Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2026, **5**, 1237

Enhancing molecular structure elucidation with reasoning-capable LLMs

Martin Priessner, *[a] Richard J. Lewis, [b] Magnus J. Johansson, [a] Jonathan M. Goodman, [c] Jon Paul Janet [d] and Anna Tomberg *[a]

We introduce a novel workflow integrating reasoning-capable language models with specialized chemical analysis tools to enhance molecular structure determination using nuclear magnetic resonance spectroscopy. Generally, structure elucidation involves generating candidate molecular structures, comparing their predicted spectral features to experimental data, and identifying the best-fitting structure. Our workflow systematically generates diverse molecular candidates through chemical synthesis predictions, regioisomer exploration, and direct spectral-based methods. The language model bridges the gap between quantitative data and chemical insight by evaluating candidates through a reasoning process that analyzes spectral evidence, explains discrepancies, and assesses overall structural plausibility, moving beyond simple numerical error. This LLM-driven reasoning stage proved crucial, increasing correct top-ranked structure identification accuracy by 26.4%. Simulated spectral data with introduced noise artifacts and solvent peaks further highlighted the robustness of our method, showing accuracy improvements by 35.3%. The language model's confidence scores effectively correlated with prediction accuracy, facilitating efficient triage of results. While currently focused on HSQC data, this framework offers a flexible foundation for next-generation structure elucidation tools combining chemical expertise with advanced reasoning capabilities.

Received 14th August 2025
Accepted 19th January 2026

DOI: 10.1039/d5dd00359h

rsc.li/digitaldiscovery

1 Introduction

Molecular structure elucidation remains a critical yet challenging task in chemistry, particularly in drug discovery and high-throughput synthesis, where rapid and precise structural identification is crucial.[1,2] Traditionally, chemists rely on manually analyzing complex spectroscopic data—such as Nuclear Magnetic Resonance (NMR) and Mass Spectrometry (MS)—meticulously comparing observed spectral peaks against hypothesized structures.[3,4] While effective, this approach is both time-intensive and dependent on significant expert knowledge.

Computer-Assisted Structure Elucidation (CASE) systems have emerged as valuable tools to accelerate this process by generating candidate structures based on spectral constraints and ranking them according to predicted *versus* experimental spectral deviations.[5–10] However, these systems often struggle with interpretability, especially in cases where multiple candidate structures align similarly well with spectral data.[11] The lack of clarity regarding how individual spectral features contribute to structural assignments makes it difficult for chemists to critically evaluate and refine these automated proposals, limiting the practical utility of CASE in ambiguous or complex structural elucidation scenarios. Rather than replacing these established tools, our work explores how reasoning-capable LLMs can provide a complementary analysis layer applicable to candidates from any source — whether generated by commercial CASE systems or specialized ML models.

Recent years have witnessed the rise of Large Language Models (LLMs) such as GPT-4,[12] Claude[13] and Gemini,[14,15] which have demonstrated remarkable capabilities across various domains, including chemistry. These models have evolved to become increasingly multimodal,[16,17] capable of analyzing text, images, and structured data simultaneously, making them particularly promising for spectroscopic analysis.

In the field of chemistry, LLMs have demonstrated notable results across various tasks, from property prediction to reaction planning.[18,19] Two distinct approaches have emerged for applying LLMs to chemical problems. The first focuses on domain-specific fine-tuning, exemplified by ChemLLM,[20] which achieves GPT-4-comparable performance across essential

[a]*Medicinal Chemistry, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Pepparedsleden 1, 43183 Mölndal, Sweden. E-mail: martin.priessner@gmail.com; anna.tomberg@astrazeneca.com*

[b]*Department of Medicinal Chemistry, Research & Early Development, Respiratory & Immunology, BioPharmaceuticals R&D, AstraZeneca, Pepparedsleden 1, 43183 Mölndal, Sweden*

[c]*Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK*

[d]*Molecular AI, Discovery Sciences, R&D, AstraZeneca, Pepparedsleden 1, 43183 Mölndal, Sweden*

chemistry tasks through specialized training. The second approach employs multi-agent systems with foundation models, as demonstrated by ChemCrow,[21] which coordinates multiple specialized LLM instances and leverages tool-calling capabilities to interface with chemical software for complex tasks like synthesis planning and reaction prediction. The success of these multi-agent systems in integrating different types of chemical information suggests a promising direction for structure elucidation, where various spectroscopic data must be analyzed in concert.

Particularly relevant to structure elucidation is the recent advancement in LLMs' reasoning capabilities. Advanced prompting techniques, such as Chain of Thought (CoT) reasoning[22] and step-by-step problem-solving approaches, have enabled LLMs to tackle complex tasks by breaking them down into manageable steps. This capability has been further enhanced in recent models like OpenAI's o1/o3, Google's Gemini-Thinking, and DeepSeek's open-source R1 model,[23] which use reinforcement learning to improve their reasoning processes. These models can systematically evaluate problems, consider alternative approaches, and even backtrack when necessary, closely mirroring the analytical process of expert chemists in structure elucidation. While these reasoning-focused models have not yet been extensively applied to chemistry tasks, their ability to provide clear reasoning paths and step-by-step analysis makes them particularly promising for structure determination, where understanding the logic behind structural assignments is crucial.

In our workflow, we specifically harness these reasoning capabilities to perform the critical evaluation step. The LLM is tasked not just with processing data, but with interpreting spectral patterns, identifying inconsistencies, and constructing a logical argument for or against each candidate structure. Our method needs a guess structure and the corresponding spectra (^{1}H, ^{13}C, HSQC, COSY and MS) as input. The structure elucidation process consists of three main stages: candidate generation, spectral analysis, and LLM-driven reasoning.

For candidate generation, we implemented three complementary approaches. First, Chemformer[24–26] generates synthetic analogues *via* retrosynthesis and forward reaction predictions, grounding structural predictions in synthetic feasibility. Mol2Mol[27–30] systematically explores regioisomers and analogues, emphasizing structural diversity. Finally, MultiModalSpectralTransformer (MMST)[31] directly derives candidate structures from spectral patterns, dynamically adapting to the chemical context through on-the-fly fine-tuning.

For the spectral analysis stage, we extend our previous work[32] on atom-specific HSQC peak matching by quantitatively assessing carbon–hydrogen connectivity predictions of the generated analogue molecules against experimental spectra. This approach provides precise structural validation at the atomic level by explicitly identifying structural mismatches for individual carbon–hydrogen bonds. For example, when a predicted methylene group (CH_2) shows a simulated HSQC peak at δ_H 2.7 ppm, δ_C 32 ppm, but the experimental spectrum reveals a significantly different chemical shift environment (δ_H 3.5 ppm, δ_C 45 ppm), this indicates a local structural error—perhaps the carbon is adjacent to an electronegative atom rather than being in a purely aliphatic environment. Such atom-level analysis complements global similarity measures with essential local structural information that can pinpoint specific regions where candidate structures deviate from reality. HSQC peak matching evaluates the similarity of two spectra, by calculating the error between their peaks. These values can be used to rank candidate structures by how well they fit experimental data, referred to as HSQC peak matching.

For the final LLM-driven analysis stage, we employ a two-step process. Initially, Claude 3.5 Sonnet's multimodal capabilities are used to evaluate molecular images alongside spectral data, performing preliminary assessments of structural consistency. Subsequently, DeepSeek R1 applies chemical reasoning to systematically assess all accumulated evidence. It moves beyond the initial HSQC ranking by interpreting the significance of specific spectral features, weighing evidence for and against each structure, generating confidence scores grounded in this analysis, identifying regions of uncertainty, and providing detailed, step-by-step explanations. The workflow, illustrating these stages and their interconnections, is presented in Fig. 1.

2 Results and discussion

To evaluate the effectiveness of our integrated structure elucidation framework, we conducted comprehensive assessments using both experimental and simulated NMR data. Our evaluation focused on the accuracy of structure identification, the interpretability of results, and the practical utility of confidence scores for prioritizing candidates.

2.1 Comparative effectiveness of candidates generation approaches

We first evaluated the individual contributions and comparative strengths of our three candidate-generation approaches to understand how effectively each method could identify the correct molecular structure under controlled conditions using simulated NMR data.

To evaluate these approaches, we used different inputs depending on the method. For Mol2Mol, we used only the target molecule structure as input. For the Chemformer approach, the target molecule was retrosynthetically broken down, and the synthetic pathways then used to produce candidate structures. For MMST, we utilized both simulated NMR spectral data and initial structural information. For this initial comparison, performance was based solely on whether the correct structure was present within the generated pool of candidates (see Methodology for the specific number of considered molecules for each method).

