MOLECULAR STRUCTURE ELUCIDATION WITH FRONTIER MODELS: A BENCHMARK ON LITERATURE-REPORTED IR, ^{1}H, AND ^{13}C NMR PEAK LISTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frontier models are often presented as near-solved structure elucidators on curated spectra. We ask an operational question: given the molecular formula and literature-reported IR, ^{1}H, and ^{13}C NMR peak lists, printed wavenumbers and shift tables, not digitised spectra images, can an out-of-the-box frontier model in an agentic workflow recover the correct constitution, and which stage fails when it does not? We release **IRSpectra-Bench** as a frozen roster of 500 problems, an RDKit InChIKey-14 contract, and the prediction deposits a mechanical scorer can replay. The object is a factorised diagnosis of one closed-book harness. One trained NMR-only baseline is scored on the same roster: NMRTrans, formula and printed ^{1}H/^{13}C lists, IR unused (79/500 top-1; Appendix A.13). It is not a full multi-modal peer of the IR+^{1}H+^{13}C task. On one closed-book frontier harness, generation top-1 is 45.4% (227/500) and recall (the fraction of problems whose true structure enters the proposed pool) is 49.8% (249/500). Self-ranking places the true structure first in 91.2% of the cases ($227/249$). The generation wall is 227 exact top-1 / 22 in-pool not top-1 / 251 never proposed. The roster is near-50/50 simple/complex; reweighted to the eligible corpus mix, top-1 on a validate-clean subset is 26.5%. Forward verification of the same pools is a secondary drop, not a second wall: top-1 falls from 227/500 to 204/500 (-23; diagnostic $204/45/251$; Appendix A.8). On a 60-compound arm only, formula-only recovery is 3/60, accuracy is flat in source-paper year, and recall stays below precision across four vendor families.

1 INTRODUCTION

Predicting a compound's molecular structure from spectral information, such as mid-infrared spectra and one-dimensional NMR chemical shifts, is routine in synthetic chemistry, as it validates reaction steps. Infrared band positions encode functional groups and conjugation, for example, carbonyl stretches, alcohols and amines envelopes, aromatic C=C, and the paired nitro bands that survive even a terse experimental section. NMR chemical shifts, multiplicities and integrals encode the number and neighbourhood of hydrogen (^{1}H NMR) and carbon environments (^{13}C NMR). The form that enters the literature is rarely the instrument file. Experimental journals commonly report **peak lists**, wavenumbers in cm^{-1} and shift tables in ppm, often with coupling constants, and only rarely with intensities or lineshape, and not digitized spectra. Furthermore, 2D correlations (COSY, HSQC, HMBC) that classical computer-assisted structure elucidation (CASE) uses to constrain connectivity (Elyashberg et al., 2012; Elyashberg & Williams, 2015) are frequently absent from those tables. A model that reads the open literature must therefore invert constitution from formula plus three printed lists, not from a digitised trace or a full CASE constraint set.

Learned systems now consume those same reports as sequence or tool-using tasks. Encoder–decoder models train on paired spectra–structure corpora (Anonymous, 2024; Ottomano et al., 2025; Alberts et al., 2024); puzzle benches and search systems report high recovery on curated/simulated or education spectra (Kamber, 2026; Guo et al., 2024; Zhuang et al., 2025; Espejo Morales et al., 2026). Reported recovery for the same model can invert with the harness (GPT-4o, 1.4% vs. 57.8% on MolPuzzle (Guo et al., 2024; Zhuang et al., 2025)), which already shows that a single end-to-end number is not a stable property of the chemistry. The operational setting that matches how chemists

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Closed-book generation on the frozen n=500 roster. Peak lists feed candidate generation, then selection from that pool. Lower panel: exact top-1, in-pool but not first, and never proposed.

read papers is stricter: an out-of-the-box model, a molecular formula $f_{\mathcal{M}}$ (as from high-resolution mass spectrometry), IR/^{1}H/^{13}C lists as printed, a fixed molecular-representation scoring contract, and a measurement of **which stage binds** when recovery fails. That stage is either proposal of the true constitution or ranking among candidates that were already proposed.

Any system that returns a ranked molecular candidate list factorizes as

$$\text{top-1} \;=\; \underbrace{P(\text{true} \in \text{pool})}_{\text{generation recall}} \times \underbrace{P(\text{rank} = 1 \mid \text{true} \in \text{pool})}_{\text{verification precision}}. \tag{1}$$

The same split is the classical CASE architecture: generate constitutional isomers, then match predicted to observed spectra (Elyashberg & Williams, 2015; Priessner et al., 2026). Recall is a fraction of n problems; precision | recall is a fraction of the recalled subset. The two rates have different denominators and must not be different (Eqs. 3). A better ranker cannot raise recall, because re-ranking cannot recover a structure that never entered the pool (Priessner et al., 2026). We report both factors on a frozen literature-peak-list roster at n=500, so that a high top-1 cannot hide a missing-candidate failure and a high verifier score cannot be read as a solved elucidator.

Contributions.

1. **A frozen roster and a mechanical scorer.** IRSpectra-Bench is a literature peak-list problems and a fixed InChIKey-14 contract (Eq. 2). Later systems deposit ranked SMILES. NMRTrans is the one trained NMR-only baseline scored below (IR unused; not a full multi-modal peer). The other systems in Section 2 are not scored in this work.
2. **The generation diagnosis** on one closed-book harness: recall $249/500$, self-rank $227/249$, wall $227/22/251$ (Figure 1), and corpus-reweighted top-1 26.5% beside the balanced-roster 45.4% (Table 1).
3. **Bounds on that headline.** Formula-only and recency; four vendor families on a 60-compound arm only.
4. **Systems not scored.** Section 2 names the published elucidators that were not run on this input. The factorisation above is what this harness measures without them.

Peak-list resource. Band and shift lists are taken from (Anonymous, 2026), an anonymized review copy is in Appendix A.12, a spectral peak-list resource of 121,233 records (57,646 structure-linked; 39,118 full IR+^{1}H+^{13}C+structure quadruples). The n=500 roster is drawn from the structure-complete `irexp_resolved` split; Listing 1 shows one JSON+molecule record.

2 Related Work

Paired corpora have produced strong in-distribution elucidators, and they set the default reporting language for the field. For example, (Anonymous, 2024; Ottomano et al., 2025) fuse infrared with

NMR in encoder–decoder or generative setups (Anonymous, 2024; Ottomano et al., 2025); