Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Towards Generalizable and Evidential Nuclear Magnetic Resonance-Based Molecular Structure Elucidation via Large Language Model Agent

Jun Xia

junxia@hkust-gz.edu.cn

The Hong Kong University of Science and Technology (Guangzhou) https://orcid.org/0000-0002-7993-0803

Zheng Fang

Information Hub, The Hong Kong University of Science and Technology (Guangzhou) https://orcid.org/0009-0009-9735-4350

Chen Yang

Information Hub, The Hong Kong University of Science and Technology (Guangzhou)

Yusen Tan

Information Hub, The Hong Kong University of Science and Technology (Guangzhou)

Yunpeng Zhao

The University of Hong Kong

Fanjie Xu

Xiamen University https://orcid.org/0009-0007-1007-5427

Hongxin Xiang

Hunan University https://orcid.org/0000-0001-8345-8735

Hanyu Sun

State Key Laboratory of Bioactive Substances and Functions of Natural Medicines, Institute of Materia Medica, Peking Union Medical College and Chinese Academy of Medical Sciences

Hanyu Gao

The Hong Kong University of Science and Technology https://orcid.org/0000-0002-6346-0739

Xiaojian Wang

State Key Laboratory of Bioactive Substances and Functions of Natural Medicines, Institute of Materia Medica, Peking Union Medical College and Chinese Academy of Medical Sciences

Wenjie Du

University of Science and Technology of China

Yuqiang Li

Shanghai Artificial Intelligence Laboratory https://orcid.org/0000-0001-6756-6154

Brief Communication

Keywords:

Posted Date: August 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10343432/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Towards Generalizable and Evidential Nuclear Magnetic Resonance-Based Molecular Structure Elucidation via Large Language Model Agent

Zheng Fang[1], Chen Yang[1], Yusen Tan[1], Yunpeng Zhao[4], Fanjie Xu[3], Hongxin Xiang[7], Hanyu Sun[5], Hanyu Gao[6], Xiaojian Wang[5], Wenjie Du[2], Yuqiang Li[8], Jun Xia[1, 6]

[1]Information Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511432, P. R. China
[2]State Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei 230026, P. R. China
[3]State Key Laboratory of Physical Chemistry of Solid Surface, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, P. R. China
[4]The University of Hong Kong, Hong Kong SAR, P. R. China
[5]State Key Laboratory of Bioactive Substances and Functions of Natural Medicines, Institute of Materia Medica, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing 100050, P. R. China
[6]The Hong Kong University of Science and Technology, Hong Kong SAR, P. R. China
[7]School of Computer Science, Hunan University, Changsha 410082, P. R. China
[8]AI for Chemistry Center, Shanghai Artificial Intelligence Laboratory, Shanghai 200235, P. R. China

Abstract

Nuclear Magnetic Resonance (NMR) spectroscopy is the gold standard for molecular structure elucidation, yet interpreting complex spectra for unknown molecules remains a bottleneck reliant on human expertise. While artificial intelligence has advanced this field, current methods face a critical trade-off: database retrieval cannot identify novel scaffolds, while *de novo* molecular structure elucidation models operate as black boxes, lacking the atom-level interpretability required for rigorous scientific validation. Here, we present NMRAgent, an evidential reasoning agent powered by large language models (LLMs) that bridges this gap through a suite of specialized tools for database retrieval, *de novo* structure generation, spectral simulation, and molecular editing. Mimicking the deductive reasoning of human experts, NMRAgent takes experimental NMR spectra and a molecular formula as input, plans the elucidation process, proposes candidate structures, verifies peak–atom consistency, and selectively refines spectrally unsupported substructures under molecular formula constraints. Enabled by its evidential reasoning, NMRAgent outperforms state-of-the-art methods, improving top-1 accuracy by 46.5% and Tanimoto similarity by 0.502 on a scaffold-split benchmark with novel scaffolds in the test set. Besides, we demonstrate the agent's practical utility by elucidating the structures of two previously unknown natural products isolated from *Hydrangea davidii* and *Vitex trifolia*, and by correcting structural misassignments in established literature. By combining high-accuracy prediction with transparent and evidence-based reasoning, NMRAgent establishes a new paradigm for interpretable AI in analytical chemistry.

Correspondence:

Jun Xia (junxia@hkust-gz.edu.cn)

1 Introduction

Nuclear Magnetic Resonance (NMR) spectroscopy is a cornerstone of molecule structure elucidation, providing rich information about molecular connectivity, stereochemistry, and local chemical environments [6]. Among NMR techniques, ^{1}H and ^{13}C NMR spectra are particularly widely used in routine organic and natural-product analysis because of their accessibility and diagnostic value [14]. Traditional NMR structure elucidation combines empirical spectral interpretation, quantum-chemical prediction, and database-assisted matching. While chemists use chemical-shift trends, multiplicity patterns, and coupling information to infer structural fragments [8, 21], Density Functional Theory (DFT)-based NMR calculations provide accurate candidate validation and stereochemical ranking [5, 31], and databases such as nmrshiftdb enable spectrum–structure dereplication [27]. Nevertheless, these approaches remain constrained by expert labor, computational cost, and limited database coverage, motivating scalable data-driven methods for automated NMR-based structure elucidation.

To address these limitations, recent studies have increasingly explored *de novo* molecular structure elucidation models based on deep learning, reframing automated NMR-based structure elucidation as an inverse mapping from NMR spectra to molecular structures. A common line of work formulates this inverse problem as sequence-to-sequence generation, where NMR spectra are encoded into neural representations and molecular structures are decoded as SMILES strings [30]. For instance, CLAMS [28] processes NMR spectra as 2D images using CNNs, whereas NMRTransformer [2] employs a 1D-CNN architecture to automatically learn spectral tokenization. Models such as NMR2Struct [12] and NMRMind [34] provide more explicit representations by encoding specific chemical shift ranges and intensities for transformer attention. Recently, inspired by breakthroughs in generative modeling [11, 18], diffusion-based paradigms have emerged. DiffNMR [36] progressively denoises and generates molecular 2D graphs, while ChefNMR [32] generates 3D atomic conformations. Concurrently, retrieval-based methods such as NMRSolver [13] improve chemical validity by searching large-scale simulated spectral databases generated with fast machine-learning NMR predictors [15, 33], followed by fragment-based optimization.

Despite these rapid advancements, existing deep learning-based NMR structure elucidation methods still face two critical challenges that hinder their practical deployment. First, retrieval-based methods often achieve higher accuracy than *de novo* models by searching over chemically valid reference databases. However, their performance is limited by the coverage and quality of these databases. When the target molecule contains a rare scaffold or represents a newly isolated natural product outside the retrieval space, these methods cannot easily propose genuinely novel structures. Second, most *de novo* molecular structure elucidation models operate largely as black-box systems. They directly output molecular structures while bypassing the peak-to-atom assignment step that is central to conventional NMR structure verification, limiting the interpretability of how the observed spectral signals support the predicted structure.