Our evaluation revealed distinct performance patterns that highlight the complementary nature of these approaches across different experimental scenarios (Fig. 2). This evaluation was conducted on our test set of 34 diverse organic molecules, described in the methodology section, for which we had both the correct structures and complete experimental data. Each method demonstrated unique strengths and limitations that

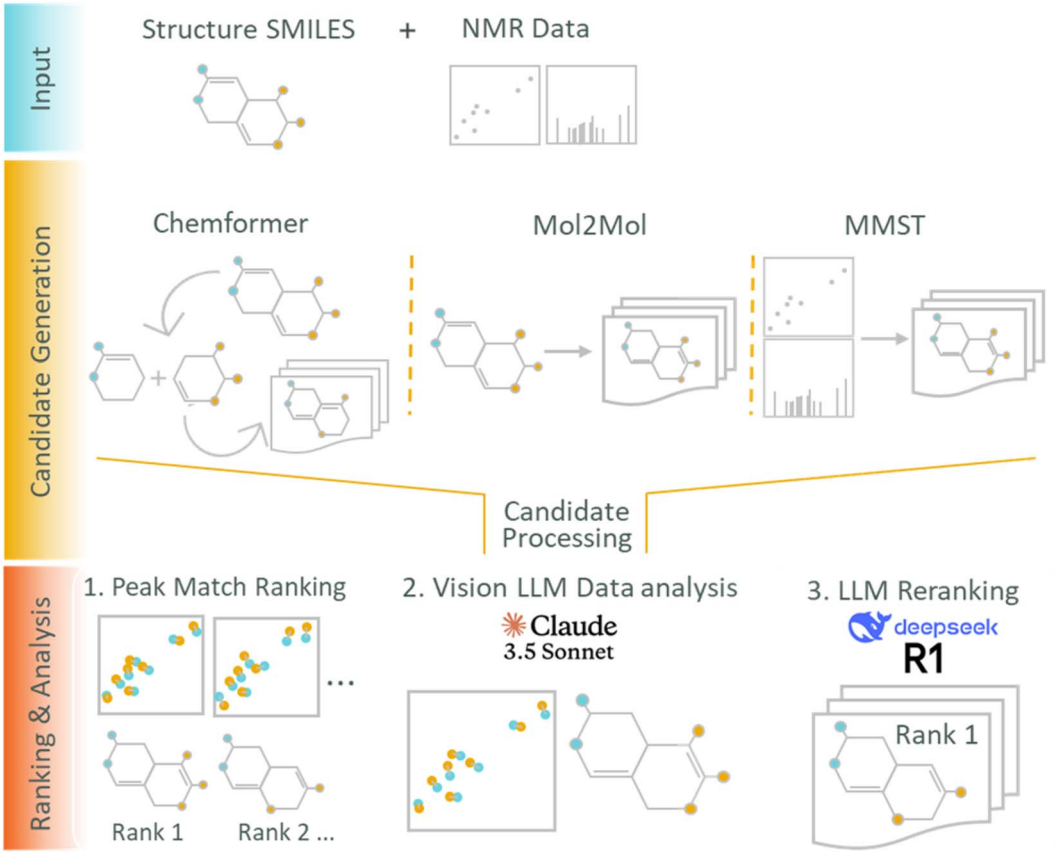

Fig. 1 Structure elucidation pipeline. Input consists of the guess molecule and the corresponding NMR spectra (1D and 2D). Three complementary approaches for candidate generation were implemented. Chemformer uses retrosynthesis & forward predictions, Mol2Mol generates structural analogues, and MMST makes a direct spectral-to-structure prediction. Candidates undergo a three-stage analysis process starting with HSQC peak match ranking, Claude 3.5 Sonnet's visual-spectral analysis, and finally DeepSeek R1's reasoning-enhanced final re-ranking. This integrated approach leverages both specialized chemical tools and advanced LLM reasoning to improve molecular structure identification.

proved valuable in different contexts of the structure elucidation process.

With correct initial structures as input, Chemformer achieved perfect performance (100.0%), successfully recovering all original molecules through its synthesis prediction pathway. The MMST-driven generation likewise reached 100.0% accuracy without requiring structural biasing, demonstrating its ability to derive correct structures directly from spectral features. In contrast, the Mol2Mol approach generated no correct molecules (0.0%) in this scenario, which aligns with its design principles—Mol2Mol intentionally uses the target structure as its starting point for generating structural variations and is not configured to reproduce its input structure.

Performance dynamics shifted dramatically when incorrect regioisomeric structures were provided as starting points. Chemformer completely failed to recover correct structures (0.0%), revealing its inability to reconstruct correct connectivity patterns from incorrect starting hypotheses. In contrast, the MMST-driven approach maintained nearly identical performance (97.1%), demonstrating robustness to input quality. However, we note that MMST is not fully *de novo*—it benefits from approximate structural context to guide its on-the-fly fine-tuning cycle and is therefore best suited for structure verification and regioisomer discrimination rather than true *de novo* elucidation from spectral data alone. The Mol2Mol approach demonstrated modest recovery capability (11.8%), successfully transforming some incorrect structures into correct ones through its systematic structural modification methodology.

These results, obtained from simulated NMR data, highlight the complementary capabilities that justify our multi-pronged approach to structure generation. Chemformer excels when provided with reliable starting structural information but fails with incorrect initial hypotheses; MMST delivers consistent performance regardless of starting conditions, serving as a robust backbone for spectral-based structure prediction; and Mol2Mol, despite its relatively weak performance in direct structure recovery tasks, still adds value by systematically expanding the chemical space with diverse regioisomers and structural variants that maintain molecular weight constraints but explore alternative connectivity patterns. Importantly, these methods are designed for complementary failure modes: MMST is probabilistic and not guaranteed to succeed in all cases, while Chemformer deterministically generates synthetically feasible products given correct chemistry. This distribution of strengths

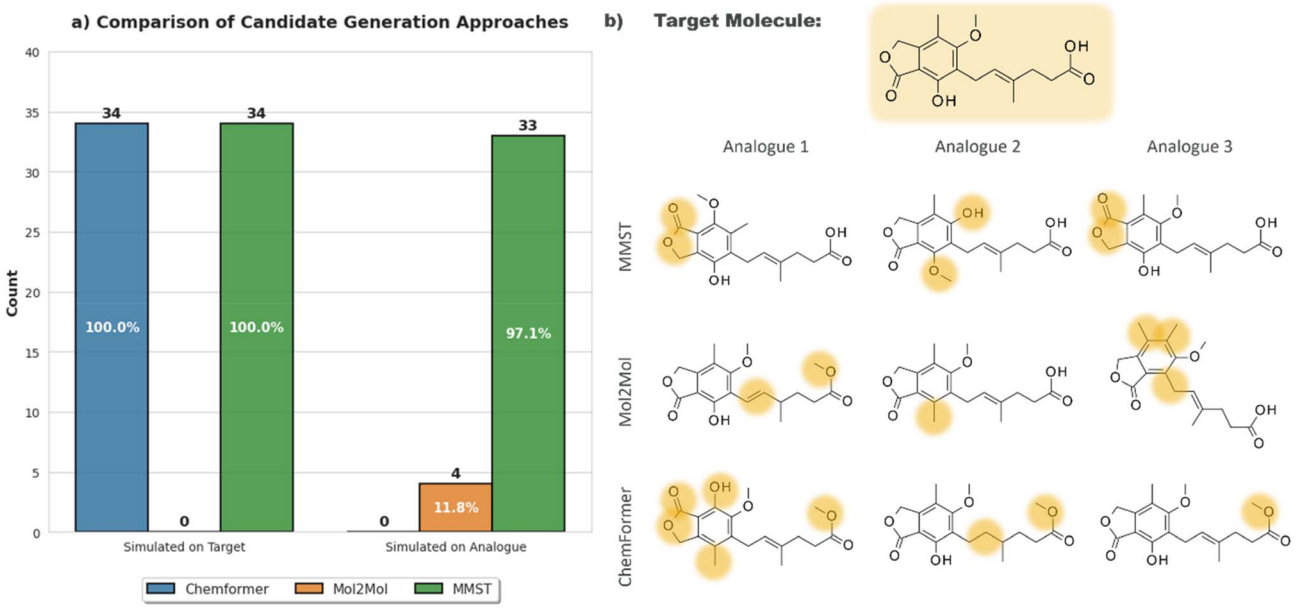

Fig. 2 Structure generation approaches and example outputs. (a) Performance comparison of three approaches across correct initial structure and regioisomeric analogue conditions using simulated NMR data. (b) Representative molecules generated for target molecule, showing MMST (top), Mol2Mol (middle), and Chemformer (bottom) outputs. Structural differences in the generated candidates compared to the target molecule are highlighted in orange, illustrating each method's distinct exploration strategy.

across methods enables our pipeline to address a broader range of structure elucidation challenges than any single approach could manage independently. To capitalize on these diverse generation strategies, the subsequent steps of our workflow operate on a unified candidate pool, created by aggregating the unique molecules generated by Chemformer, Mol2Mol, and MMST. This combined set of structures is then subjected to the ranking and LLM-driven analysis detailed below.

2.2 Ranking and analysis performance comparison

After evaluating the effectiveness of structure generation approaches, we next assessed how LLM-enhanced analysis improved molecular structure identification compared to baseline spectral matching methods. Baseline spectral matching in this context refers to our previously established methodology that ranks candidate molecules by the mathematical error between their simulated HSQC peaks and experimental data, with lower numerical errors indicating better structural matches.[32] It is important to note that our LLM-enhanced approach is constrained by two fundamental limitations: the quality of the candidate pool generated by the HSQC ranking process, and the accuracy of the simulated NMR data. The LLM can only identify the correct structure if it appears within the top-ranked candidates from the initial HSQC matching. As we will show, DeepSeek-R1 effectively converts top-5 HSQC rankings to accurate top-1 predictions, but cannot overcome cases where the correct structure is entirely absent from the candidate pool or where simulated spectral data significantly deviates from experimental realities.

Our experimental design explored five key conditions to test the robustness of both approaches: simulated data with target structures (Sim Target), simulated data with analogue structures (Sim Analogue) where attachment points of functional groups were systematically modified, simulated data with target structures and deliberately introduced spectral artifacts (Sim Target + Noise) including both a DMSO solvent peak (δ 2.52 ppm, 39.57 ppm) and a random noise peak (δ 2.52 ppm, 103.42 ppm), experimental data with target structures (Exp Target), and experimental data with analogue structures (Exp Analogue). This design allowed us to systematically evaluate performance across increasing levels of real-world complexity.

