A) The agent correctly disambiguates two structurally similar isomers with the same molecular formula and never confuses them with each other. (B) Across a set of varied molecules, the agent predicts the correct structure in every trial. Bottom: representative failure modes. (C) A near-miss prediction with high structural similarity to the ground truth. (D) A structural failure with low similarity to the ground truth. (E) A stereochemical error, where the predicted connectivity is correct but the stereochemistry differs.

5. Discussion

This work reframes NMR elucidation for small organic molecules as an agentic, constraint-driven search problem rather than an end-to-end model that maps spectra to structures. Our results support this hypothesis: an agent with access to a curated environment can achieve competitive performance against models trained on large simulated datasets. It achieves a 80% top-1 accuracy on the Chemistry Education dataset, performs on par with graduate-level chemistry students on Alberts, and reaches 19% top-1 accuracy on the more challenging AstraZeneca dataset, which contains complex drug-like molecules measured under laboratory conditions. The failure cases show where the current approach remains limited. Some errors occur when the agent commits too early to a molecular family and cannot revise it later, especially when spectra contain noise, impurities, artifacts, or overlapping peaks. As molecular complexity increases, the reasoning traces become longer and harder to navigate. Robustness across repeated independent runs is another important limitation: some tasks converge consistently to the same answer, while others do not. Understanding this variability and quantifying uncertainty are important future directions.

A useful aspect of the agentic approach is that it can scale by improving the environment around the model: more reliable preprocessing, stronger verification checks, better search strategies, and additional modalities such as IR, mass spectrometry, EPR, UV-vis, microwave, Raman, and 2D NMR. Overall, these results suggest that agentic puzzle

solving is a promising framework for small-molecule NMR structure elucidation. Going forward, we are extending the framework into a multi-agent system that can integrate additional modalities while improving context management. This framework should be viewed as a supporting tool for chemists, and not as a substitution; it can generate candidates, expose its reasoning, and help prioritize structures for further experimental validation.

References

Aguilar, J. A., Faulkner, S., Nilsson, M., and Morris, G. A. Pure shift 1h nmr: A resolution of the resolution problem? *Angewandte Chemie International Edition*, 49(23):3901–3903, 2010.

Alberts, M., Zipoli, F., and Vaucher, A. C. Learning the language of nmr: Structure elucidation from nmr spectra using transformer models. *ChemRxiv*, 2023 (0814), 2023. doi: 10.26434/chemrxiv-2023-8wxcz. URL https://chemrxiv.org/doi/abs/10.26434/chemrxiv-2023-8wxcz.

Alberts, M., Laino, T., and Vaucher, A. C. Leveraging infrared spectroscopy for automated structure elucidation. *Commun. Chem.*, 7(1):268, nov 2024a.

Alberts, M., Schilter, O., Zipoli, F., Hartrampf, N., and Laino, T. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. In *NeurIPS 2024 Datasets and Benchmarks Track*, 2024b. URL https://openreview.net/forum?id=xjxqWYyTfR.

Alberts, M., Hartrampf, N., and Laino, T. Automated Structure Elucidation at Human-Level Accuracy via a Multimodal Multitask Language Model, 2025a.

Alberts, M., Zipoli, F., and Laino, T. Setting new benchmarks in ai-driven infrared structure elucidation. *Digital Discovery*, 4:1936–1943, 2025b. doi: 10.1039/D5DD00131E. URL http://dx.doi.org/10.1039/D5DD00131E.

Bohde, M., Manjrekar, M., Wang, R., Ji, S., and Coley, C. W. DiffMS: Diffusion Generation of Molecules Conditioned on Mass Spectra, 2025. arXiv:2502.09571.

Breitmayer, E. *Structure Elucidation by NMR in Organic Chemistry: A Practical Guide*. John Wiley & Sons, Ltd, 2002.

Chen, L., Weng, Z., Goh, L., and Garland, M. An efficient algorithm for automatic phase correction of nmr spectra based on entropy minimization. *Journal of Magnetic Resonance*, 158(1):164–168, 2002.

Funatsu, K. and Sasaki, S.-i. Recent Advances in the Automated Structure Elucidation System, CHEMICS. Utilization of Two-Dimensional NMR Spectral Information and Development of Peripheral Functions for Examination of Candidates. *Journal of Chemical Information and Computer Sciences*, 36(2):190–204, 1996.