To address these challenges, we propose **NMRAgent**, a large-language-model-powered agentic framework designed for NMR structure elucidation. By leveraging the reasoning capability of large language model (LLM) together with specialized NMR analysis tools, NMRAgent introduces an evidential reasoning paradigm that explicitly identifies peak-atom assignments for structure verification. Unlike existing black-box models, this design grounds structural hypotheses in interpretable spectral evidence and supports more chemically reliable automated elucidation. More specifically, we design a three-level hierarchical architecture. At the ***Knowledge-Level***, a LLM planner utilizes Retrieval-Augmented Generation (RAG) over external chemical knowledge graph to formulate logical step-by-step elucidation plans. Guided by these plans, the ***Molecular-Level*** constructs a broad yet chemically plausible candidate space by combining *de novo* generation with retrieval tools. Retrieval provides database-grounded candidates, while *de novo* generation enables exploration beyond direct database matches. Using fast NMR prediction models, generated structures are annotated with atom chemical shifts and then reranked with retrieved candidates using a learned spectral-structural representation that considers both NMR consistency and molecular similarity. The resulting candidate pool is then examined by an LLM-powered Peak-Atom Verifier, which produces explicit peak–atom assignments to support interpretable structural reasoning. Finally, if structural discrepancies persist, the ***Fragment-Level*** performs evidence-guided structural optimization. Leveraging the LLM's chemical reasoning capability, NMRAgent interprets verifier-provided peak-atom mismatches as refinement hypotheses and applies a formula-aware heuristic to select and recombine compatible fragments from candidate pools. This work brings a new perspective to NMR-based molecular structure elucidation by introducing the first agentic framework that coordinates chemical knowledge, specialized tools, and spectral evidence for interpretable structural reasoning, offering a foundation for future advances in automated spectroscopy-driven chemical analysis.

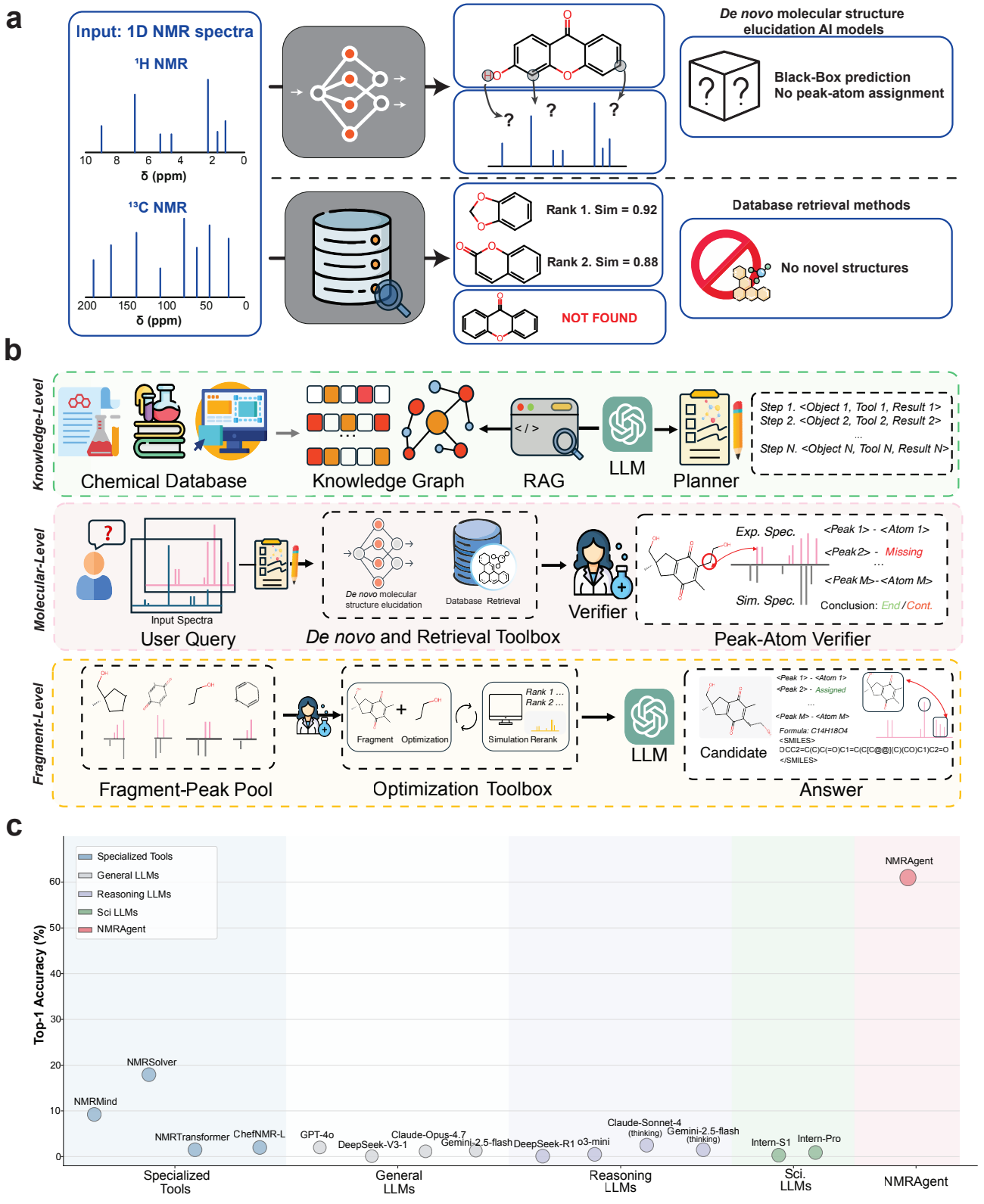


Figure 1 | Overview of previous methods and our proposed NMRAgent. a. Limitations of previous NMR structure elucidation methods. **b.** The overall workflow of NMRAgent. (1) Knowledge-Level: A Planner leverages RAG over chemical knowledge graph to formulate step-by-step elucidation plans. (2) Molecular-Level: A diverse *de novo* and retrieval toolbox generates candidates which are then rigorously validated by a peak-atom verifier through experimental-to-simulated spectral matching. (3) Fragment-Level: For unresolved structures, the agent autonomously invokes an optimization toolbox to perform fragment optimization, guided by missing peak evidence. **c.** NMRAgent achieves state-of-the-art performance on real-world datasets. Comparison of Top-1 Accuracy (%) across five categories: Specialized Tools, General LLMs, Reasoning LLMs, Science LLMs, and our proposed NMRAgent. While existing methods—including specialized computational tools and LLMs—struggle to exceed a 20% accuracy threshold, NMRAgent significantly outperforms all baselines, achieving a breakthrough SOTA performance of over 60% in real-world scenarios.

2 Results and Discussion

2.1 Overview of NMRAgent

As shown in Figure 1b, we developed NMRAgent as an agentic framework for NMR-based molecular structure elucidation, designed to bridge candidate generation and evidence-based structural verification. Rather than treating elucidation as a single-step prediction problem, NMRAgent decomposes the task into a hierarchical workflow that mirrors expert chemical reasoning: retrieving relevant prior knowledge, proposing candidate structures, verifying atom-level spectral consistency, and refining unresolved regions.