A critical metric for practical structure elucidation is whether the top-ranked candidate is the correct structure, as this determines the system's ability to autonomously identify molecules without expert intervention. Fig. 3 illustrates how effectively DeepSeek R1 converts HSQC top-5 predictions into accurate top-1 rankings—a key capability for automated structure determination.

Under ideal conditions with simulated NMR data and target structures (Sim Target), the baseline HSQC peak matching demonstrated strong performance with a top-1 accuracy of 85.3% (29/34 molecules). Applying DeepSeek-R1's reasoning-based analysis to re-rank the top HSQC candidates further improved this to 94.1% (32/34 molecules), showing modest but meaningful gains even in this favorable scenario. Similar performance patterns were observed with simulated data using analogue initial structures (Sim Analogue), where both approaches maintained comparable accuracy levels (85.3% for both HSQC matching and DeepSeek-R1).

The most significant performance differentials emerged when analyzing imperfect spectral data—a crucial test for real-world applicability. For simulated data containing additional

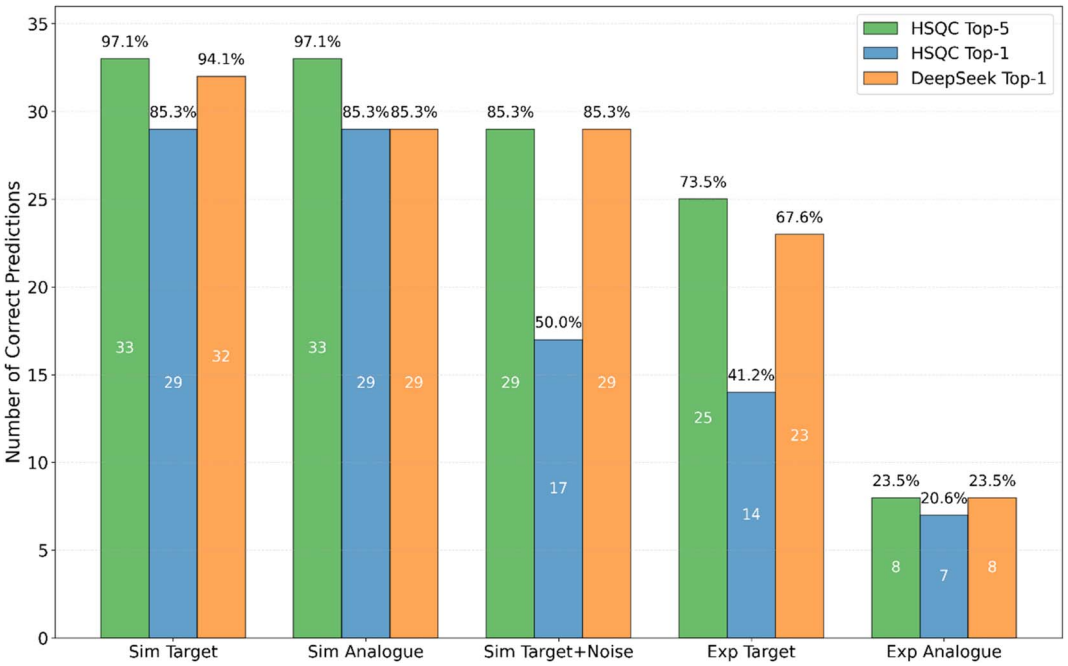

Fig. 3 Comparison of HSQC ranking and DeepSeek-R1 enhanced performance. Comparative accuracy across five experimental conditions (Sim Target, Sim Analogue, Sim Target + Noise, Exp Target, Exp Analogue), showing HSQC top-5 (green), HSQC top-1 (blue), and DeepSeek-R1 top-1 (orange). Numbers indicate correctly identified molecules out of 34. Most significant improvements occur in Sim Target + Noise and Exp Target conditions, where DeepSeek-R1 effectively converts HSQC top-5 candidates to accurate top-1 predictions. "Target" refers to cases where the correct molecular structure was used as the starting point, while "Analogue" indicates cases where regioisomeric variants were used as initial structural hypotheses.

noise peaks (Sim + Noise), the baseline HSQC peak matching approach's top-1 accuracy declined dramatically to 50.0% (17/34 molecules), while DeepSeek-R1 maintained robust performance with 85.3% accuracy (29/34 molecules). This 35.3% improvement highlights the LLM's effectiveness in applying chemical reasoning to navigate the inherent variability. The advantages of LLM enhancement became even more pronounced with experimental NMR data (Exp Target), where the baseline HSQC peak matching approach achieved only 41.2% top-1 accuracy (14/34 molecules), while DeepSeek-R1 substantially improved this to 67.6% (23/34 molecules). This 26.4% increase highlights the LLM's effectiveness in navigating the variability and complexity of real-world spectral data.

In the most challenging scenario—experimental data with analogue initial structures (Exp Analogue)—baseline HSQC peak matching achieved a modest 20.6% top-1 accuracy (7/34 molecules). Even here, the DeepSeek-R1 approach provided improvement, reaching 23.5% (8/34 molecules) accuracy in this particularly difficult context.

Notably, we observed comparable performance across all tested LLM systems, including commercial models such as Claude 3.5 Sonnet, Claude 3.7 Sonnet-Thinking, Gemini 2.0 Flash-Thinking, KIMI 1.5 and o3-mini, as detailed in Fig. S1–S5. The performance consistency across different LLM architectures suggests that structural reasoning capabilities are well-distributed among current state-of-the-art language models. We selected DeepSeek-R1 as our primary model for detailed analysis due to its open-source nature, which offers greater flexibility for future domain-specific fine-tuning. We anticipate that targeted fine-tuning with chemistry-specific data or reinforcement learning could further enhance performance, particularly for the most challenging experimental scenarios.

These results demonstrate that the LLM-enhanced approach adds the most value in precisely the scenarios that challenge traditional methods: when dealing with noisy spectral data, experimental artifacts, or imperfect initial hypotheses. The ability to recover correct structures despite these challenges represents a significant advancement for automated structure elucidation workflows in realistic settings. Furthermore, our reproducibility analysis on a representative subset ($N = 6$ replicates) confirmed that while the system is generally robust, the non-deterministic nature of the LLM can lead to variations in candidate ranking across independent runs (see Methodology (Reproducibility and stochasticity analysis) and SI Table S1). The results indicate that while the model exhibits high reasoning stability in the majority of cases (with 60% showing zero variance), ranking oscillations occur in scenarios with ambiguous spectral data. This confirms that the workflow is not strictly deterministic and that prediction outcomes can be sensitive to token generation probabilities.

To isolate the impact of multimodal integration, we performed an ablation study ($N = 6$) removing the visual analysis provided by Claude 3.5 Sonnet. While visual context was critical for solving complex cases (improving rank in 40%), it reduced performance in one instance by inducing 'over-rationalization,' where the model prioritized qualitative structural narratives

over superior quantitative error scores (see SI Table S2). This specific failure mode—where the AI struggles to appropriately weigh conflicting quantitative and qualitative evidence—contrasts with expert chemical intuition. It underscores that while reasoning models can automate data synthesis, a human-in-the-loop remains essential to adjudicate cases where narrative plausibility diverges from hard spectral metrics.

2.3 Confidence score analysis

To assess the reliability of LLM-based structure elucidation, we analyzed DeepSeek-R1's confidence scores across all experimental conditions. This analysis included a total of 203 prediction instances derived from our 34 molecules across all five experimental conditions (Sim Target, Sim Analogue, Sim Target + Noise, Exp Target, and Exp Analogue), with some molecules having fewer than 5 candidate structures generated.

For each molecule, DeepSeek-R1 assigned confidence scores (0–1) to candidate structures based on spectral evidence analysis and then ranked them from highest to lowest confidence. Fig. 4 shows the distribution of these confidence scores for correct *versus* incorrect structure predictions at each ranking position.

Our analysis of the top-1 candidates revealed that correct structures received a higher mean confidence score (0.92) compared to incorrect ones (0.87). While this difference in mean confidence is relatively small and the distributions for correct and incorrect predictions show substantial overlap (as shown in Fig. 4), we do observe a meaningful pattern in how confidence scores correlate with the overall ranking. The separation between correct and incorrect structures generally decreased progressively through positions 2–5, with position 5 containing only a single correct structure among 188 candidates (0.5%). Notably, confidence scores showed a clear stepwise decrease from position 1 (median ~0.92) to position 5 (median ~0.20), indicating that DeepSeek-R1's confidence values correlate with ranking position. While the overlap in confidence distributions for correct and incorrect structures at position 1 limits the ability to use confidence scores alone as a reliable differentiator, the overall trend suggests that higher confidence scores are generally associated with better candidates. A combination of confidence thresholds and other metrics would likely be needed to effectively automate the verification process and identify cases requiring expert review.

2.4 Case studies: resolving structural ambiguity through LLM reasoning

A critical advantage of LLM-enhanced structure elucidation is the ability to identify the correct molecular structure even when traditional numeric error metrics suggest alternative candidates. We present two case studies that demonstrate how DeepSeek-R1's reasoning capability overcomes the limitations of purely quantitative matching approaches.

