ANTHROP\C

How Claude performs on NMR prediction and structure elucidation

Assessing Claude Opus 4.7 against ChemDraw 25.0.2 and MestReNova 17.0.0

Published

June 5, 2026

Author

David Kamber

Summary

One of Anthropic's goals is to accelerate scientific discovery. This involves training Claude to take on researchers' more time-intensive tasks and to give scientists capabilities they didn't previously have. A bench chemist's day is full of bottlenecks: literature searches, experimental writing, data interpretation, and spectral analysis. Spectral analysis, in particular, is one of the most time-consuming steps in synthetic chemistry, and improving its efficiency enables researchers to spend their time elsewhere. For this paper, we tested whether a frontier general-purpose AI model can perform spectral analysis as well as the dedicated Nuclear Magnetic Resonance (NMR) software chemists rely on today. We measured three Claude models (Opus 4.7, Opus 4.6, Sonnet 4.6) against two dedicated software (ChemDraw and MestReNova) on 20 compounds curated from synthetic chemistry preprints. We found that for routine data prediction Opus 4.7 is now as good as or better than ChemDraw and MestReNova on average. Dedicated structure-elucidation software has existed for decades, but it typically requires 2D NMR (a spectrum with two axes, and the output is a contour map rather than a row of peaks), specialized training, and licensed tools. We also found that Claude can work the problem in reverse, proposing a structure from NMR data alone. Claude works from exactly what a chemist would paste into a chat: the standard readouts from a routine mass spec and NMR run, with no setup.

Introduction

Chemical reactions rarely produce one product cleanly, and confirming that the resultant product is what the chemist had intended consumes the bulk of characterization time. For chemists, the standard confirmation of products from a chemical reaction is an NMR spectrum: a list of peaks, one per chemically distinct hydrogen or carbon, each shifted by its surrounding atoms.

Software helps interpret these peaks in two directions. Forward prediction is the more routine of these: draw the expected structure, predict its spectrum, and compare against what was measured. ChemDraw and MestReNova are two tools on essentially every chemist's desktop that are commonly used for forward prediction of NMR spectra.

The harder direction is inverse prediction (structure elucidation): given a spectrum, determine the structure. This requires expert reasoning about which fragments are present and how they connect. While ChemDraw does not have this capability at all, MestReNova helps assign peaks to a known structure but does not generate candidates from a peak list.

We assessed Claude in both directions: 20 compounds for forward shift prediction against ChemDraw and MestReNova, and 15 for inverse structure elucidation. The forward task tests whether a general-purpose model is competitive with the dedicated predictors chemists already use; the inverse task tests whether it can take on the expert-reasoning step those tools were not designed to perform.

We found that:

- ^{1}H NMR *shift prediction* (Figure 4) and *multiplicity reporting* (Figure 7): Opus 4.7 is the most accurate tool tested[1]

[1] Chemical shift (ppm): where a signal sits in the spectrum, reflecting an atom's chemical environment; predicted for ^{1}H and ^{13}C.

- ^{13}C NMR *shift prediction.* Opus 4.7's accuracy is comparable to MestReNova's (Figure 2).
- *J-coupling accuracy* (Figure 8): Claude models uniformly outperform ChemDraw and MestReNova.
- *Peak coverage.* ChemDraw's main strength is breadth: it maintains the widest peak coverage across both nuclei (Figure 2), even where its coupling accuracy lags.

On the inverse task (NMR/HRMS → structure), Opus 4.7 was asked to propose candidate SMILES for 15 published targets drawn from the same ChemRxiv preprint pool, spanning a deliberate range of scaffold density. At the simpler end: anilino-chloropyrimidines, a Boc-N-aryl maleimide, an oxazolidinone-iodobenzamide, and a triethylsilyl sulfonamide. At the more complex end: fused chloropyridazines, an N-acyl oxazinane, spirocyclic ketones with phenacyl or acetyl pendants, an α-aryl triethylsilyl sulfonamide, and a gem-disilyl sulfonamide.

From 1D NMR and HRMS alone, Opus 4.7 recovered the published target on every attempt for the simpler scaffolds. For the structurally denser targets, supplementing the spectra with the starting-material SMILES, with no other reaction context (no reagents, conditions, mechanism, or product class), was sufficient for Opus 4.7 to reach the correct structure on most or all attempts across the set. This extends automated structure elucidation from 1D data into the scaffold-density regime where chemists currently default to 2D experiments or manual interpretation, using only information already in hand at the point of a confirmatory NMR. Opus 4.7 spans both forward and inverse: it leads or ties for the lead on shift accuracy, multiplicity, and J-coupling, and it makes inverse elucidation tractable in a setting where ChemDraw and MestReNova cannot operate.

Evaluation

To set up this assessment, we drew compounds from ChemRxiv synthetic chemistry preprints and selected them by hand per the Section 3 protocol: a chemist read each preprint, kept compounds whose reported NMR and HRMS data were complete and self-consistent, and transcribed the peak lists manually (section 3). We produced two sets of compounds: 20 for forward shift/coupling prediction (4 scaffold classes × 5 compounds; Figure 1) and 15 for inverse structure elucidation. The four forward-task scaffolds are

---

Multiplicity: a ^{1}H signal's splitting pattern (singlet, doublet, triplet, etc.), set by its number of neighboring hydrogens.

J-coupling: the coupling constant (J, in Hz), the spacing between split sub-peaks, measuring how strongly two nuclei interact through bonds.

Peak coverage: the fraction of observed atoms a tool actually predicts.

referred to as P1 (chloropyridazines), P2 (Boc-N-aryl maleimides + N-Boc ynamides), P3 (spiroketones), and P4 (α-silyl methanesulfonamides) below.

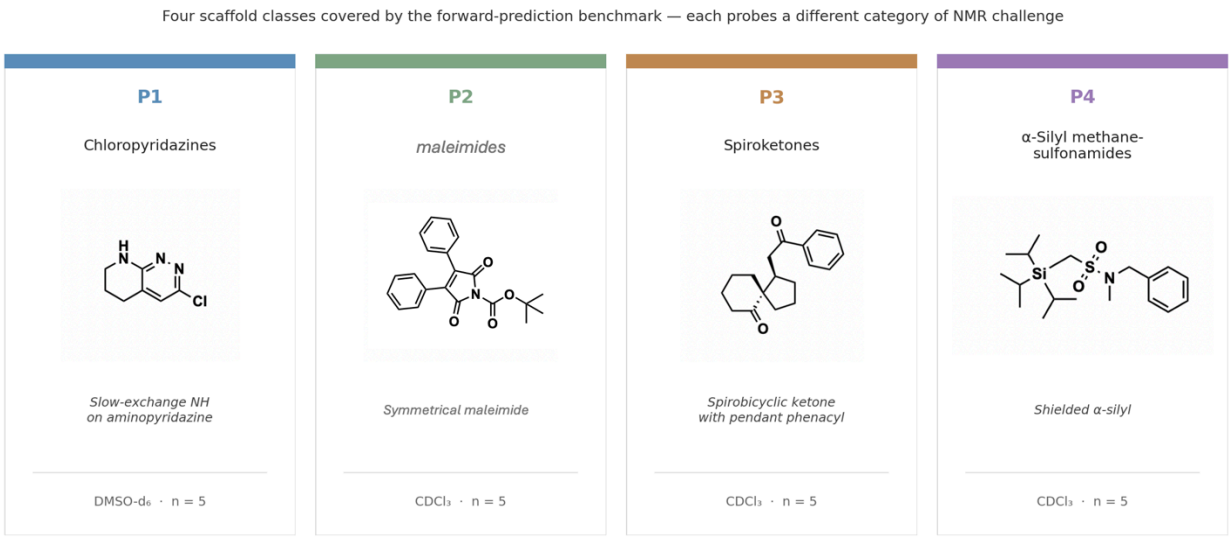

Figure 1. Four scaffold classes covered by the forward-prediction benchmark. Each probes a different category of NMR challenge. P1 chloropyridazines have a slow-exchange NH on aminopyridazine in DMSO-d_6; P2 Boc-N-aryl maleimides and N-Boc ynamides exercise α-vinyl-imide carbonyls and the rare ynamide α/β-carbon pair; P3 spiroketones are spirobicyclic ketones with phenacyl or acetyl pendants and diastereotopic CH_2; P4 α-silyl methanesulfonamides have shielded silicon-α carbons. Five compounds per class, n = 20 total.

Selection protocol

To avoid selection bias, we chose and locked the compounds before generating any predictions from Claude or the comparison tools. From ChemRxiv synthetic chemistry preprints with full ^{1}H and ^{13}C characterization in the supporting information (SI), we took the first fully characterized novel compounds per preprint, excluding compounds reported as mixtures of rotamers, extracting SMILES, ^{1}H and ^{13}C peak lists, and NMR solvent. Forward (n=20) and inverse (n=15) sets were drawn separately under the same protocol.

Forward prediction: structure → NMR

Each tool was given a SMILES and asked to predict ^{1}H and ^{13}C shifts—with multiplicity and scalar spin-spin coupling constants (J, in Hz) where applicable—in the solvent of the original report. The three Claude models were queried three times per compound to characterize run-to-run variability; ChemDraw and MestReNova are deterministic, so were run once per compound. Predicted peaks were matched to experimental atoms one-to-one by minimum $|\Delta\delta|$ (Hungarian assignment), with ^{1}H multiplets reported as ranges wider than

0.3 ppm were dropped from the shift-error calculation but retained for multiplicity. We report MAE, RMSE, and median |Δδ| against fixed denominators of 401 ^{1}H and 225 ^{13}C atoms; the headline tolerance metric is the fraction of atoms predicted within ±0.20 ppm (^{1}H) or ±1.0 ppm (^{13}C), and the per-compound win rate counts compounds where each tool gives the lowest MAE, computed once per LLM replicate and reported with its min–max range.

On ^{1}H, Opus 4.7 has an MAE of 0.079 ppm, the lowest of any tool tested, and ranks first on the fraction of atoms within the ±0.20 ppm tolerance window (Figures 2–4). On ^{13}C, Opus 4.7 (1.37 ppm MAE) and MestReNova (1.48 ppm) are comparable across the 20-compound set; the remaining tools hold the same rank order as on ^{1}H. The per-compound win rate makes the same point at the sample level: Opus 4.7-wins on most compounds on both nuclei, while MestReNova's wins are split across chloropyridazines and spiroketones, with spiroketones the largest single bucket on ^{13}C and the only scaffold on which MestReNova wins any ^{1}H compounds (Figure 4, bottom). Note that MestReNova's ^{1}H MAE is computed on partial coverage (267/401 atoms; the >0.3 ppm experimental multiplets it skips are dropped from its score), so it is not directly comparable to fully covered tools.

The slow-exchange NH proton sits in a narrow 6.8–7.9 ppm window experimentally; it drifts upfield in Opus 4.7, scatters across several ppm in Opus 4.6, and is misplaced into the 10–13 ppm regime by Sonnet 4.6. The aromatic ring carbons are predicted systematically low by both Opus models, while Sonnet 4.6's errors on those atoms are large but undirected. Outside chloropyridazines, the bias picture is benign: Boc-N-aryl maleimides, spiroketones, and α-silyl methanesulfonamides give Opus 4.7 either the lowest scaffold MAE or the lowest residual spread (Figures 5–6). One common-mode artifact is worth flagging: every tool predicts carbonyl carbons slightly low, so a two-tool sanity check that requires agreement on δ(C=O) will not detect it.

Beyond shift accuracy, Claude models hold their edge on coupling-pattern reporting. Opus 4.7 returns the highest exact-multiplicity rate of any tool under one-way leniency (Figure 7); ChemDraw's low exact rate is a scoring artifact rather than a capability gap, since it routinely emits concrete multiplicity calls where the experimentalist reported only 'm'. On J-values (Figure 8), all three Claude models cluster around 0.5 Hz mean |ΔJ| and reach 80–84% of pairs within ±0.5 Hz; the classical tools sit at 1.9–2.0 Hz mean |ΔJ| and reach only 26–35% within ±0.5 Hz, with ChemDraw's template defaults inflating its error on aromatic systems—most visibly the 12.4 Hz geminal coupling, which appears in 5 of its 31 J calls; aromatic vicinal couplings are emitted near 7.0–7.1 Hz.

Across replicates, Opus 4.7 is the most stable model: its run-to-run MAE varies by less than the gap to the next tool (Figure 2), while the per-compound winner shifts often enough across the 20-compound set that single-replicate assessments would substantially misstate win rates (Figure 4, bottom).

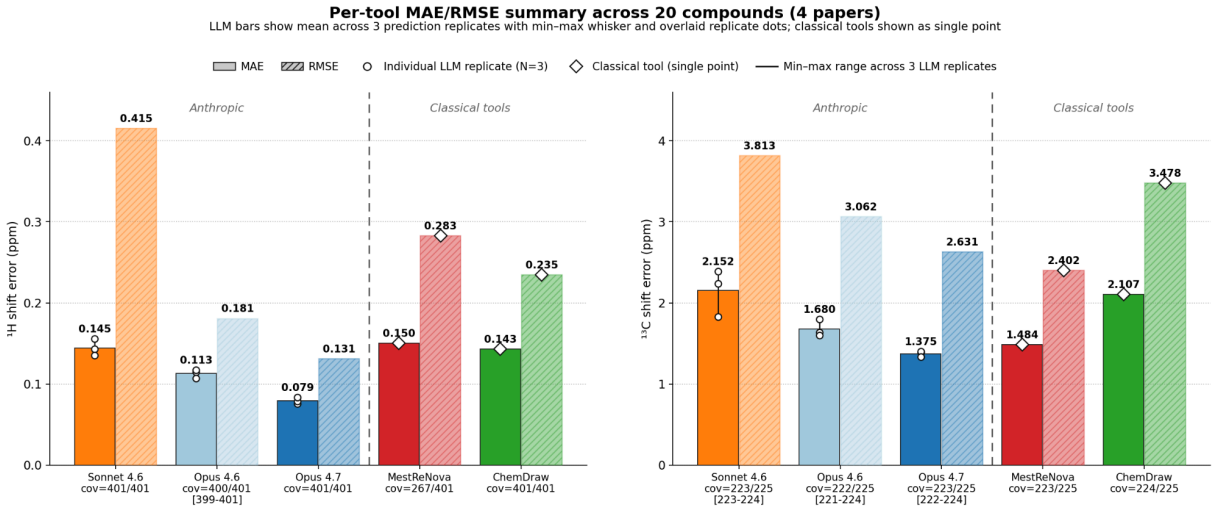

Figure 2. Per-tool MAE (solid) and RMSE (hatched) for ¹H (left) and ¹³C (right) shift errors across 20 compounds for forward prediction, with coverage shown beneath each tool. Claude bars: mean across three replicates with min–max range and overlaid replicate dots. Classical tools: single-point predictions (no range).

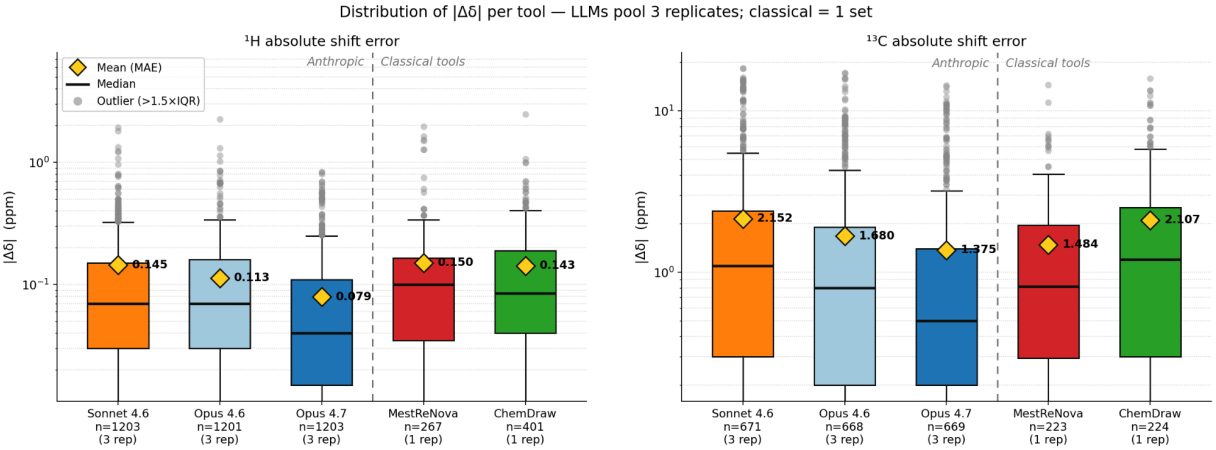

Figure 3. Distribution of absolute shift error |Δδ| per tool across all matched pairs, shown for ¹H (left) and ¹³C (right) on a log scale. Box = IQR; whiskers = 1.5×IQR; small grey circles = outliers; gold diamond = mean; thick black bar = median. Claude models pool 3 replicates; classical tools are a single observation set.