Guo, K., Nan, B., Zhou, Y., Guo, T., Guo, Z., Surve, M., Liang, Z., Chawla, N. V., Wiest, O., and Zhang, X. 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*, 2024.

Halabalaki, M., Vougogiannopoulou, K., Mikros, E., and Skaltsounis, A. L. Recent advances and new strategies in the nmr-based identification of natural products. *Current Opinion in Biotechnology*, 25:1–7, 2014.

Helmus, J. J. and Jaroniec, C. P. Nmrglue: an open source Python package for the analysis of multidimensional NMR data. *Journal of Biomolecular NMR*, 55(4):355–367, 2013.

Howarth, A., Ermanis, K., and Goodman, J. M. Dp4-ai automated nmr data analysis: straight from spectrometer to structure. *Chemical Science*, 11:4351–4359, 2020. doi: 10.1039/D0SC00442A. URL https://doi.org/10.1039/D0SC00442A.

Hu, F., Chen, M. S., Rotskoff, G. M., Kanan, M. W., and Markland, T. E. Accurate and efficient structure elucidation from routine one-dimensional nmr spectra using multitask machine learning. *ACS Central Science*, 10(11):2162–2170, 2024. doi: 10.1021/acscentsci.4c01132. URL https://doi.org/10.1021/acscentsci.4c01132.

Jin, Y., Wang, J.-J., Xu, F., Ji, X., Gao, Z., Zhang, L., Ke, G., Zhu, R., and E, W. Nmr-solver: Automated structure elucidation via large-scale spectral matching and physics-guided fragment optimization. *arXiv preprint arXiv:2509.00640*, 2025. URL https://arxiv.org/abs/2509.00640.

Jonas, E. Deep imitation learning for molecular inverse problems. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Lederberg, J. How DENDRAL was conceived and born. In *Proceedings of ACM conference on History of medical informatics*, pp. 5–19. Association for Computing Machinery, 1987.

Lindsay, R. K. Applications of artificial intelligence for organic chemistry: the dendral project. *McGraw-Hill Companies*, 1980.

Munk, M. E. Computer-Based Structure Determination: Then and Now. *Journal of Chemical Information and Computer Sciences*, 38(6):997–1009, 1998.

Noh, H., Lee, N., Na, G. S., Kim, K., and Park, C. Ir-agent: Expert-inspired llm agents for structure elucidation from infrared spectra, 2025. URL https://arxiv.org/abs/2508.16112.

Priessner, M., Lewis, R., Janet, J. P., Lemurell, I., Johansson, M., Goodman, J., and Tomberg, A. Advancing structure elucidation with a flexible multi-spectral ai model. *ChemRxiv*, 2025(0812), 2025. doi: 10.26434/chemrxiv-2024-zmmnw-v2. URL https://chemrxiv.org/doi/abs/10.26434/chemrxiv-2024-zmmnw-v2.

Priessner, M., Lewis, R. J., Johansson, M. J., Goodman, J. M., Janet, J. P., and Tomberg, A. Enhancing molecular structure elucidation with reasoning-capable llms. *Digital Discovery*, 2026a. doi: 10.1039/d5dd00359h. URL https://doi.org/10.1039/d5dd00359h. Open Access.

Priessner, M., Lewis, R. J., Lemurell, I., Johansson, M. J., Goodman, J. M., Janet, J. P., and Tomberg, A. Advancing structure elucidation with a flexible multi-spectral ai model. *Angewandte Chemie International Edition*, 65(2): e17611, 2026b. doi: 10.1002/anie.202517611.

Rowlands, J. B., Jonsson, L., Goodman, J. M., Howe, P. W. A., Czechtizky, W., Leek, T., and Lewis, R. J. Towards automatically verifying chemical structures: the powerful combination of 1h nmr and ir spectroscopy. *Chem. Sci.*, 16:21590–21599, 2025. doi: 10.1039/D5SC06866E. URL http://dx.doi.org/10.1039/D5SC06866E.

Steinbeck, C. SENECA: A Platform-Independent, Distributed, and Parallel System for Computer-Assisted Structure Elucidation in Organic Chemistry. *Journal of Chemical Information and Computer Sciences*, 41(6): 1500–1507, 2001.

Stravs, M. A., Dührkop, K., Böcker, S., and Zamboni, N. MSNovelist: de novo structure generation from mass spectra. *Nature Methods*, 19(7):865–870, 2022.