2.4.1 Case study 1: reasoning through spectral noise in simulated data. In our first example, we analyzed a molecule where baseline HSQC peak matchings were confounded by added spectral noise. The target molecule contains

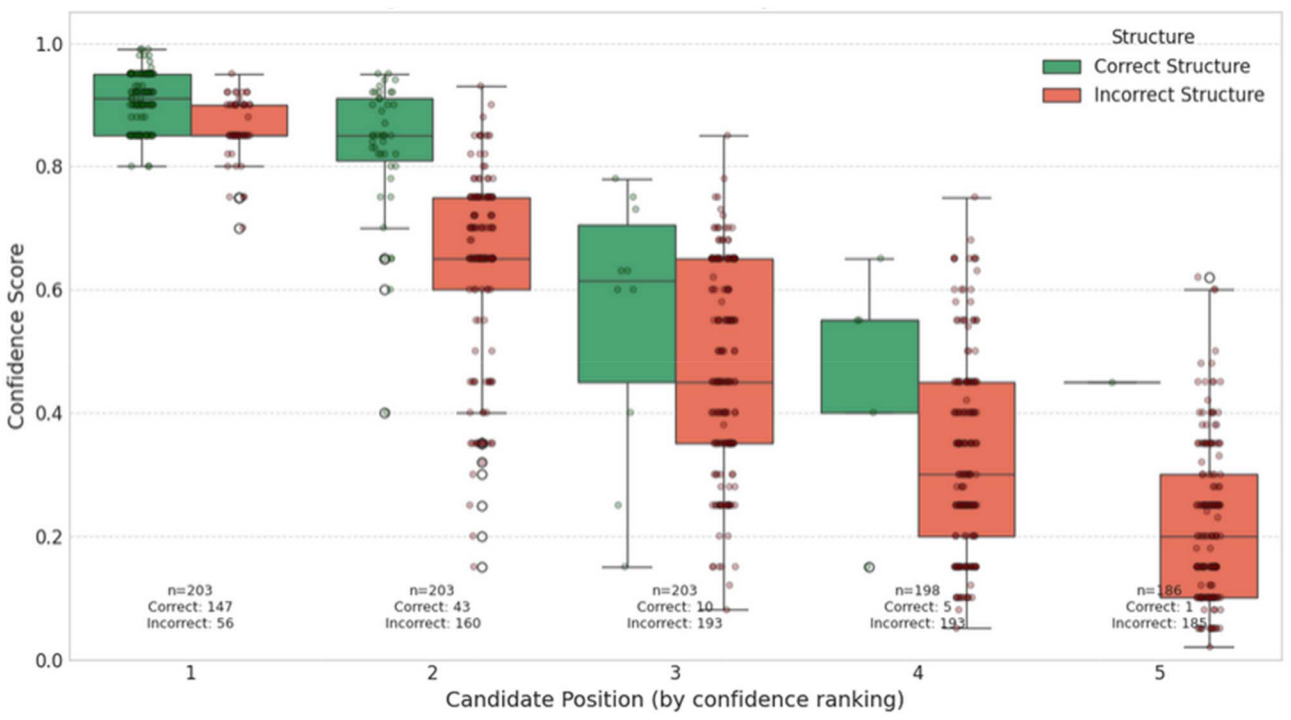

Fig. 4 DeepSeek-R1 confidence scores by candidate position and structural correctness. Box plots showing confidence scores for correct (green) and incorrect (red) structure predictions across positions 1–5, where candidates are ranked by DeepSeek-R1 confidence scores (position 1 = highest confidence). Position 1 represents the highest-confidence candidate. Numbers below boxes indicate correct/incorrect structures at each position. Positions 4–5 have fewer candidates ($n = 198$ and $n = 188$) as some molecules had less than 5 candidates generated.

a tetrahydroindole core with a morpholine substituent, but several structurally similar candidates were proposed (see Fig. 5a). All five candidate structures shared the molecular formula $C_{16}H_{24}N_2O_2$ but differed in ring systems and functional groups.

While baseline HSQC peak matching ranked candidate 5 (the correct structure) last due to the highest numerical error (4.547), DeepSeek-R1 employed a reasoning process that looked beyond this single metric. Instead of being solely driven by the error score, the LLM interpreted the spectral evidence in the context of chemical structure. As its internal analysis reveals, it identified and prioritized diagnostic signals consistent with key structural motifs:

"*Candidate 5 has the highest HSQC error (4.547), but… despite the higher error, it aligns well with morpholine groups. The SMILES includes a morpholine ring (C1CCOCC1), and the HSQC peaks (2.56, 60.69; 3.64, 65.99) match morpholine's CH_2 groups… Comparing all candidates, candidate 5 has the correct substituents (morpholine vs. oxazolidine or oxazinan in others). The seven-membered ring in candidates 2 and 3 introduces strain and shifts that don't fit, whereas candidate 5's six-membered rings (tetrahydroindole and morpholine) have expected shifts… The error might be higher due to a specific atom's deviation but overall structural fit is better.*"

Here, the model reasoned that the presence of the correct morpholine substituent, confirmed by specific HSQC peaks, constituted strong positive evidence. It actively compared this key feature against the incorrect substituents in other candidates. Furthermore, the LLM applied chemical knowledge regarding ring strain to discount the plausibility of candidates with seven-membered rings, noting their predicted shifts would not fit the data as well as the six-membered rings in candidate 5. Critically, it weighed the conflicting evidence—the high numerical error *versus* the strong diagnostic spectral matches and structural plausibility—and hypothesized a justification for the discrepancy, suggesting the error might be localized rather than indicative of an overall poor fit.

This interpretive analysis, prioritizing diagnostic spectral features and chemical plausibility over a potentially misleading numerical score, led DeepSeek-R1 to correctly identify candidate 5 and assign it a high confidence score (0.85), showcasing the power of its reasoning approach in navigating noisy data.

2.4.2 Case study 2: resolving ambiguity in experimental data. This second case study illustrates the LLM's capacity to prioritize diagnostic spectral features and chemical plausibility over purely numerical error scores when dealing with real experimental data. We examine a natural product featuring a benzofuran core (Fig. 5b), where the baseline HSQC peak matching algorithm ranked candidate 1 highest due to the lowest error score (3.245). The correct structure (candidate 3) possessed a slightly higher error (3.657), presenting a scenario where a purely quantitative approach would fail.

DeepSeek-R1, however, employed chemical reasoning to interpret the underlying spectral evidence rather than relying solely on the overall HSQC error. Its analysis involved scrutinizing the details of the peak matches for both candidates and weighing the significance of specific deviations *versus* overall structural consistency. The LLM's internal reasoning highlights this process:

"*Starting with candidate 1: its HSQC error is 3.245, which is the lowest among all. The detailed analysis mentions that most peaks match well except for a significant deviation in atom 5. The problem*

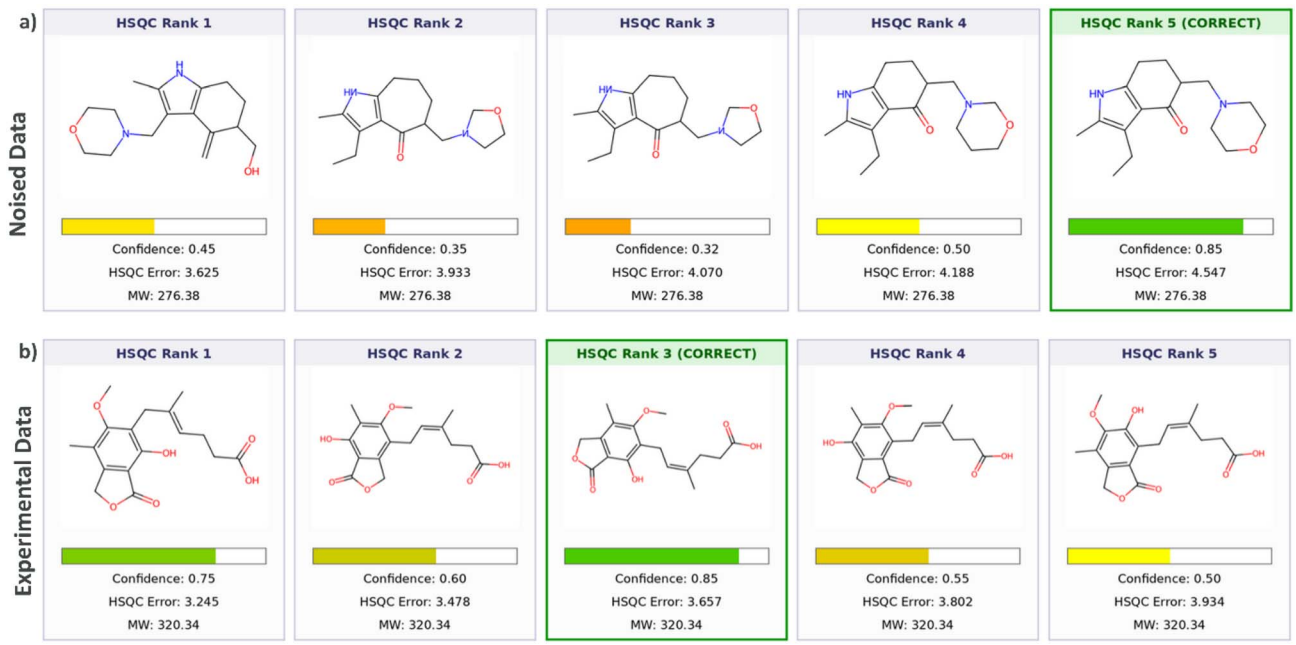

Fig. 5 DeepSeek-R1 structure elucidation performance on simulated and experimental data. (a) Case study 1: simulated data with noise showing correct identification of candidate 5 despite highest HSQC error (4.547). (b) Case study 2: experimental data with correct identification of candidate 3 over the lower-error candidate 1, demonstrating chemically-informed reasoning over numeric metrics alone.

there is a methylene group with an error of 0.504… candidate 3 has an HSQC error of 3.657, slightly higher than candidate 1 but ranked first in the overall analysis. The detailed analysis shows moderate deviations in some atoms but no extreme errors. The methoxy group aligns well, and the core structure fits the data… The final decision weighs structural compatibility over minor error differences. Since the HSQC error difference between candidates 1 and 3 is small, but the overall analysis strongly supports candidate 3's substituent positions, that's likely the right choice."