At the Knowledge-Level, NMRAgent uses a retrieval-augmented planner to introduce external chemical knowledge into the elucidation process. The planner queries a chemical knowledge graph to retrieve relevant chemical context, including compound-level annotations, scaffold-related records, functional-group clues. This retrieval step supplements the LLM with domain-specific chemical context that are complementary to and coupled with NMR spectral evidence. The LLM then organizes the retrieved information into a reasoning plan, specifying which tools should be invoked and what structural evidence should be examined. This level provides high-level chemical context and constrains subsequent molecular search beyond unconstrained enumeration.

At the Molecular-Level, NMRAgent constructs a broad candidate pool from the input spectra by integrating retrieval and *de novo* generation. We use a 158M spectrum-structure database based on PubChem-NMRNet [13] and a large-scale NMR spectral vector library [34, 35], and adopt deep-learning-based dense spectral representations [35] to enable scalable molecular search. To move beyond conventional spectrum-only retrieval, NMRAgent first retrieves spectra-level neighbors using Gaussian-based global NMR vectors and then re-ranks the corresponding structures with molecular dense embeddings, prioritizing candidates that are both spectrally consistent and chemically similar. Meanwhile, the *de novo* generator proposes additional structures beyond database entries. Fast machine-learning NMR predictors [33] are then used to produce atom-annotated spectra for retrieved and generated candidates, which are further examined by the LLM-powered Peak-Atom Verifier to establish explicit peak-atom correspondences.

At the Fragment-Level, NMRAgent uses the verifier output to perform targeted structural refinement when full-molecule candidates are not sufficiently supported. The candidate molecules are cleaved into chemically meaningful fragments, and the LLM interprets unresolved peak-atom mismatches to decide which fragments should be preserved and which regions should be exposed to repair. The optimization process is constrained by the target molecular formula and BRICS assembly rules, allowing bounded multi-fragment recombination while retaining spectrally supported substructures. Newly assembled candidates are then returned to the Peak-Atom Verifier, forming an iterative loop between atom-level spectral evidence and fragment-level optimization.

2.2 Generalization of NMRAgent Across Benchmarks

2.2.1 Evaluation on Scaffold-Split Benchmark

We fine-tuned NMRAgent on the training split of NMRGym and evaluated its structure elucidation performance on the held-out scaffold-split test set. NMRGym is a benchmark for NMR-based molecular structure elucidation designed to assess generalization to unseen molecular scaffolds [7]. Unlike random-split datasets, NMRGym separates molecules by scaffold, requiring models to infer structures that are topologically distinct from those seen during training or retrieval construction. This design better reflects real-world natural product elucidation, where target molecules often contain rare or previously unseen scaffolds rather than close database analogues. Additional details of NMRGym, including the fine-tuning setup, scaffold-split protocol, and molecular complexity statistics, are provided in the Supplementary Information.

We first assessed whether existing large language models can directly solve this task. As shown in Figure 1c, general-purpose, reasoning-oriented, and science-oriented LLMs exhibit near-zero Top-1 accuracy in this setting. Although these models can occasionally succeed on simple cases and often produce fluent chemical explanations, they generally lack the spectral grounding required to recover exact molecular structures from NMR peaks. This result highlights a clear gap between language-level chemical reasoning and executable NMR structure elucidation: successful elucidation requires not only chemical knowledge, but also specialized molecular search, spectral simulation, and quantitative evidence verification.

We then compared NMRAgent with representative task-specific baselines on NMRGym, including retrieval-based methods and DL-based *de novo* methods. As shown in Figure 2a, *De novo* methods show limited accuracy under