Van Bramer, S. E. and Bastin, L. D. Spectroscopy data for undergraduate teaching. *Journal of Chemical Education*, 100(10):3897–3902, 2023. doi: 10.1021/acs.jchemed.3c00046. URL https://doi.org/10.1021/acs.jchemed.3c00046.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Wang, X. et al. Nmrmind: A transformer-based model enabling structure elucidation from multidimensional nmr. *Analytical Chemistry*, 2025. doi: 10.1021/acs.analchem.5c03783. URL https://doi.org/10.1021/acs.analchem.5c03783.

Weininger, D. 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, 1988.

Wu, W., Leonardis, A., Jiao, J., Jiang, J., and Chen, L. Transformer-Based Models for Predicting Molecular Structures from Infrared Spectra Using Patch-Based Self-Attention. *The Journal of Physical Chemistry A*, 129 (8):2077–2085, 2025.

Xiong, Z., Zhang, Y., Alauddin, F., Cheng, C. X., An, J. S., Seyedsayamdost, M. R., and Zhong, E. D. Atomic diffusion models for small molecule structure elucidation from nmr spectra, 2025. URL https://arxiv.org/abs/2512.03127.

Supplementary Material

See extensive examples of correct, incorrect, near misses and mixed predictions by our agent. Combination of all three datasets, with Alberts and AstraZeneca being ran for three independent runs.