This excerpt shows the LLM identifying the specific source of error reduction in candidate 1 (a localized deviation in atom 5) while simultaneously recognizing the superior overall spectral consistency and alignment of key functional groups (like the methoxy group) in candidate 3. Crucially, the model reasoned that the small difference in total HSQC error was less significant than the better explanation of the overall spectral pattern provided by candidate 3's specific structural arrangement, particularly concerning substituent positions which often yield diagnostic NMR signals.

By interpreting the spectral data through the lens of chemical structure and prioritizing overall evidence quality over a single numerical metric, DeepSeek-R1 correctly identified candidate 3. It assigned a higher confidence score of 0.85 to candidate 3, compared to 0.75 for candidate 1, reflecting its reasoned assessment that candidate 3 represented the chemically more plausible structure despite the marginally higher HSQC error.

Our case studies reveal crucial advantages of using LLM-enhanced structure elucidation methods. The model demonstrates an ability to look beyond simplistic numeric error metrics, recognizing when these can be misleading due to isolated deviations in specific atoms. DeepSeek-R1 applies chemical reasoning that mimics expert analysis, evaluating structural features based on their expected NMR characteristics across various molecular frameworks. Particularly impressive is how the model prioritizes diagnostic spectral features that strongly indicate specific structural elements, even when contradicted by overall error scores. The transparent reasoning process provides chemists with insight into the structural assignment logic, facilitating verification of the proposed structures. These findings demonstrate that LLM integration brings a sophistication to structure elucidation that traditional scoring systems lack, especially when dealing with complex scenarios involving spectral noise or subtle structural variations.

2.5 Limitations

We acknowledge several boundaries of our current method. First, our test set of 34 molecules, while chemically diverse, is insufficient for statistically meaningful subgroup analyses across molecular classes (*e.g.*, electron-withdrawing groups, polycyclic systems). Performance variation across specific structural features remains an open question requiring larger, balanced datasets.

Second, our validation was limited to molecules of 180–420 Da, representative of typical small-molecule pharmaceutical compounds. Effectiveness for larger molecules (*e.g.*, peptides, natural products >420 Da) with more complex spectral patterns remains untested and would likely require adapted approaches.

Third, we tested robustness using a controlled noise scenario (one DMSO solvent peak plus one artifact peak). Real-world spectra may contain multiple overlapping impurity signals, and systematic evaluation of performance degradation with increasing spectral complexity is needed.

Fourth, our current workflow focuses primarily on HSQC for ranking, with $^{1}H/^{13}C$/COSY as supporting data. Extension to other 2D techniques (*e.g.*, HMBC for long-range correlations) would require adapted prompts and analysis strategies, representing valuable future development.

2.6 Conclusion

In conclusion, our study demonstrates the significant potential of integrating reasoning-capable Large Language Models (LLMs) to perform nuanced evaluation of spectral evidence alongside specialized chemical tools to enhance NMR-based molecular structure elucidation. By combining complementary candidate generation methods—Chemformer's retrosynthesis-forward synthesis predictions, Mol2Mol's regioisomer exploration, and MMST's direct spectral-to-structure predictions—we achieved robust and accurate structure identification across a range of challenging scenarios. Crucially, the LLM-driven analysis stage, where DeepSeek R1 applied chemical reasoning to interpret spectral data, weigh conflicting evidence, and assess structural plausibility, substantially improved top-ranked structural accuracy, particularly with experimental data (67.6% *vs.* 41.2% baseline) and in the presence of noise (85.3% *vs.* 50.0% baseline). Moreover, LLM-generated confidence scores effectively correlated with prediction accuracy, enabling targeted expert intervention only when necessary.

Two detailed case studies illustrated how DeepSeek R1 successfully navigated structural ambiguities beyond numeric error metrics alone, prioritizing chemically-informed reasoning and diagnostic spectral features. This interpretable reasoning process, explicitly articulated by the LLM, facilitates greater trust and practical utility in real-world chemical settings.

Our flexible, reasoning-enhanced workflow thus represents a significant advancement over traditional structure elucidation methods, offering a foundation for next-generation tools where advanced AI reasoning actively interprets complex chemical data, augmenting traditional methods and expert analysis. Importantly, the LLM reasoning layer is modular and tool-agnostic—while demonstrated here with our candidate generation pipeline, the same approach could enhance outputs from commercial CASE systems such as ACD/Structure Elucidator or MestreNova. Future work could explore several promising directions: (1) extension to additional spectral modalities, particularly HMBC for long-range C–H correlations, enabling cross-modal consistency checking across ^{1}H, HSQC, and HMBC data; (2) domain-specific fine-tuning on experimental spectra; (3) training open-source reasoning models with chemistry-specific reinforcement learning to improve spectral

interpretation capabilities; and (4) systematic benchmarking of newer reasoning models as LLM capabilities continue to advance. These avenues could further push the boundaries of automated, interpretable, and accurate molecular structure elucidation.

3 Methodology

3.1 Overview of experimental design and dataset

3.1.1 Experimental dataset. For our validation studies, we utilized a collection of 34 organic molecules with comprehensive spectroscopic data from a previously published dataset.[33] Molecules were selected based on availability of complete spectral data across all required modalities (^{1}H NMR, ^{13}C NMR, HSQC, and COSY). NMR peaks were manually picked based on analysis of the correct molecular structures, following the previously described procedure. No additional preprocessing or filtering was applied.

The dataset encompasses organic molecules with molecular weights ranging from 180–420 Da, representing diverse structural features. The molecules contain various functional groups, heteroatoms (N, O, S, F, Cl, Br), fused and non-fused ring systems, and different degrees of unsaturation. This structural diversity provides a rigorous test set representative of challenges typically encountered in pharmaceutical and synthetic chemistry. Detailed molecular structures and corresponding spectral data are presented in Fig. S6.

3.1.2 Regioisomer generation for methodology testing. To evaluate our pipeline's ability to recover correct structures from imperfect initial hypotheses, we developed a set of regioisomeric analogues for the molecules in our dataset. These analogues were created by systematically modifying the original structures through the disconnection of side branches from core scaffolds and their reconnection at alternative attachment points. This process generated regioisomers with identical molecular formulas and molecular weights but different connectivity patterns (for more details see ref. 31).

These manually curated regioisomers served as challenging "starting guesses" in our experimental design, allowing us to test whether our methodology could successfully identify the correct molecular structure even when initialized with an incorrect but structurally related hypothesis. This aspect of our experimental design addresses real-world structure elucidation scenarios where initial structural proposals may contain inaccuracies in atomic connectivity. The full set of regioisomeric analogues are provided in Fig. S7.

3.1.3 Experimental design. Our validation approach employed an experimental design to systematically evaluate the performance of our structure elucidation pipeline across different data types and methodological variations. As illustrated in Fig. 6, we explored five distinct experimental conditions by varying the following key parameters:

(1) Data type: we utilized both simulated NMR data (generated using our SGNN model) and experimental NMR data acquired under standard laboratory conditions. This allowed us to evaluate the robustness of our approach when transitioning from idealized to real-world spectral data.

(2) Initial structure guess: two different approaches were employed for the initial structure hypotheses:

• Correct target molecule: using the actual structure as the initial guess.

• Regioisomeric analogue: using a regioisomer of the target molecule as the initial guess to simulate scenarios where the initial hypothesis contains structural inaccuracies.

(3) Data augmentation: for all simulated data, we introduced controlled noise to test system robustness. We augmented HSQC spectra with two specific peaks: a DMSO solvent peak (δ_H 2.52 ppm, δ_C 39.57 ppm) and a consistent artifact peak (δ_H 3.25 ppm, δ_C 103.42 ppm). While we placed this artifact at a fixed position across all spectra for experimental control and reproducibility, it represents the type of unpredictable spectral artifact commonly encountered in real-world NMR analysis. This standardized approach enabled us to systematically evaluate how both traditional and LLM-based methods handle well-defined spectral interference, simulating real-world experimental conditions where solvent signals and various artifacts are commonly encountered.

For our main analysis, we employed a molecular weight delta filter of $\Delta = 0.5$ Da, constraining candidates to those with near-identical molecular weights. This focused approach allowed us to evaluate structure elucidation performance in scenarios where the chemical space is more precisely defined, as is often the case in targeted synthesis verification.

For each experimental condition, we analyzed performance using both our baseline HSQC peak matching approach and the LLM-enhanced evaluation process. While we evaluated multiple LLMs (DeepSeek-R1, Claude 3.5 Sonnet, Claude 3.7 Sonnet-Thinking, Gemini 2.0 Flash-Thinking, o3-mini, and Kimi 1.5), our main text focuses primarily on results from DeepSeek-R1 due to its strong performance and open-source nature, which enables potential future fine-tuning for specialized chemical applications. Comprehensive results for all evaluated models are provided in the SI Section S1.

This experimental design allowed us to systematically assess the impact of each variable on elucidation accuracy and identify the conditions under which our LLM-augmented approach provides the greatest advantage over traditional methods, with particular emphasis on realistic scenarios involving experimental data and imperfect structural hypotheses.