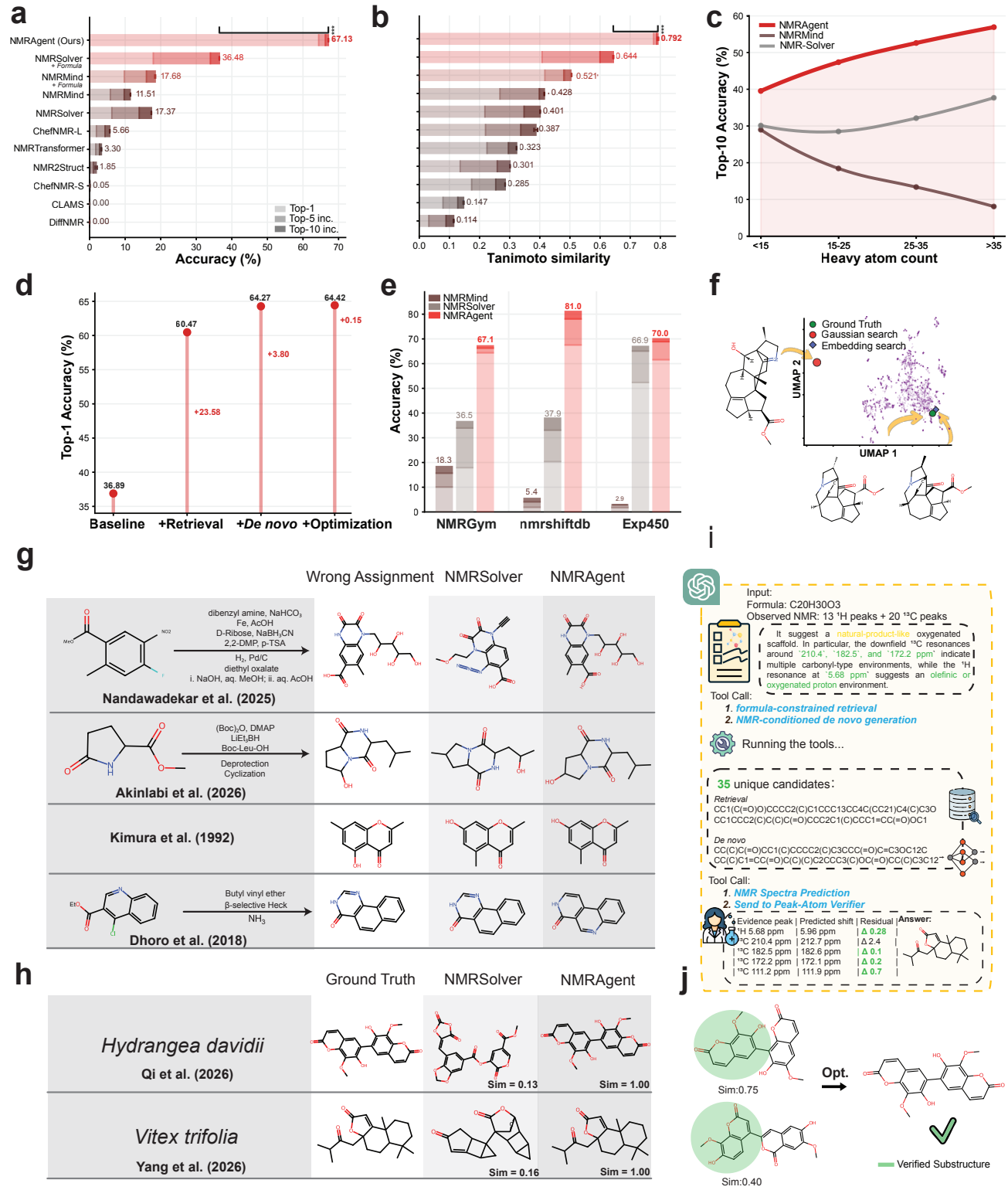


Figure 2 | Benchmark and real-world evaluation of NMRAgent. a. Top-k accuracy on NMRGym, showing that NMRAgent outperforms previous methods. **b.** Tanimoto similarity between predicted and ground-truth molecules on NMRGym, indicating improved structural proximity of NMRAgent predictions. **c.** Accuracy stratified by molecular size. Existing methods exhibit degraded performance as heavy atom count increases, whereas NMRAgent maintains stable accuracy on larger and more complex molecules. **d.** Ablation study of retrieval, *De novo* generation, and fragment-level optimization, demonstrating the complementary contribution of each module. **e.** Evaluation across NMRGym, nmrshiftdb, and Exp450 benchmarks, highlighting the robustness of NMRAgent across diverse test sets. **f.** Embedding-space analysis showing that dense embedding retrieves structures closer to the ground truth. **g.** Representative example where NMRAgent corrects a literature-reported human peak-assignment error. **h.** Real-world natural product elucidation examples. **i.** An evidential reasoning trajectory of a real-world case. **j.** A real-world optimization case.

this scaffold-split setting, suggesting that direct spectrum-to-structure generation remains challenging for complex molecules with unseen scaffolds. Retrieval-based methods perform better because they search over chemically valid structures, but their performance remains constrained by database coverage and retrieval-induced structural divergence. In contrast, NMRAgent achieves a Top-10 accuracy of 67.13%, substantially outperforming the strongest baseline, NMRSolver+Formula, which reaches 36.48%. This corresponds to an absolute improvement of 30.65 percentage points and a relative improvement of approximately 84.0%. The same trend is observed in structural similarity evaluation: NMRAgent achieves the highest Tanimoto similarity to the ground-truth molecules, indicating that its predictions are not only more often exact, but also closer to the correct structures in chemical space when exact recovery is not achieved (Figure 2b).

We further analyzed model performance as molecular size increases. A major challenge in NMR elucidation is that larger molecules usually contain more complex scaffolds, denser peak patterns, and more ambiguous local chemical environments. As shown in Figure 2c, baseline methods exhibit a clear degradation in accuracy as the heavy atom count increases. This trend indicates that existing methods struggle to maintain reliable structure recovery when the molecular search space becomes larger and the spectra become more difficult to disambiguate. In contrast, NMRAgent maintains stable performance across molecular-size groups, suggesting that the integration of retrieval, *de novo* generation, atom-level verification, and fragment-level optimization provides a more robust strategy for complex natural-product-like molecules.

To further understand the source of NMRAgent's improvement, we conducted ablation studies over the major components of the framework. As shown in Figure 2d, retrieval, *de novo* generation, and fragment-level optimization contribute complementary gains. Retrieval provides database-grounded chemically valid candidates, *de novo* generation expands the search beyond direct database matches, and fragment-level optimization repairs unresolved regions using molecular formula constraints and verifier-provided mismatch evidence. The progressive improvement across these settings indicates that NMRAgent benefits from an integrated elucidation workflow rather than a single dominant module. The ablation also shows that the gain is not merely a consequence of enlarging the retrieval database. When the NMRSolver-style retrieval baseline is evaluated with the expanded database, its Top-1 accuracy already exceeds the Top-10 accuracy obtained using the original database, confirming the importance of database coverage. Nevertheless, NMRAgent further improves over this expanded-database baseline, suggesting that the proposed representation, verification, and optimization mechanisms provide additional benefits beyond candidate pool expansion. We further provide a qualitative case study to analyze the retrieval behavior. As shown in Figure 2f, dense embedding search retrieves a candidate that is structurally closer to the ground-truth molecule. This observation suggests that the learned embedding space better preserves molecular semantic similarity while maintaining spectral consistency, thereby reducing retrieval-induced structural divergence. Figure 2j further illustrates this evidence-guided refinement process in a real-world case: NMRAgent preserves the substructures already supported by the spectrum and selectively optimizes the unmatched regions, yielding a final structure that is more consistent with the observed NMR evidence. Additional case study visualizations and peak-atom correspondence results are provided in the Supplementary Information.

2.2.2 Evaluation on Additional Benchmarks

We further assessed NMRAgent on nmrshiftdb and Exp450 to evaluate its robustness across different data sources. nmrshiftdb is a commonly used public NMR database [27], whereas Exp450 is a literature-derived experimental benchmark [13]. Compared with NMRGym, these datasets present different challenges: nmrshiftdb contains many entries with incomplete spectral information, often including only a single spectrum type, while Exp450 reflects curated experimental cases collected from published reports.

As shown in Figure 2e, NMRAgent achieves the strongest performance on both datasets. On nmrshiftdb, NMRAgent reaches a Top-10 accuracy of 81.0%, compared with 37.9% for NMRSolver and 5.4% for NMRMind. The large improvement indicates that NMRAgent can effectively handle heterogeneous public-database spectra, even when complete paired ^{1}H and ^{13}C information is not always available. This robustness is consistent with the design of the dense NMR encoder, which learns continuous spectral representations rather than relying solely on fixed sparse peak-vector matching. On Exp450, NMRAgent obtains a Top-10 accuracy of 70.0%, outperforming NMRSolver and NMRMind. Notably, NMRSolver remains competitive on Exp450, reflecting the advantage of retrieval-based methods in literature-derived benchmarks. However, NMRAgent further improves over this strong retrieval baseline by integrating coarse-to-fine retrieval, *de novo* candidate generation, and optimization. These results show that NMRAgent is not limited to a single benchmark distribution, but can maintain strong performance across public database settings and experimentally curated literature cases.

2.3 Real-world Application of NMRAgent

NMR Assignment Correction and Structural Revision. Incorrect NMR-based structural assignments remain a persistent challenge in natural product chemistry, especially when regioisomers, scaffold isomers, or closely related analogues exhibit highly similar one-dimensional NMR spectra. To evaluate whether NMRAgent can assist structural revision, we therefore re-analyzed four representative structural revision cases, including Altechromone A [16, 17], Samoquasine A [4, 19, 29], a natural tetrahydroquinoxaline-6-carboxylic acid from Caulis Sinomenii [20], and C5-hydroxy-cyclo(L-Pro-L-Leu) [1]. These cases cover common sources of NMR misassignment, including positional ambiguity, regioisomeric uncertainty, and insufficient discrimination between structurally close candidates. For each case, NMRAgent was provided only with the reported ^{1}H and ^{13}C NMR chemical shifts, without access to the subsequent synthetic, crystallographic, or DFT evidence used in the human revisions. Despite this restricted input, NMRAgent recovered the revised structures rather than the originally proposed assignments. These results suggest that NMRAgent can serve as a practical tool for detecting potentially incorrect NMR assignments and prioritizing chemically plausible structural revisions.

Novel Natural Product Structure Elucidation. Beyond correcting previously misassigned structures, we further evaluated NMRAgent in novel natural product elucidation scenarios. We considered two representative novel-structure cases: one coumarin dimer isolated from *Hydrangea davidii* [23], and one compound isolated from *Vitex trifolia* [35]. These cases represent more realistic application settings, where the target molecules are not simple benchmark compounds but structurally complex natural products requiring interpretation from experimental NMR evidence. NMRAgent successfully recovered structures consistent with the literature-reported assignment and the collaborating group's experimentally validated solution, respectively. These results demonstrate that NMRAgent can extend beyond retrospective benchmark evaluation and support practical structure elucidation of newly isolated natural products. We show that NMRAgent demonstrates evidential reasoning capability, explicitly tracing how observed NMR signals support the predicted structure, as illustrated in Figure 2i.

In addition to structure-level elucidation, we also explored whether the learned NMR representation could provide tentative metadata annotations for newly isolated compounds. Specifically, we integrated the knowledge-graph database described above with a fine-tuned NMR encoder, using approximately 1,000 labeled examples per class to train lightweight classifiers for chemical superclass, natural product pathway, and genus-level metadata. Rather than serving as definitive annotations, these predictions are intended to provide auxiliary hypotheses that may help contextualize novel molecules in broader natural-product chemical space. The resulting embedding-space visualizations, including t-SNE plots, are provided in the Supplementary Information.

3 Discussion and Limitations

Despite its strong performance, NMRAgent still has several limitations. First, the current framework focuses on one-dimensional ^{1}H and ^{13}C NMR spectra, while expert elucidation often relies on richer spectroscopic evidence, including 2D NMR, IR, UV, and stereochemical measurements. Future versions should therefore incorporate additional tools and modalities to support more comprehensive spectral reasoning. Second, although the expanded database substantially improves candidate coverage, NMRAgent remains partially constrained by database completeness and still does not fully reach the reliability of experienced human experts, especially for rare scaffolds, noisy spectra, or highly ambiguous natural products. Third, the LLM component does not yet intrinsically understand NMR spectral information; its current effectiveness mainly comes from tool grounding, retrieval, and verifier-provided evidence. Future work should explore agentic learning and post-training strategies that allow LLMs to learn from verifier feedback, failed reasoning trajectories, and expert corrections, thereby internalizing spectroscopic knowledge more effectively. Despite these limitations, NMRAgent establishes an LLM-powered agentic framework for evidence-grounded NMR structure elucidation. By moving beyond direct black-box prediction toward interpretable candidate verification, our work provides a scalable foundation for incorporating richer spectral modalities, broader chemical knowledge, and stronger spectroscopic reasoning in future systems.

4 Method

4.1 Preliminaries

In standard structure elucidation pipelines, the exact molecular formula $\mathcal{F}$ is typically determined prior to NMR analysis, often utilizing mass spectrometry (MS) [21]. Given $\mathcal{F}$ as a global structural constraint, the

subsequent task is to deduce the precise atomic connectivity. We formally define the input NMR information as a collection of peak sets. Specifically, let $\mathcal{S}_H = \{s_1^H, s_2^H, \ldots, s_{N_H}^H\}$ and $\mathcal{S}_C = \{s_1^C, s_2^C, \ldots, s_{N_C}^C\}$ denote the ^{1}H and ^{13}C NMR spectra, respectively, where each peak $s_i \in \mathbb{R}$ represents a chemical shift value (in ppm). Detailed chemical definitions are provided in Supplementary Information. Since the physical measurement of NMR peaks is invariant to permutation, we treat each spectrum strictly as a set rather than a sequence. For methodological simplicity in the following sections, we denote the unified input as $\mathcal{S} = \mathcal{S}_H \cup \mathcal{S}_C$, assuming that both spectra are concurrently provided to the model. The corresponding molecular structure is represented as a SMILES [30] sequence $Y = \{y_1, \ldots, y_L\}$, where each element $y_i \in \mathcal{V}$ denotes a discrete token from a predefined chemical vocabulary $\mathcal{V}$ (e.g., atom symbols, bonds, and structural indicators).

4.2 Overall Workflow

To elucidate a molecular structure from NMR spectra, human chemists typically follow a deductive workflow that starts from global constraints such as the molecular formula, interprets local chemical environments from NMR peaks, and iteratively verifies structural hypotheses against experimental evidence [8, 21, 25]. Inspired by this process, NMRAgent adopts a three-level hierarchical workflow consisting of Knowledge-Level planning, Molecular-Level candidate generation and verification, and Fragment-Level optimization. An LLM acts as the reasoning controller throughout the workflow, coordinating knowledge retrieval, tool invocation, candidate evaluation, and evidence-based refinement.

As illustrated in Figure 1, the ***Knowledge-Level*** uses a retrieval-augmented planner to formulate a target-specific elucidation plan from external chemical knowledge. The ***Molecular-Level*** constructs full-molecule candidates conditioned on the input spectra $\mathcal{S}$ and molecular formula $\mathcal{F}$, using a *de novo* and retrieval toolbox together with an LLM-powered Peak-Atom Verifier that establishes explicit peak-atom correspondences. If unresolved spectral evidence remains, the ***Fragment-Level*** performs targeted fragment optimization guided by the missing or inconsistent peak evidence identified by the verifier. Detailed descriptions of each module are provided below.

4.3 Knowledge-Level Planner

In chemical research, expert chemists rely on standardized analytical workflows for molecular structure elucidation, grounded in established spectroscopic heuristics and empirical rules [8, 21, 25]. To emulate this human expertise, we propose the **Knowledge-Level Planner**, a central reasoning module within NMRAgent that uses Retrieval-Augmented Generation (RAG) to formalize complex structure elucidation processes. To establish a comprehensive knowledge base, we integrate heterogeneous scientific resources, including natural product repositories [22, 24, 26] and bioactivity-related databases [3, 10]. In addition to external chemical knowledge, the planner maintains an NMR evidence memory that stores previously verified spectroscopic reasoning cases, such as confirmed peak–atom assignments, solvent conditions, multiplicity patterns, confidence scores, and expert or user feedback. This memory allows the planner to reuse reliable assignment evidence from prior elucidation cases, thereby making the reasoning process closer to the way chemists draw on accumulated spectroscopic experience.

Formally, the planning process is triggered by the molecular formula $\mathcal{F}$, input spectra $\mathcal{S}$, and experimental metadata $\mathcal{M}$, where $\mathcal{M}$ may include available contextual information such as reaction precursors, reagents, catalysts, biological source, or isolation conditions. We denote the external chemical knowledge graph as $\mathcal{K}$ and the NMR evidence memory as $\mathcal{H}$. Given $(\mathcal{F}, \mathcal{S}, \mathcal{M})$, the planner retrieves relevant chemical context from $\mathcal{K}$ and assignment-level evidence from $\mathcal{H}$:

$$\mathcal{C}_K = \mathcal{R}_K(\mathcal{K}; \mathcal{F}, \mathcal{M}), \qquad \mathcal{C}_H = \mathcal{R}_H(\mathcal{H}; \mathcal{S}, \mathcal{F}, \mathcal{M}), \tag{1}$$

where $\mathcal{C}_K$ denotes retrieved chemical knowledge and $\mathcal{C}_H$ denotes retrieved memory evidence from previous NMR assignment cases. The LLM then synthesizes these retrieved contexts into a procedural guideline $\mathcal{G}$:

$$\mathcal{G} = \mathrm{LLM}(\mathcal{F}, \mathcal{S}, \mathcal{M}, \mathcal{C}_K, \mathcal{C}_H). \tag{2}$$

To translate the high-level guideline $\mathcal{G}$ into an executable workflow, the planner integrates it with a predefined toolset $\mathcal{T}$ and corresponding operation descriptions $\mathcal{A}$. It generates a molecule-specific elucidation plan $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n\}$, where each reasoning step is represented as:

$$\mathcal{P}_i = \langle t_i, x_i, e_i, y_i \rangle, \qquad t_i \in \mathcal{T}, \tag{3}$$

where t_i is the selected tool, x_i is the input to the tool, e_i is the supporting evidence retrieved from $\mathcal{C}_K$ or $\mathcal{C}_H$, and y_i is the expected structural inference or intermediate result. In this framework, the inclusion of experimental metadata and NMR evidence memory helps NMRAgent avoid blind structural enumeration. Instead, the planner produces a chemically grounded and evidence-supported elucidation plan that reflects both external chemical knowledge and reusable spectroscopic assignment experience.

4.4 Molecular-Level Candidate Construction and Verification

Recent *de novo* models directly generate molecular strings or graphs while bypassing the peak-to-atom assignment process that is central to conventional NMR verification. Second, retrieval-based methods provide chemically valid candidates, but their performance depends on the coverage of the reference spectrum–structure database and on the quality of the retrieval representation. In particular, retrieval-based methods are inherently constrained by database coverage and therefore cannot propose novel structures that are absent from the reference database. In addition, conventional spectrum-only retrieval may identify candidates with similar global peak distributions but distinct molecular scaffolds, especially for complex natural products with rare scaffolds or subtle local variations.

To address these issues, NMRAgent constructs a hybrid candidate pool by integrating database-grounded retrieval with *de novo* generation. We expand Pubchem-NMRNet [13] to 158M spectrum–structure pairs by augmenting the original database with additional compounds, including complex scaffold-containing and natural-product-like molecules. For the expanded portion of the database, NMR spectra are simulated *in silico* using Gaussian software [9], enriching the reference space beyond the original database coverage. This expanded database provides a broader pool of chemically valid candidates for downstream retrieval and reasoning. In parallel, the *de novo* and retrieval toolbox can invoke NMR-conditioned molecular generators, such as UltraNMR [35], ChefNMR [32] and NMRMind [34], to propose additional structures beyond database entries. The retrieved and generated candidates are then merged into a unified candidate pool for verification.

To support more informative retrieval, we use UltraNMR [35], a self-supervised representation model for NMR spectra. Given heterogeneous ^{1}H and ^{13}C NMR inputs, UltraNMR maps spectra into a dense continuous embedding space that captures both individual peak patterns and contextual dependencies across spectra. The model is pre-trained with a masked spectral modeling objective, where a subset of chemical-shift tokens is masked and predicted from the remaining spectral context. In addition, UltraNMR is trained to estimate structural similarity between molecules from their spectral embeddings:

$$\mathcal{L}_{\mathrm{SSL}} = \mathcal{L}_{\mathrm{mask}} + \lambda_{\mathrm{FP}}\mathcal{L}_{\mathrm{FP}}, \tag{4}$$

where $\mathcal{L}_{\mathrm{mask}}$ denotes the masked chemical-shift prediction loss and $\mathcal{L}_{\mathrm{FP}}$ denotes the fingerprint-similarity prediction loss. This objective encourages the spectral representation to preserve structural information useful for candidate ranking.

To further distinguish structurally similar candidates, especially formula-identical isomers, we further incorporates a cross-modal contrastive objective:

$$\mathcal{L}_{\mathrm{NCE}} = -\log \frac{\exp(\cos(\boldsymbol{z}, \boldsymbol{v}^{+})/\tau)}{\exp(\cos(\boldsymbol{z}, \boldsymbol{v}^{+})/\tau) + \sum_{k=1}^{K} \exp(\cos(\boldsymbol{z}, \boldsymbol{v}_k^{-})/\tau)}, \tag{5}$$

where $\boldsymbol{z}$ is the projected NMR embedding, $\boldsymbol{v}^{+}$ is the fingerprint representation of the matched molecule, $\boldsymbol{v}_k^{-}$ denotes the k-th hard-negative fingerprint representation, and τ is the temperature hyperparameter. The final representation learning objective is:

$$\mathcal{L}_{\mathrm{UltraNMR}} = \mathcal{L}_{\mathrm{SSL}} + \lambda_{\mathrm{NCE}}\mathcal{L}_{\mathrm{NCE}}. \tag{6}$$

For database-grounded candidate search, NMRAgent performs coarse-to-fine retrieval. At the coarse stage, the query spectrum is encoded into a global NMR vector and used for high-recall retrieval of spectra-level neighbors following the global vector retrieval protocol [13]. This step rapidly narrows the 158M-scale database to a manageable candidate set with similar overall NMR patterns. At the fine stage, the retrieved candidates are re-ranked using UltraNMR dense embeddings, prioritizing structures whose learned spectral representations are closer to the query. Thus, the coarse stage emphasizes efficient spectrum-level recall, whereas the fine stage refines the ranking with a representation that better reflects structural similarity.

In parallel with retrieval, UltraNMR is paired with an autoregressive molecular decoder for *de novo* candidate generation. Given a target SMILES sequence $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_L\}$ and the projected NMR embedding $\boldsymbol{z}$, the decoder models:

$$p_\theta(\boldsymbol{Y} \mid \boldsymbol{z}) = \prod_{t=1}^{L} p_\theta(y_t \mid y_{<t}, \boldsymbol{z}), \tag{7}$$

with the training objective:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{L} \log p_\theta(y_t \mid y_{<t}, \boldsymbol{z}). \tag{8}$$

This generation branch complements retrieval by proposing additional structures that may not appear in the reference database.

After candidate construction, NMRAgent applies an LLM-powered peak-atom verifier to convert candidate scoring into explicit atom-level evidence. Given the unified candidate pool $\mathcal{C}_{\text{pool}} = \{c_1, \ldots, c_M\}$ and the query spectrum $\mathcal{S}$, the verifier first invokes a fast machine-learning NMR predictor [33, 37] to generate atom-annotated chemical shifts for each candidate. For a candidate molecule c, this process produces $\hat{\mathcal{S}}^{\text{atom}}(c) = \{(\hat{\delta}, a, \alpha)\}$, where $\hat{\delta}$ is a predicted chemical shift, a is the associated atom index, and $\alpha \in \{^1\text{H}, ^{13}\text{C}\}$ denotes the nucleus type.