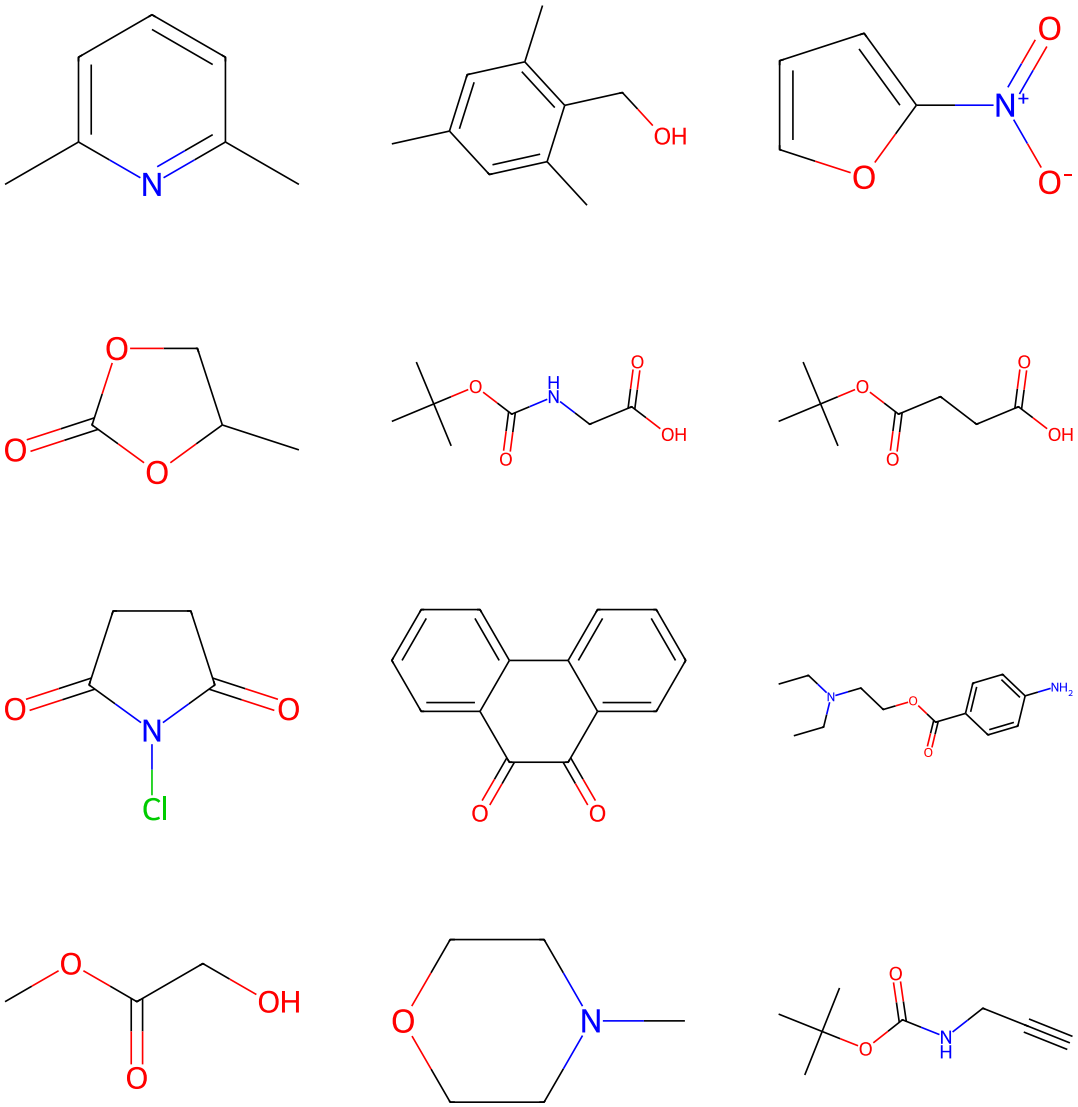

Figure 4. **Additional correct predictions by our agent.** Examples of correct top-1 predictions from `kimi-k2.6`.