3.2 Structure generation pipeline

After defining our experimental dataset and conditions, we implemented a multi-stage structure generation pipeline designed to produce diverse candidate molecules for evaluation. This pipeline employs three complementary approaches to systematically explore the chemical space around a target structure, ensuring comprehensive coverage of potential structural isomers and related molecules (Fig. 1).

3.2.1 Chemformer-based candidate generation. The first approach in our structure generation pipeline leverages computational synthetic chemistry to identify plausible structural analogues. This two-step process begins with retrosynthesis analysis of the target molecule to identify potential

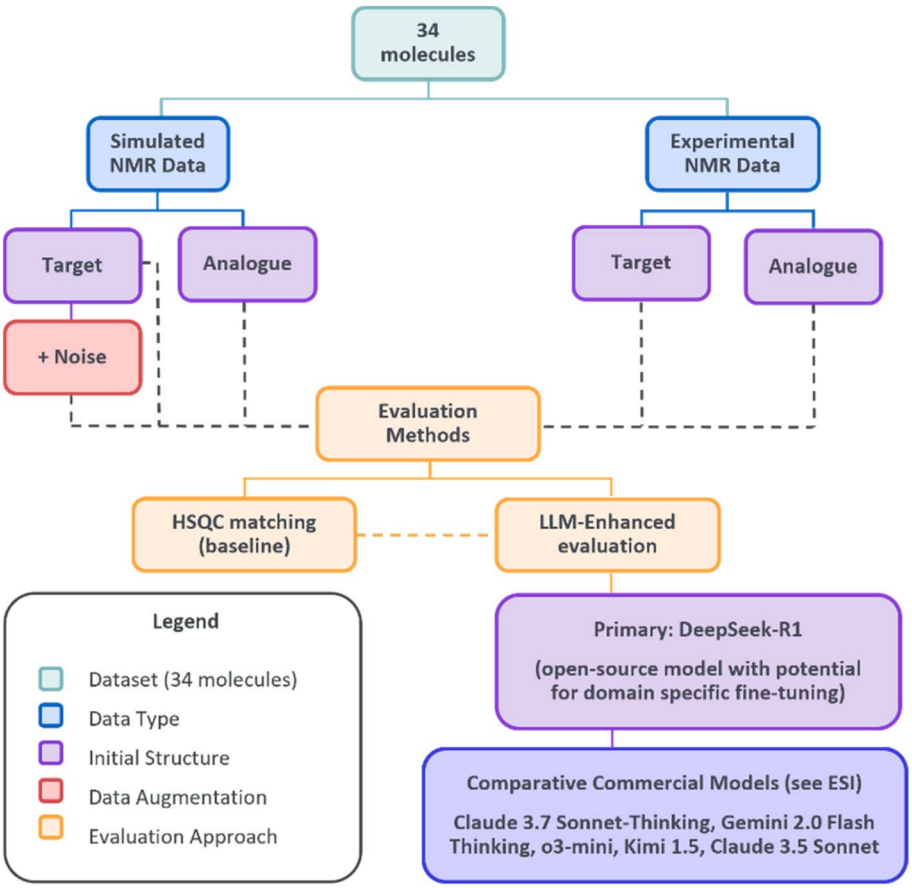

Fig. 6 Experimental design matrix. Our validation approach tested five conditions across 34 molecules, comparing simulated and experimental NMR data with different initial structure hypotheses and noise levels. Each condition was evaluated using both baseline HSQC peak matching and LLM-enhanced approaches, primarily using DeepSeek-R1.

precursors, followed by forward synthesis prediction to generate structurally related molecules.

For retrosynthesis prediction, we employed a Chemformer model[24–26] configured to generate up to 20 retrosynthesis suggestions for the target molecule. The model analyzes the target structure (provided as a SMILES string) and proposes potential disconnections, yielding a diverse set of starting materials from just one synthetic step backwards. These starting materials are canonicalized and filtered for uniqueness to ensure a non-redundant set of precursors for the subsequent forward synthesis step.

In the forward synthesis phase, the same Chemformer model is repurposed to predict possible products that could be formed from each of the identified starting materials. For each starting material, the model generates up to 20 potential product molecules, expanding the search space to include synthetically accessible analogues that maintain chemical similarity to the original target. This approach grounds our structure elucidation in synthetic reality, prioritizing molecules that could plausibly be formed through established chemical transformations.

3.2.2 Mol2Mol analogue generation. The second approach employs the Mol2Mol framework[27–30] to systematically generate structural analogues with controlled diversity. Unlike the synthesis-driven approach, Mol2Mol directly manipulates molecular structures to generate variants while maintaining core scaffold features.

The Mol2Mol model is configured with specific parameters to control the generation process: DELTA_WEIGHT = 0.5 to constrain molecular weight deviation from the target; TANIMOTO_FILTER = 0.2 to ensure a minimum level of structural similarity; NUM_GENERATIONS = 100 to specify the total number of analogues to generate and MAX_TRIALS = 500 to limit generation attempts.

This approach is particularly valuable for generating regioisomers and molecules with alternative functional group arrangements, complementing the synthesis-based approach by exploring regions of chemical space that are structurally plausible.

3.2.3 MMST-based candidate generation. The third and most sophisticated approach employs our Multi-Modal Spectral Transformer (MMST) framework to directly generate candidate structures based on spectral data. The base MMST model used in this workflow was pre-trained following the methodology detailed in our previous work, with the specific exclusion of IR spectra. The training was conducted on a large dataset of 4

million molecules from the ZINC database using simulated ^{1}H NMR, ^{13}C NMR, HSQC, COSY, and MS-derived molecular weight data. The training process consisted of two sequential stages designed to maximize accuracy and robustness:

• Base training: the model was first trained for five epochs to predict SMILES strings from the corresponding spectral data. This stage utilized a cross-entropy loss function for the SMILES tokens.

• Dropout training: following the base training, a second stage introduced a 50% spectral data dropout, where individual spectra were randomly omitted during training. This dropout was applied uniformly across all modalities, including HSQC. This forced the model to become more robust and less reliant on any single data modality, improving generalization to real-world scenarios where spectra may be incomplete.

This entire pre-training was performed on four Nvidia V100 GPUs using the AdamW optimizer and a ReduceLROnPlateau learning rate scheduler. The resulting pre-trained model serves as the starting point for the subsequent improvement cycle.

The MMST workflow begins with an initial test using simulated NMR data of the provided target molecule. We assess the pre-trained model's prediction capability by comparing its output to the known target structure using the RDKit framework, which evaluates molecular graph isomorphism and produces a similarity score between 0 and 1. If this score exceeds our predefined threshold (IC_THRESHOLD = 0.5), the model is deemed sufficiently accurate for the current chemical space and proceeds directly to generating candidate structures using the experimental data.

However, if the accuracy falls below 0.5, indicating the pre-trained model struggles with this particular chemical class, an iterative improvement cycle is initiated before tackling the experimental data:

(1) Generation of similar molecules to the target using the Mol2Mol model with parameters optimized for fine-tuning data generation (MF_GENERATIONS = 200, MF_DELTA_WEIGHT = 100).

(2) Simulation of NMR spectra (^{1}H, ^{13}C, COSY, HSQC) for these generated molecules using the SGNN model.[34]

(3) Fine-tuning of the pre-trained MMST model on this simulated data (NUM_EPOCHS = 15, LEARNING_RATE = 0.0002).

(4) Deployment of the fine-tuned model to generate and sample new candidate structures (MULTINOM_RUNS = 30).

This cycle is repeated up to three times (IMPROVEMENT_CYCLES = 3), allowing the MMST model to progressively refine its predictions based on the specific spectral characteristics of the chemical space surrounding the target molecule. This comprehensive approach generates a diverse pool of candidate molecules that are subsequently ranked using HSQC peak matching and then subjected to in-depth LLM-enhanced analysis using DeepSeek-R1, as described in the following section.

3.2.4 Candidate analysis and ranking. Following the generation of candidate molecules through these three complementary approaches, we perform a unified analysis and ranking process to identify the most promising candidates for detailed evaluation.

For each candidate molecule, we simulate NMR spectra using the SGNN model[34] and perform quantitative HSQC peak matching against the experimental spectra.[32] We prioritize HSQC data for ranking due to its high information content and reliability for structural discrimination.

Beyond overall HSQC matching scores, we calculate per-atom error metrics for each carbon–hydrogen bond, providing a detailed view of structural agreement or discrepancy at the atomic level. These granular error metrics prove particularly valuable for subsequent LLM-driven analysis, enabling focused assessment of specific structural features.

The ranked candidates from all three generation approaches are combined into a unified pool, with the top-ranked molecules (the top 5) selected for detailed evaluation in the subsequent LLM-enhanced analysis stage using DeepSeek-R1.

3.2.5 LLM-enhanced structure evaluation. After generating and ranking candidate structures through the three complementary approaches described in the previous section, our pipeline employs an advanced LLM-enhanced evaluation workflow to determine the most likely correct structure. This phase leverages the reasoning capabilities of state-of-the-art language models to analyze spectroscopic data in conjunction with structural information, providing expert-level evaluation that mimics the approach of experienced spectroscopists (Fig. 7). Complete system prompts for all LLM components are provided in Section S2: individual molecule visual analysis (Fig. S8), comparative multi-molecule analysis and the full DeepSeek R1 reasoning and confidence scoring framework (following Fig. S9).