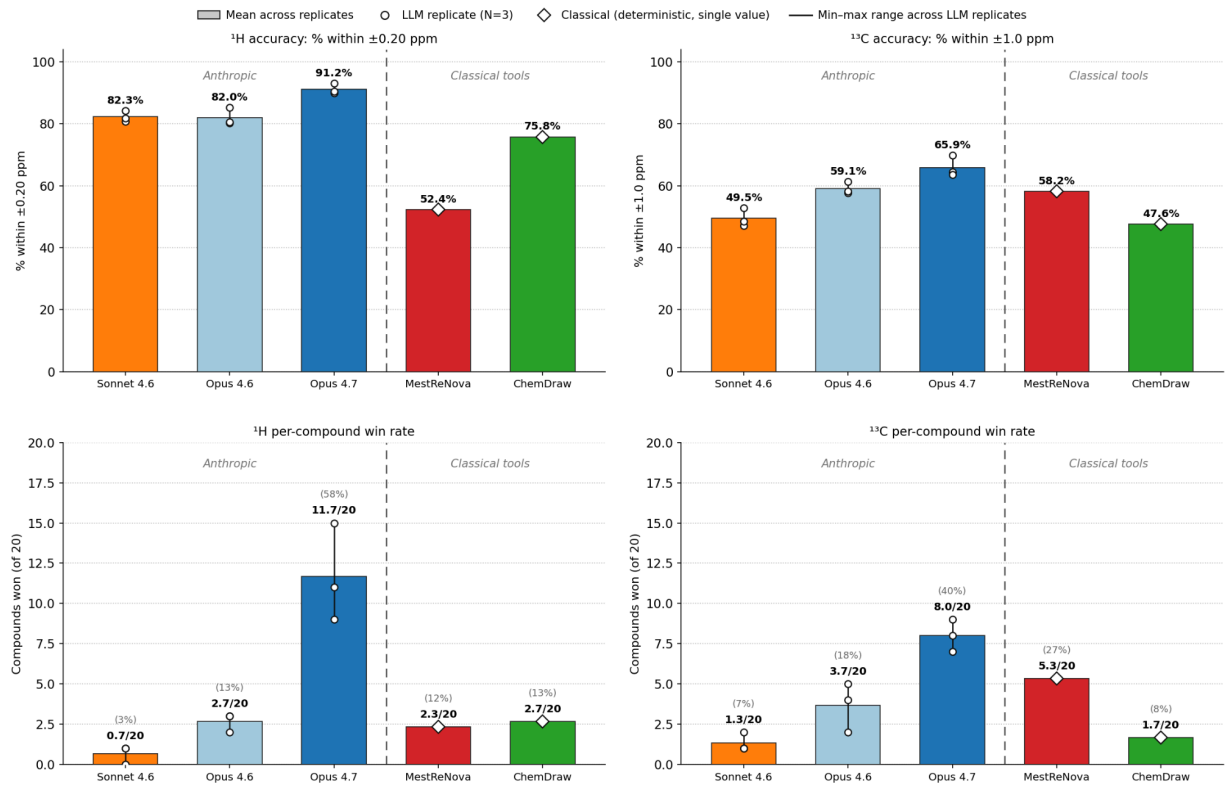

Figure 4. Top: % of experimental atoms within ±0.20 ppm (¹H, left) and ±1.0 ppm (¹³C, right). Bottom: per-compound win rate (compounds where the tool had the lowest per-compound MAE, out of 20). Claude bars: mean across three replicates with min–max range; classical tools: single-point predictions.

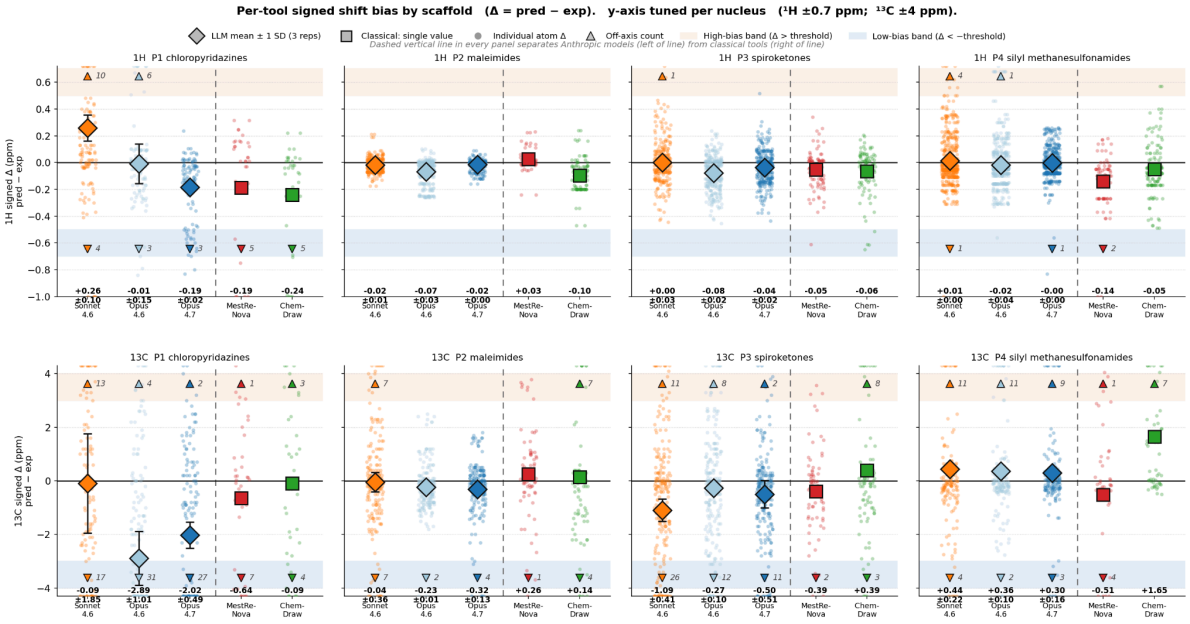

Figure 5. Per-tool signed shift bias by scaffold (Δ = pred − exp). One panel per (nucleus, scaffold) combination. Translucent dots are individual matched atoms; diamonds (Claude models) and squares (classical tools) mark the mean signed Δ, with error bars showing ±1 SD across the three LLM replicates. Mean ± SD values are printed below each tool. Colored bands flag bias regions (¹H |Δ| > 0.5 ppm; ¹³C |Δ| > 3.0 ppm); points with |Δ| beyond the displayed y-axis range (¹H 0.7 ppm; ¹³C 4.0 ppm) are summarized as triangles at the panel edges with their counts. For LLMs, error bars and ±SD labels show the spread of the per-replicate mean across 3 runs; classical tools are deterministic single-point predictions and carry no replicate SD.