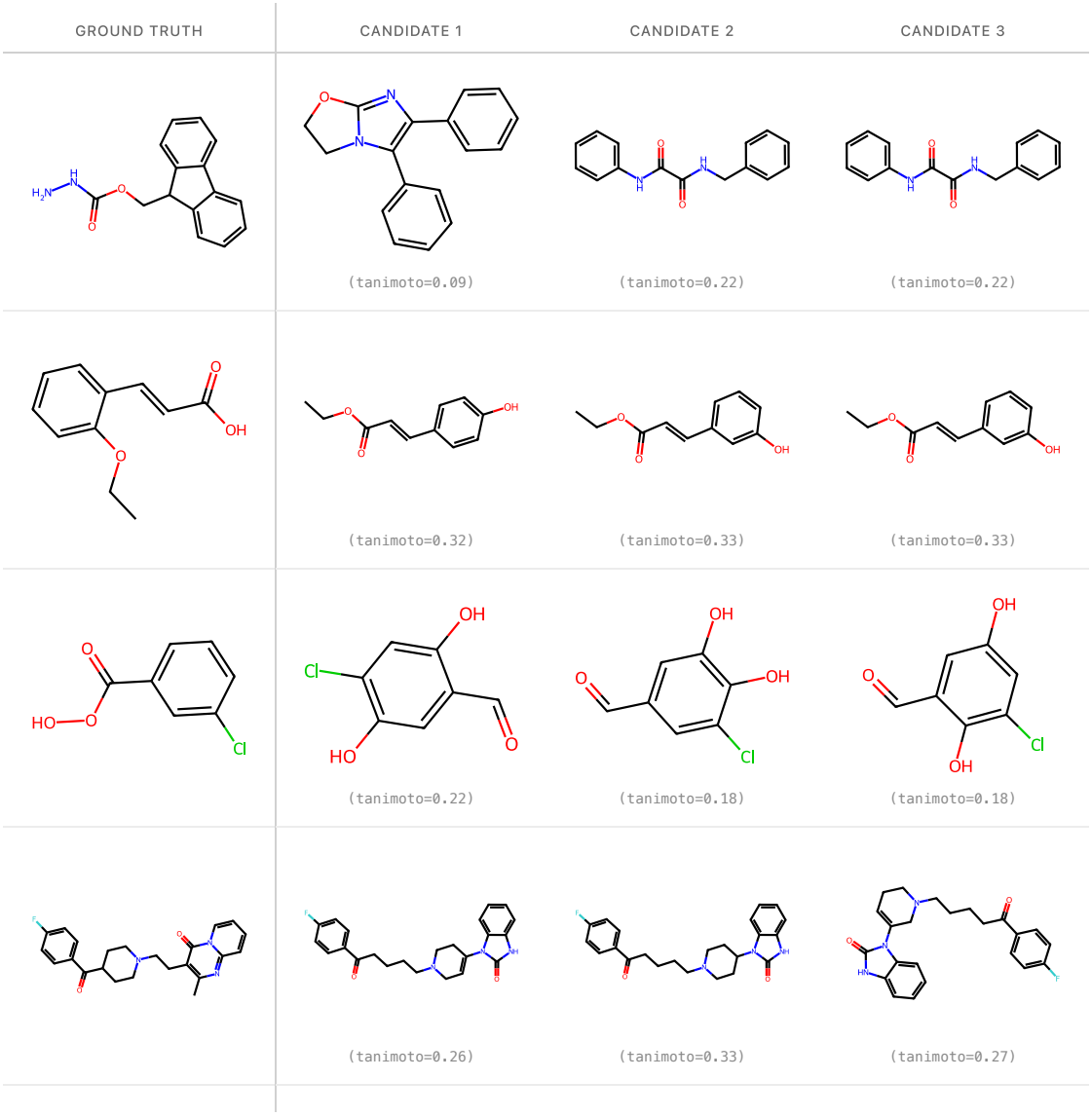

Figure 5. **Additional incorrect predictions by our agent.** Representative failure cases from `kimi-k2.6`.

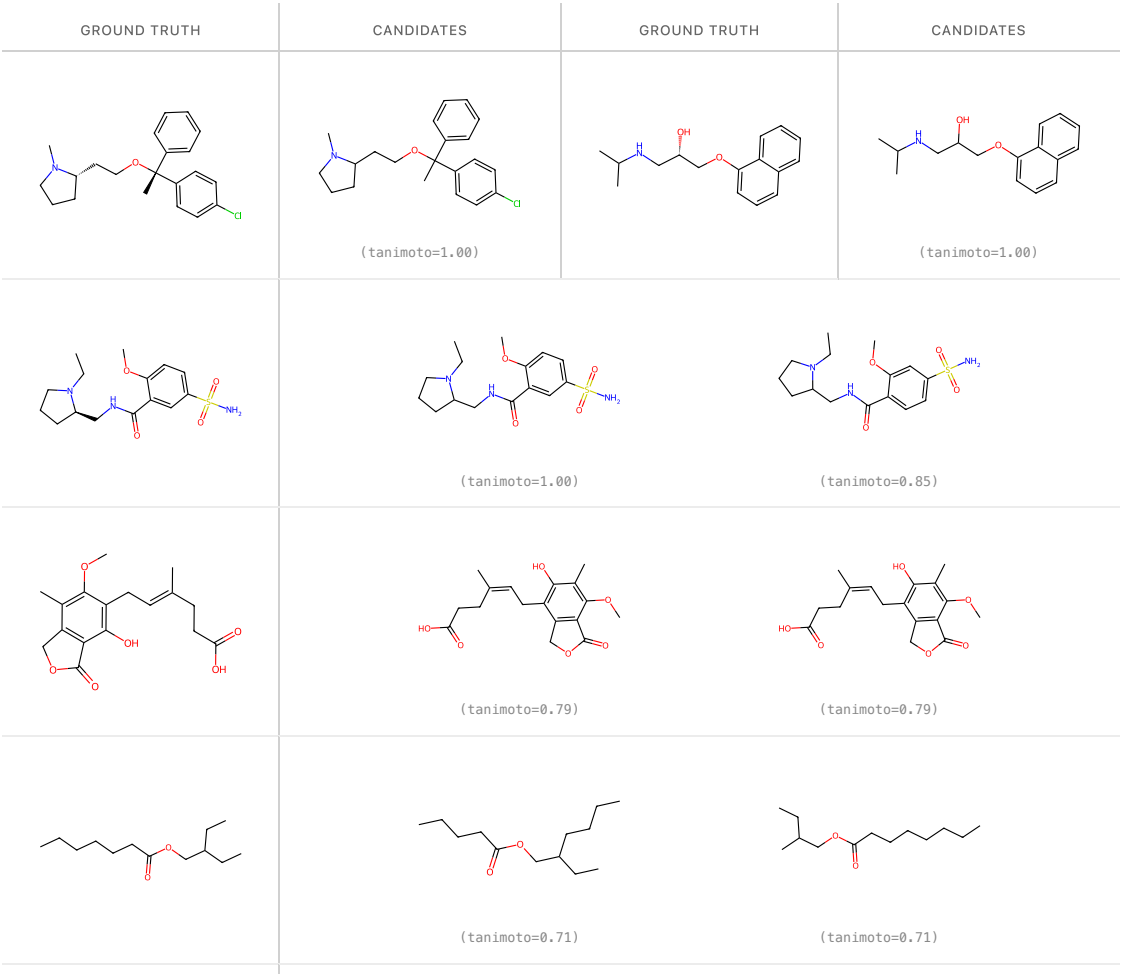

Figure 6. **Additional near-miss predictions by our agent.** Representative near-miss cases from `kimi-k2.6`.