3.2.6 Structural analysis and spectral comparison. The LLM-enhanced evaluation begins with comprehensive analysis of top-ranked candidate structures from our HSQC peak matching. For each candidate, we first prepare a complete structural representation through two preprocessing steps: (1) generating RDKit visualizations with labeled atomic indices to provide spatial and connectivity information, and (2) converting SMILES to standardized IUPAC nomenclature using the STOUT v2 model.[35,36]

Using these enriched structural representations, Claude 3.5 Sonnet performs a two-stage spectroscopic analysis. In the first stage, each candidate undergoes individual assessment where the LLM analyzes the molecule's visual representation and IUPAC name alongside experimental HSQC data and per-atom error metrics. Through a chain-of-thought process, the LLM evaluates structural-spectral alignment, considering peak patterns, functional group contributions to chemical shifts, and potential structural anomalies explaining spectral discrepancies.

The second stage involves comparative evaluation, where Claude 3.5 Sonnet examines all top five candidates simultaneously. This side-by-side comparison, mimicking the approach of expert spectroscopists, enables the LLM to identify distinguishing spectral features and structural elements that differentiate candidates. The resulting analyses provide chemically-informed insights that highlight specific structural elements supporting or contradicting each candidate structure based on the spectral evidence.

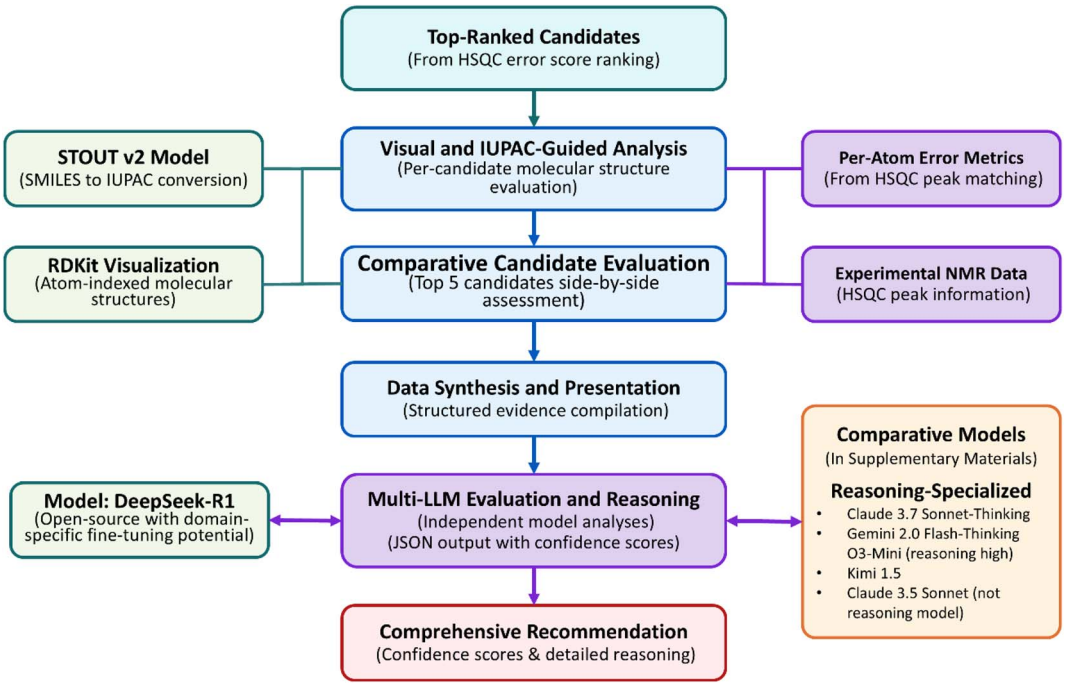

Fig. 7 LLM analysis workflow. Our multi-stage pipeline begins with HSQC-ranked candidates undergoing visual analysis using RDKit visualizations and STOUT v2 nomenclature. After comparative assessment of top candidates, DeepSeek-R1 performs final structure determination. SI includes benchmarks against five additional reasoning-capable LLMs, all providing structured outputs with confidence.

3.2.7 Multi-model LLM analysis workflow. Our workflow implements the LLM-enhanced analysis through three sequential technical steps:

(1) Data synthesis: for each top candidate molecule, our system aggregates comprehensive information including IUPAC name, molecular properties, HSQC error scores, and spectral analyses from earlier evaluation stages.

(2) Evidence analysis: using chemistry-specific chain-of-thought prompting, DeepSeek-R1 systematically evaluates each candidate by analyzing spectral data (particularly HSQC shift comparisons), structural features, and molecular properties. The model justifies its assessments with specific data references and assigns a confidence score (0–1) to each candidate.

(3) Structured output: analysis results are formatted as standardized JSON containing confidence scores, detailed reasoning, and notes on data quality or structural ambiguities. This structured approach enables rigorous evaluation of model performance in molecular structure determination.

Additionally, to contextualize DeepSeek-R1's effectiveness, supplementary evaluations were conducted using other reasoning-focused models (Claude 3.7 Sonnet-Thinking, Gemini 2.0 Flash-Thinking, o3-mini, Kimi 1.5, and standard Claude 3.5 Sonnet). These comparative benchmarks confirmed broadly similar reasoning capabilities among leading LLMs, reinforcing the robustness of our DeepSeek-R1-based primary evaluation approach. Comprehensive benchmarking details are provided in the Fig. S1–S5.

3.3 Evaluation framework

To rigorously assess our structure elucidation pipeline's performance, we implemented a comprehensive evaluation framework that quantifies both prediction accuracy and confidence reliability.

3.3.1 Performance metrics. We evaluated our system's effectiveness primarily through top-1 accuracy—the percentage of cases where the correct structure was identified as the highest-ranked candidate. This metric directly measures the pipeline's ability to autonomously select the correct structure, making it the most relevant measure for high-throughput screening automation applications.

For completeness, we also examined the distribution of correct structures across ranking positions (first through fifth) for both the baseline HSQC peak matching and LLM-enhanced approaches. These distributions, presented as histograms in the Fig. S1–S5, provide insights into how each method shifts the ranking of correct structures, particularly in challenging scenarios with experimental data or added noise.

The comparative analysis between HSQC matching's top-5 accuracy and LLM-enhanced top-1 accuracy proved especially informative, revealing the LLM's ability to effectively re-rank candidates and elevate correct structures to the top position. This relationship demonstrates that while the LLM is constrained by the candidate pool generated through HSQC matching, it substantially improves prioritization within that pool.

3.4 HSQC error score calculation

The HSQC error score, which forms the basis of our initial candidate ranking, was calculated using our previously developed peak matching methodology:

(1) For each candidate molecule, we matched simulated and experimental HSQC peaks using a nearest-neighbor double assignment algorithm with Hungarian distance optimization as described in our previous publication.[32]

(2) This algorithm optimally pairs each experimental peak with its closest corresponding simulated peak, minimizing the overall matching distance across all peak pairs.

(3) For each matched peak pair, we calculated the Euclidean distance between corresponding peaks in the 2D space defined by the carbon and proton chemical shifts.

(4) The overall HSQC error score for a candidate molecule was calculated as the sum of these distances across all matched peak pairs.

This approach ensures optimal peak matching even in cases with complex or overlapping signals. The resulting error score provides a quantitative measure of how well a candidate structure's predicted HSQC spectrum matches the experimental data, with lower scores indicating better matches. These scores were then used to rank candidate molecules for subsequent LLM-based analysis.

3.5 Reproducibility and stochasticity analysis

To assess the consistency of the pipeline given the inherent stochasticity of generative models, we conducted a reproducibility analysis on five representative molecules selected from the dataset. Each molecule was processed across six independent runs ($N = 6$) using identical input data and prompt parameters. We quantified stability by tracking the rank assignment of the ground truth structure across these replicates (see "Analysis of model stochasticity and rank stability" in SI).

3.6 Multimodal ablation study

To quantify the specific contribution of the visual-structural analysis provided by Claude 3.5 Sonnet, we performed an ablation experiment on the same subset of five representative molecules ($N = 6$ replicates, see "Multimodal ablation and the weighting problem" in SI). In this configuration, the prompt provided to DeepSeek R1 was modified to exclude the detailed visual structural assessments and comparative spectral reasoning generated by Claude 3.5 Sonnet. The model received only the candidate metadata (SMILES, IUPAC name, molecular formula, molecular weight) and the raw quantitative HSQC peak matching data (JSON format). This setup allowed us to isolate the impact of the qualitative visual context on the final reasoning process, separating the model's ability to interpret raw spectral data from its ability to leverage pre-digested visual insights.

3.6.1 LLM confidence scores and reliability analysis. A unique aspect of our LLM-enhanced approach is the generation of confidence scores for structural predictions. DeepSeek-R1, our primary model, provides a confidence score (0–1) for each candidate structure based on its analysis of spectral data and structural features.

To evaluate the reliability of these confidence scores, we performed a correlation analysis between:

- LLM-assigned confidence scores for candidate structures.
- Whether the candidate was the correct structure (ground truth).

This analysis, visualized as a confidence–accuracy correlation plot, allowed us to assess whether the LLMs' self-reported confidence levels were reliable indicators of prediction accuracy. High correlation between confidence and accuracy suggests that the system can effectively self-evaluate the reliability of its predictions—a critical feature for practical applications where knowing when to trust automated results is essential.

The sample size for our confidence score analysis included all molecules across experimental conditions ($n = 204$), with slightly smaller samples for positions 4 and 5 ($n = 198$ and $n = 188$, respectively) as some molecules had fewer than 5 candidate structures generated.