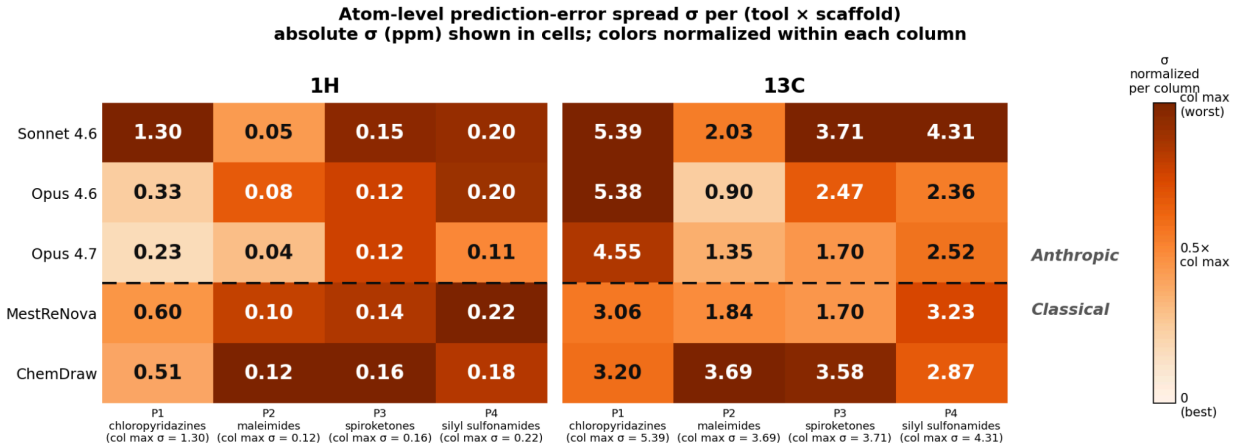

Figure 6. Atom-level prediction-error spread per (tool × scaffold). σ = standard deviation of signed Δ across atoms within the scaffold (LLM cells pool 3 replicates). σ is the headline metric for spread: a tool with low σ predicts the scaffold consistently (tight residuals around its mean bias), even if the mean bias itself is non-zero. Two panels (¹H, ¹³C); cell color is normalized within each column to its maximum σ (darker = more spread relative to the worst tool on that scaffold), with absolute σ printed in each cell.

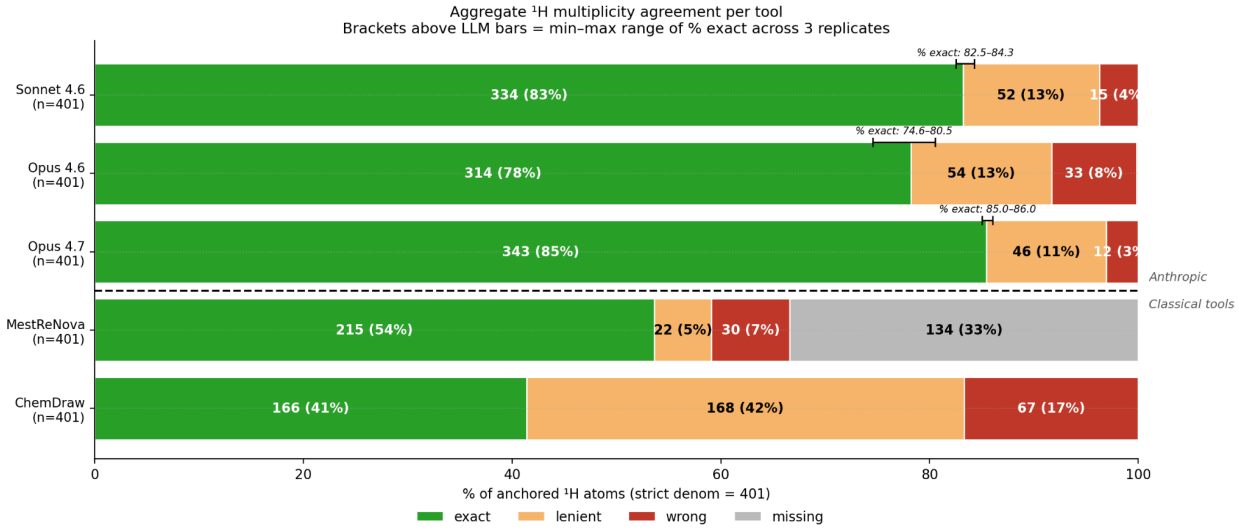

Figure 7. ¹H multiplicity agreement across 20 compounds. Stacked bars per tool show how each predicted multiplicity was scored against the experimental call, in four categories: exact (the predicted multiplicity matches experiment), lenient (experiment was reported as "m" but the tool committed to a specific call), wrong (concrete disagreement, or experiment was concrete and the tool predicted "m"), and missing (no predicted peak aligned to this experimental atom). Scoring uses a strict denominator of 401. Every anchored ¹H atom across the 20 compounds is counted, including atoms a tool failed to predict (which fall into missing), so coverage gaps cannot be hidden by simply omitting hard atoms. Brackets above the LLM bars span the min–max range of % exact across the three independent replicates per compound; classical tools (MestReNova, ChemDraw) are deterministic single predictions and have no bracket.