The verifier compares $\hat{\mathcal{S}}^{\text{atom}}(c)$ with the observed spectrum $\mathcal{S}$ using a set-similarity metric [13]. It then identifies compatible peak–atom correspondences, producing an assignment set $\mathcal{A}(c)$, a consistency score $s(c)$, and unresolved mismatch regions $\mathcal{U}(c)$. Together, these outputs form an evidence object:

$$\mathcal{E}(c) = \{c,\ s(c),\ \mathcal{A}(c),\ \mathcal{U}(c)\}. \tag{9}$$

The peak-atom verifier aggregates the verifier evidence across $\mathcal{C}_{\text{pool}}$ to produce either a ranked elucidation list or targeted mismatch descriptions for subsequent Fragment-Level optimization.

4.5 Fragment-Level Optimization

When the peak-atom verifier indicates insufficient spectral support for the current molecular candidates, NMRAgent activates the *Fragment-Level Optimization* module for targeted structural refinement. Rather than treating refinement as unconstrained molecular generation, this module uses verifier-provided evidence to localize uncertain structural regions and explores chemically valid fragment substitutions or recombinations under molecular-formula constraints. This design complements molecular-level candidate construction by preserving spectrally supported regions as structural anchors and focusing refinement on unresolved or mismatched regions.

Given the candidate pool evaluated by the peak-atom verifier, NMRAgent constructs a fragment pool $\mathcal{P}_{\text{frag}}$ by decomposing relevant candidates according to chemically defined fragmentation rules, such as BRICS-based cleavage or other rule-based bond-disconnection strategies. Each fragment is associated with its sub-formula and local spectral evidence inherited from the verifier. Specifically, atom-level shift annotations indicate which local regions are supported by matched experimental peaks and which regions remain ambiguous or inconsistent. The LLM acts as an evidence interpreter rather than a direct structure generator: it uses verifier outputs to identify fragments that should be preserved, regions that require modification, and alternative fragment compositions that should be explored.

Formally, let $\mathcal{F}_{\text{target}}$ denote the target molecular formula. Based on the peak–atom evidence, NMRAgent identifies a set of supported fragments $\mathcal{P}_{\text{keep}}$ and a candidate repair space $\mathcal{P}_{\text{edit}}$ associated with unresolved spectral evidence. Fragment-Level Optimization searches for chemically assembleable fragment combinations that satisfy the formula constraint and improve agreement with the experimental spectrum:

$$S^* = \arg\min_{S \subseteq \mathcal{P}_{\text{frag}}} d_{\text{set}}\left(\hat{\mathcal{Y}}_{\text{NMR}}(\text{Assemble}(S)), \mathcal{Y}_{\text{exp}}\right), \tag{10}$$

subject to:

$$\sum_{f \in S} \text{Formula}(f) = \mathcal{F}_{\text{target}}, \quad \mathcal{P}_{\text{keep}} \subseteq S, \quad S \setminus \mathcal{P}_{\text{keep}} \subseteq \mathcal{P}_{\text{edit}}. \tag{11}$$

Here, $\text{Assemble}(\cdot)$ denotes rule-guided chemical assembly, $\hat{\mathcal{Y}}_{\text{NMR}}(\cdot)$ denotes the predicted NMR shifts of the assembled molecule, $\mathcal{Y}_{\text{exp}}$ is the experimental spectrum, and $d_{\text{set}}(\cdot, \cdot)$ measures spectral disagreement. The formula

constraint acts as a hard filter, while the verifier-guided preservation and repair constraints focus the search on regions with insufficient peak–atom evidence. The resulting formula-compatible and chemically assembleable candidates are then passed back to the peak–atom verifier for iterative evidence-based assessment.

5 Code and Data Availability

The code is available at https://github.com/Meaw0415/NMRAgent. The checkpoints of UltraNMR [35] are available at https://huggingface.co/milesyc/ultranmr. The released NMRMind [34] checkpoints used in this work can be accessed through the official NMRMind repository at https://github.com/WJmodels/NMRMind. The released ChefNMR [32] checkpoints are available at https://zenodo.org/records/17766755. The released NMRNet [33] checkpoints for forward NMR prediction are available at https://zenodo.org/records/19142375. The source data of the Pubchem-NMRNet database [13] can be accessed through https://huggingface.co/datasets/yqj01/SimNMR-PubChem. The NMRGym [7] dataset can be accessed through https://huggingface.co/datasets/meaw0415/NMRGym. The processed nmrshiftdb data [27] can be obtained at https://zenodo.org/records/19142375. The Exp450 dataset can be obtained at https://doi.org/10.5281/zenodo.16952024. Detailed information on the datasets is provided in the Supplementary Information.

References

[1] Blessing B Akinlabi, Jared Balsz-Diaz, Mark C Walker, and Matthew R Aronoff. Synthesis-driven structural revision of c5-hydroxy-cyclo (l-pro-l-leu) using electrochemical oxidation. 2026.

[2] Marvin Alberts, Federico Zipoli, and Alain Vaucher. Learning the language of nmr: structure elucidation from nmr spectra using transformer models. In AI for Accelerated Materials Design-NeurIPS 2023 Workshop, 2023.

[3] Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. Scientific data, 10(1):67, 2023.

[4] Francis Dhoro, Jesse Parkin-Gibbs, Matthew McIldowie, Brian W Skelton, and Matthew J Piggott. Confirmation of the revised structure of samoquasine a and a proposed structural revision of cherimoline. Journal of natural products, 81(7):1658–1665, 2018.

[5] Robert Ditchfield. Self-consistent perturbation theory of diamagnetism: I. a gauge-invariant lcao method for nmr chemical shifts. Molecular Physics, 27(4):789–807, 1974.

[6] Abdul-Hamid Emwas, Kacper Szczepski, Benjamin Gabriel Poulson, Kousik Chandra, Ryan T McKay, Manel Dhahri, Fatimah Alahmari, Lukasz Jaremko, Joanna Izabela Lachowicz, and Mariusz Jaremko. Nmr as a "gold standard" method in drug design and discovery. Molecules, 25(20):4597, 2020.

[7] Zheng Fang, Chen Yang, Hai-tao Yu, Haoming Luo, Haitao He, Jiaqing Xie, Zhuo Yang, and Jun Xia. Nmrgym: A comprehensive benchmark for nuclear magnetic resonance based molecular structure elucidation. arXiv preprint arXiv:2601.15763, 2026.

[8] Leslie D Field, Sev Sternhell, and John R Kalman. Organic Structures from Spectra. John Wiley & Sons, 2012.

[9] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Jr. Montgomery, J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian 16, revision c.01. Gaussian, Inc., Wallingford CT, 2016.

[10] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. Nucleic acids research, 40(D1):D1100–D1107, 2012.

[11] Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In International conference on machine learning, pages 8867–8887. PMLR, 2022.