Author contributions

M. P. developed the methodology, implemented the agent workflow, conducted experiments, performed formal analysis, data curation, and investigation, and drafted the manuscript. A. T. supervised the project, provided guidance, co-wrote and edited the manuscript. J. P. J. and J. M. G. supervised the project and provided critical revisions to the manuscript. R. J. L. provided expertise on NMR-related topics, reviewed and corrected the manuscript, and provided feedback. M. J. J. reviewed the manuscript and provided feedback.

Conflicts of interest

This research was supported by AstraZeneca. M. P., R. J. L., M. J. J., J. P. J., and A. T. are employees of AstraZeneca and may hold stock or stock options in the company. The remaining author declares no competing financial interests.

Data availability

The code for this study is available at https://github.com/mpriessner/ChemStructLLM_ and archived on Zenodo. The experimental data and model weights, including the fine-tuned MultiModalSpectralTransformer (MMST), Mol2Mol, and Chemformer models, can be downloaded from Zenodo (https://doi.org/10.5281/zenodo.17877209).

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d5dd00359h.

Acknowledgements

We gratefully acknowledge AstraZeneca for their support and funding of the postdoctoral position, instrumental in the success of this research. For the development of the codebase and the preparation of this manuscript, we utilized AI

technologies, including OpenAI's ChatGPT and Claude for code development support and Grammarly for text refinement. These tools served as supplementary aids, providing assistance in editing and optimizing both code and content. The AI suggestions were rigorously reviewed, tested, and selectively implemented by the authors to ensure the integrity and functionality of the code, as well as the accuracy, consistency, and clarity of the manuscript. Despite the involvement of AI technologies, the responsibility for the final content, its validation, and the overall quality of the paper rests solely with the authors.

References

1 S. A. Biyani, *et al.*, Advancement in Organic Synthesis Through High Throughput Experimentation, *Chem.:Methods*, 2021, **1**, 323–339, DOI: 10.1002/cmtd.202100023.
2 S. M. Mennen, *et al.*, The Evolution of High-Throughput Experimentation in Pharmaceutical Development and Perspectives on the Future, *Org. Process Res. Dev.*, 2019, **23**, 1213–1242, DOI: 10.1021/acs.oprd.9b00140.
3 T. Kind and O. Fiehn, Advances in structure elucidation of small molecules using mass spectrometry, *Bioanalytical Reviews*, 2010, **2**, 23, DOI: 10.1007/s12566-010-0015-9.
4 D. A. Dias, *et al.*, Current and Future Perspectives on the Structural Identification of Small Molecules in Biological Systems, *Metabolites*, 2016, **6**, 46, DOI: 10.3390/metabo6040046.
5 G. Lee, Towards the automatic analysis of 1H NMR spectra: Part 2. Accurate integrals and stoichiometry, *Magn. Reson. Chem.*, 2001, **39**, 194–202, DOI: 10.1002/mrc.822.
6 D. C. Burns, *et al.*, The role of computer-assisted structure elucidation (CASE) programs in the structure elucidation of complex natural products, *Nat. Prod. Rep.*, 2019, **36**, 919–933, DOI: 10.1039/C9NP00007K.
7 M. Valli, *et al.*, Computational methods for NMR and MS for structure elucidation II: Database resources and advanced methods, *Physical Sciences Reviews*, 2019, **4**, 20180167, DOI: 10.1515/psr-2018-0167.
8 M. Valli, *et al.*, Computational methods for NMR and MS for structure elucidation I: software for basic NMR, *Volume 1 Fundamental Concepts*, ed. F. Ntie-Kang, De Gruyter, Berlin, Boston, 2020, pp. 177–204, DOI: 10.1515/9783110579352-008.
9 M. E. Elyashberg, A. J. Williams and G. E. Martin, Computer-assisted structure verification and elucidation tools in NMR-based structure elucidation, *Prog. Nucl. Magn. Reson. Spectrosc.*, 2008, **53**, 1–104, DOI: 10.1016/j.pnmrs.2007.04.003.
10 M. Elyashberg, *et al.*, Computer-assisted methods for molecular structure elucidation: Realizing a spectroscopist's dream, *J. Cheminf.*, 2009, **1**, 1–26, DOI: 10.1186/1758-2946-1-3.
11 J. Bellenger, *et al.*, An Automated Purification Workflow Coupled with Material-Sparing High-Throughput 1H NMR for Parallel Medicinal Chemistry, *ACS Med. Chem. Lett.*, 2024, **15**(9), 1635–1644, DOI: 10.1021/acsmedchemlett.4c00245.
12 OpenAI, *et al.*, GPT-4 Technical Report, *arXiv*, 2023, preprint, arXiv:2303.08774, DOI: 10.48550/arXiv.2303.08774.
13 Anthropic, *The Claude 3 Model Family: Opus, Sonnet, Haiku Anthropic*, https://platform.claude.com/docs/en/home, accessed Dec. 9, 2025.
14 G. Team, *et al.*, Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, *arXiv*, 2024, preprint, arXiv:2403.05530, DOI: 10.48550/arXiv.2403.05530.
15 G. Team, *et al.*, Gemini: A Family of Highly Capable Multimodal Models, *arXiv*, 2023, preprint, arXiv:2312.11805, DOI: 10.48550/arXiv.2312.11805.
16 Z. Yang, *et al.*, The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision), *arXiv*, 2023, preprint, arXiv:2309.17421, DOI: 10.48550/arXiv.2309.17421.
17 X. Zhang, *et al.*, GPT-4V(ision) as a Generalist Evaluator for Vision-Language Tasks, *arXiv*, 2023, preprint, arXiv:2311.01361, DOI: 10.48550/arXiv.2311.01361.
18 K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero and B. Smit, Leveraging Large Language Models for Predictive Chemistry, *Nat. Mach. Intell.*, 2024, **6**, 161–169, DOI: 10.1038/s42256-023-00788-1.
19 T. Guo, *et al.*, What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks, *arXiv*, 2023, preprint, arXiv:2305.18365, DOI: 10.48550/arXiv.2305.18365.
20 D. Zhang, *et al.*, ChemLLM: A Chemical Large Language Model, *arXiv*, 2024, preprint, arXiv:2305.18365, DOI: 10.48550/arXiv.2305.18365.
21 A. M. Bran, *et al.*, Augmenting large language models with chemistry tools, *Nat. Mach. Intell.*, 2024, **6**, 525–535, DOI: 10.1038/s42256-024-00832-8.
22 J. Wei, *et al.*, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, *Advances in Neural Information Processing Systems*, ed. S. Koyejo, *et al.*, Curran Associates, Inc., 2022, vol. 35, pp. 24824–24837.
23 D. Guo, *et al.*, DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning, *Nature*, 2025, **645**, 633–638, DOI: 10.1038/s41586-025-09422-z.
24 R. Irwin, *et al.*, Chemformer: a pre-trained transformer for computational chemistry, *Machine Learning: Science and Technology*, 2022, **3**, 015022, DOI: 10.1088/2632-2153/ac3ffb.
25 A. M. Westerlund, *et al.*, Constrained synthesis planning with disconnection-aware transformer and multi-objective search, *ChemRxiv*, 2024, preprint, DOI: 10.26434/CHEMRXIV-2024-C77P4.
26 A. M. Westerlund, *et al.*, Do Chemformers Dream of Organic Matter? Evaluating a Transformer Model for Multistep Retrosynthesis, *J. Chem. Inf. Model.*, 2024, **64**(8), 3021–3033, DOI: 10.1021/acs.jcim.3c01685.
27 J. He, *et al.*, Transformer-based molecular optimization beyond matched molecular pairs, *J. Cheminf.*, 2022, **14**, 18, DOI: 10.1186/s13321-022-00599-3.

28 A. Tibo, J. He, J. P. Janet, *et al.*, Exhaustive local chemical space exploration using a transformer model, *Nat. Commun.*, 2024, **15**, 7315, DOI: 10.1038/s41467-024-51672-4.
29 J. He, H. You, E. Sandström, *et al.*, Molecular optimization by capturing chemist's intuition using deep neural networks, *J. Cheminf.*, 2021, **13**, 26, DOI: 10.1186/s13321-021-00497-0.
30 H. H. Loeffler, J. He, A. Tibo, *et al.*, Reinvent 4: Modern AI–driven generative molecule design, *J. Cheminf.*, 2024, **16**, 20, DOI: 10.1186/s13321-024-00812-5.
31 M. Priessner, *et al.*, Advancing Structure Elucidation with a Flexible Multi-Spectral AI Model, *Angew. Chem., Int. Ed.*, 2025, e17611, DOI: 10.1002/anie.202517611.
32 M. Priessner, *et al.*, HSQC Spectra Simulation and Matching for Molecular Identification, *J. Chem. Inf. Model.*, 2023, **64**, 34, DOI: 10.1021/acs.jcim.3c01735.
33 J. B. Rowlands, *et al.*, Towards automatically verifying chemical structures: the powerful combination of 1H NMR and IR spectroscopy, *Chem. Sci.*, 2025, **16**, 21590–21599, DOI: 10.1039/D5SC06866E.
34 J. Han, *et al.*, Scalable graph neural network for NMR chemical shift prediction, *Phys. Chem. Chem. Phys.*, 2022, **24**, 26870–26878, DOI: 10.1039/D2CP04542G.
35 K. Rajan, A. Zielesny and C. Steinbeck, STOUT: SMILES to IUPAC names using neural machine translation, *J. Cheminf.*, 2021, **13**, 34, DOI: 10.1186/s13321-021-00512-4.
36 K. Rajan, A. Zielesny and C. Steinbeck, STOUT V2.0: SMILES to IUPAC name conversion using transformer models, *J. Cheminf.*, 2024, **16**, 146, DOI: 10.1186/s13321-024-00941-x.