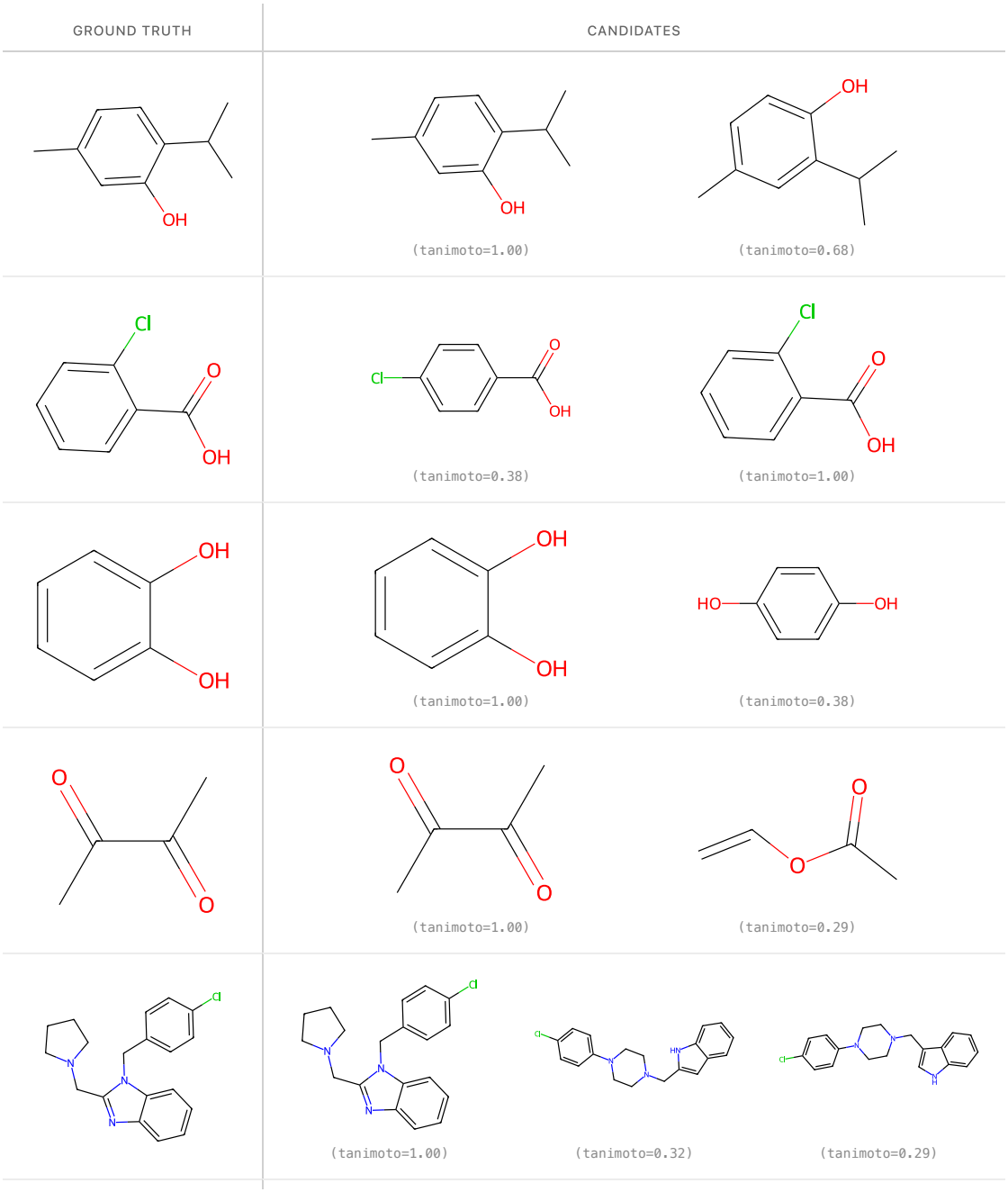

Figure 7. **Additional mixed predictions by our agent.** Representative cases where the agent predicts the correct answer in some, but not all, independent runs from `kimi-k2.6`.