[12] Frank Hu, Michael S Chen, Grant M Rotskoff, Matthew W Kanan, and Thomas E Markland. Accurate and efficient structure elucidation from routine one-dimensional nmr spectra using multitask machine learning. ACS Central Science, 10(11):2162–2170, 2024.

[13] Yongqi Jin, Jun-Jie Wang, Fanjie Xu, Xiaohong Ji, Zhifeng Gao, Linfeng Zhang, Guolin Ke, Rong Zhu, and Weinan E. Nmr-solver: automated structure elucidation via large-scale spectral matching and physics-guided fragment optimization. Nature Communications, 2026.

[14] Eric Jonas, Stefan Kuhn, and Nils Schlörer. Prediction of chemical shift in nmr: A review. Magnetic Resonance in Chemistry, 60(11):1021–1031, 2022.

[15] Hyun Woo Kim, Chen Zhang, Raphael Reher, Mingxun Wang, Kelsey L Alexander, Louis-Félix Nothias, Yoo Kyong Han, Hyeji Shin, Ki Yong Lee, Kyu Hyeong Lee, et al. Deepsat: learning molecular structures from nuclear magnetic resonance data. Journal of Cheminformatics, 15(1):71, 2023.

[16] Yasuo Kimura, Takashi Mizuno, Hiromitsu Nakajima, and Takashi Hamasaki. Altechromones a and b, new plant growth regulators produced by the fungus, alternaria sp. Bioscience, biotechnology, and biochemistry, 56(10):1664–1665, 1992.

[17] P Konigs, B Rinker, L Maus, M Nieger, J Rheinheimer, and SR Waldvogel. Structural revision and synthesis of altechromone a. Journal of natural products, 73(12):2064–2066, 2010.

[18] Alex Morehead and Jianlin Cheng. Geometry-complete diffusion for 3d molecule generation and optimization. Communications Chemistry, 7(1):150, 2024.

[19] Hiroshi Morita, Yumiko Sato, Kit-Lam Chan, Chee-Yan Choo, Hideji Itokawa, Koichi Takeya, and Jun'ichi Kobayashi. Samoquasine a, a benzoquinazoline alkaloid from the seeds of annona s quamosa. Journal of natural products, 63 (12):1707–1708, 2000.

[20] Laxman D Nandawadekar, Hiba K P, Teena P George, Choppari Thirupathi, Aparna Sahoo, Sidharth Chopra, and D Srinivasa Reddy. Structural revision of a natural tetrahydroquinoxaline-6-carboxylic acid isolated from caulis sinomenii through total synthesis of both the regioisomers. Journal of Natural Products, 88(12):2978–2986, 2025.

[21] Yong-Cheng Ning. Interpretation of Organic Spectra. John Wiley & Sons, 2011.

[22] Ella F Poynton, Jeffrey A van Santen, Matthew Pin, Marla Macias Contreras, Emily McMann, Jonathan Parra, Brandon Showalter, Liana Zaroubi, Katherine R Duncan, and Roger G Linington. The natural products atlas 3.0: extending the database of microbially derived natural products. Nucleic Acids Research, 53(D1):D691–D699, 2025.

[23] Wan-Xian Qi, Wei-Juan Zu, Gang Li, Jing Wang, Li-Juan Lang, Rong Liu, Bei Jiang, and Chao-Jiang Xiao. Coumarin dimers from hydrangea davidii and their antimalarial activities. Magnetic Resonance in Chemistry, 2026.

[24] Adriano Rutz, Maria Sorokina, Jakub Galgonek, Daniel Mietchen, Egon Willighagen, Arnaud Gaudry, James G Graham, Ralf Stephan, Roderic Page, Jiří Vondrášek, et al. The lotus initiative for open knowledge management in natural products research. elife, 11:e70780, 2022.

[25] Ondrej Socha, Zuzana Osifová, and Martin Dracinsky. Nmr-challenge. com: An interactive website with exercises in solving structures from nmr spectra, 2023.

[26] Maria Sorokina, Peter Merseburger, Kohulan Rajan, Mehmet Aziz Yirik, and Christoph Steinbeck. Coconut online: collection of open natural products database. Journal of Cheminformatics, 13(1):2, 2021.

[27] Christoph Steinbeck, Stefan Krause, and Stefan Kuhn. Nmrshiftdb constructing a free chemical information system with open-source components. Journal of chemical information and computer sciences, 43(6):1733–1739, 2003.

[28] Xiaofeng Tan. A transformer based generative chemical language ai model for structural elucidation of organic compounds. Journal of cheminformatics, 17(1):103, 2025.

[29] Cody Timmons and Peter Wipf. Density functional theory calculation of 13c nmr shifts of diazaphenanthrene alkaloids: reinvestigation of the structure of samoquasine a. The Journal of Organic Chemistry, 73(22):9168–9170, 2008.

[30] David Weininger. 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.

[31] Krzysztof Wolinski, James F Hinton, and Peter Pulay. Efficient implementation of the gauge-independent atomic orbital method for nmr chemical shift calculations. Journal of the American Chemical Society, 112(23):8251–8260, 1990.

[32] Ziyu Xiong, Yichi Zhang, Foyez Alauddin, Chu Xin Cheng, Joon Soo An, Mohammad R Seyedsayamdost, and Ellen D Zhong. Atomic diffusion models for small molecule structure elucidation from nmr spectra. arXiv preprint arXiv:2512.03127, 2025.

[33] Fanjie Xu, Wentao Guo, Feng Wang, Lin Yao, Hongshuai Wang, Fujie Tang, Zhifeng Gao, Linfeng Zhang, Weinan E, Zhong-Qun Tian, et al. Toward a unified benchmark and framework for deep learning-based prediction of nuclear magnetic resonance chemical shifts. Nature Computational Science, 5(4):292–300, 2025.

[34] Xi Xue, Hanyu Sun, Jingying Sun, Luc Patiny, Xiangying Liu, Kai Chen, Jingjie Yan, Liangning Li, Xue Liu, Shu Xu, et al. Nmrmind: A transformer-based model enabling the elucidation from multidimensional nmr to structures. Analytical Chemistry, 97(41):22603–22614, 2025.

[35] Chen Yang, Zheng Fang, Hanyu Sun, Fanjie Xu, Hongxin Xiang, Hanyu Gao, Xiangxiang Zeng, Yuqiang Li, Xiaojian Wang, and Jun Xia. A large-scale foundation model enables simulation-to-real adaptation for nuclear magnetic resonance-based molecular structure analysis, 2026. URL https://arxiv.org/abs/2606.20756.

[36] Qingsong Yang, Binglan Wu, Xuwei Liu, Bo Chen, Wei Li, Gen Long, Xin Chen, and Mingjun Xiao. Diffnmr: diffusion models for nuclear magnetic resonance spectra elucidation. Materials Futures, 5(1):015601, 2026.

[37] Zihan Zou, Yujin Zhang, Lijun Liang, Mingzhi Wei, Jiancai Leng, Jun Jiang, Yi Luo, and Wei Hu. A deep learning model for predicting selected organic molecular spectra. Nature Computational Science, 3(11):957–964, 2023.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NMRAgentBCSI.pdf