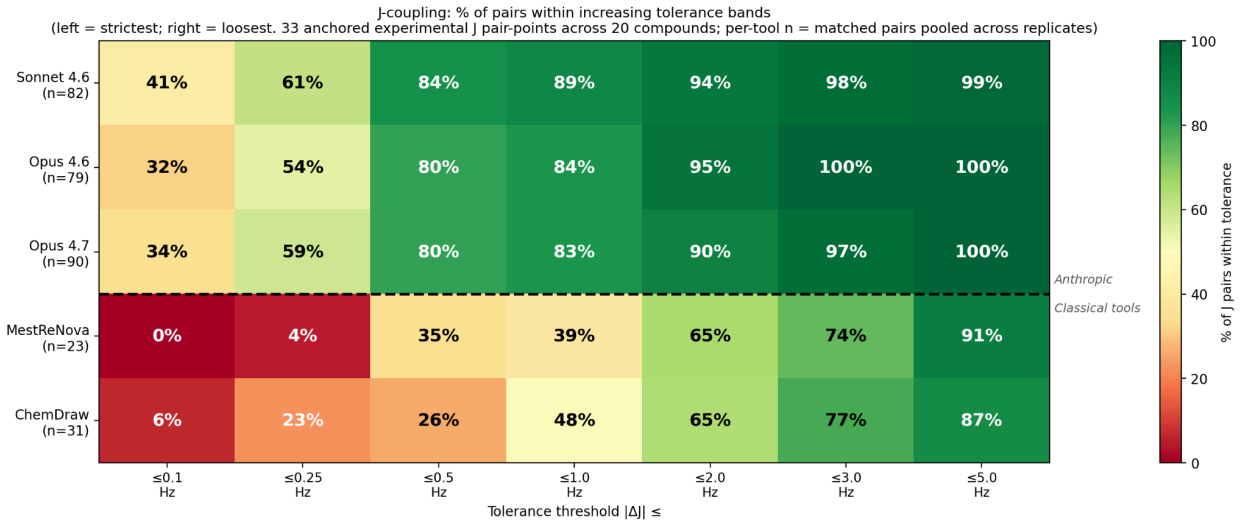

Figure 8. J-coupling accuracy across 20 compounds, summarized as a tolerance-band heatmap. Cells show the percentage of matched experimental–predicted J pairs (33 anchored experimental pair-points across the 20 compounds; per-tool matched counts are shown beside each tool label and reflect the 3-replicate pool for LLMs and the single deterministic run for classical tools) that fall within increasingly strict tolerance thresholds, from ≤ 5 Hz on the right to ≤ 0.1 Hz on the left. Color encodes the percentage on a diverging green-to-red scale (green = high pass rate, red = low). For LLMs, predictions pool the three replicate runs.

Inverse prediction: NMR → structure

Opus 4.7 was given 15 NMR/HRMS elucidation problems and asked, three times each, to return up to three ranked SMILES candidates, with stereochemistry excluded since 1D NMR alone cannot define absolute stereochemistry. The 15 problems were run under two prompt conditions, chosen by structural complexity. For 8 of the simpler targets (Q1–Q5, Q9, Q10, Q14), Opus 4.7 received only the HRMS and 1D NMR. For the remaining 7 structurally denser targets (Q6–Q8, Q11–Q13, Q15), the same data plus the starting-material SMILES were provided, since accurate prediction on these scaffolds required that

additional anchor. No other reaction context (reagents, conditions, mechanism, expected scaffold) was provided in either condition. The two conditions reflect the two situations a chemist actually faces: confirming the product of an unspecified reaction, versus confirming the product of a reaction whose inputs are known.

Recovery tracks structural complexity (Figure 9). The single-ring and two-fragment scaffolds (anilino-chloropyrimidines, Boc-N-aryl maleimide, oxazolidinone-iodobenzamide, triethylsilyl sulfonamide) were recovered on every attempt from HRMS and spectra alone. The structurally denser targets (fused chloropyridazines, an N-acyl oxazinane, spirocyclic carbonyl systems, an α-aryl triethylsilyl sulfonamide, the gem-disilyl sulfonamide) required the starting-material SMILES; under that condition, Opus 4.7 predicted the correct candidate structure on all three attempts for Q6, Q8, Q12, and Q15, and on two of three attempts for Q7, Q11, and Q13.

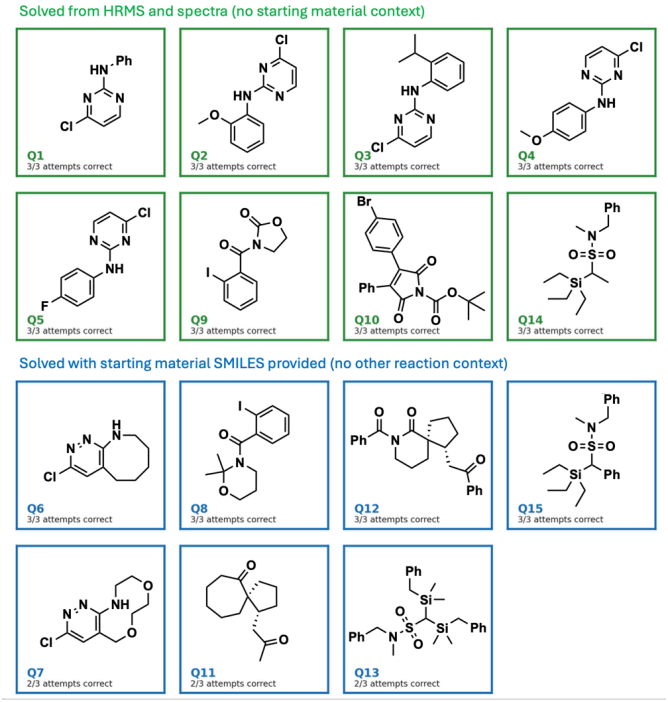

Figure 9. Structure-elucidation results across the 15 inverse-task problems. Each panel shows the published target with its success count out of 3 attempts. Border color indicates the prompt condition: green for spectra and HRMS only with no starting-material context; blue for spectra, HRMS, and the starting-material SMILES, with no other reaction context.

Takeaway

Two findings stand out. First, an LLM matches the strongest dedicated NMR tool on forward ^{13}C shift prediction (1.37 vs 1.48 ppm MAE) and records the lowest ^{1}H error of any

tool (0.079 ppm), across 20 compounds in four scaffold classes. Second, the same model makes inverse elucidation from 1D NMR + HRMS tractable, a problem that has historically been underserved by software, with dedicated elucidation packages either requiring 2D experiments or assuming a candidate structure to assign against, leaving the 1D-only case to expert reasoning. Together these change what a working chemist can ask of a general reasoning model. Confirming the product of a known reaction, ruling out a regioisomer, sanity-checking an assignment, and triaging which compounds need 2D experiments are all tasks that previously required either dedicated software or expert reasoning; a single model now handles each of them as plain text. The practical effect is fewer steps that demand expert attention on the routine cases, and more time available for the experiments and decisions that genuinely need it.

Limitations

Several caveats bound the conclusions. The assessment was small—20 compounds across four scaffolds for the forward task, 15 for the inverse task—and each scaffold contributes a single class of failure modes, so the numerical rankings should be read as indicative rather than precise. Slow-exchange NH heteroaromatics are sampled only through chloropyridazines, leaving related systems (hydroxypyridines, aminothiazoles, and other DMSO-d_6 NH-active scaffolds) untested; 2D experiments (COSY, HSQC, HMBC) and stereochemistry are out of scope by design, since 1D NMR alone cannot fix configuration. Solvent coverage is limited to DMSO-d_6, $CDCl_3$ and D_2O, so methanol-d_4, benzene-d_6, and acetone-d_6 are not assessed.

References

Kordubailo, M. V.; Borysov, O. V.; Vlasov, S. V.; Ryabukhin, S. V.; Volochnyuk, D. M. The Assembly of Fused Pyridazines via One-Pot Sequence IEDDA Reaction, Scope and Limitations. *ChemRxiv*, April 22, 2026. https://doi.org/10.26434/chemrxiv.15002316/v1.

Heymans, T.; Durant-Baudet, L.; Landrain, Y.; Evano, G. Straightforward and Modular Maleimide Synthesis from Ynamides. *ChemRxiv*, 2026. https://doi.org/10.26434/chemrxiv.15002423/v1.

Strong, C. S.; Chen, P.-J.; Bissenali, S.; Goddard, W. A., III; Stoltz, B. M. Enantioselective Michael Spirocyclization of Palladium Enolates. *ChemRxiv*, 2025. https://doi.org/10.26434/chemrxiv-2025-59lfh.

Leitch, M. A.; Ding, R.; Garrigues, S. L.; Chiong, H.; Mwiva, J.; Miller, K. E.; Liu, P.; Wang, Y.-M. Access to an Elusive Soft Deprotonation Manifold: Direct α-C(sp3) Silylation of Sulfonyl Compounds. *ChemRxiv*, April 21, 2026.
https://doi.org/10.26434/chemrxiv.15002274/v1.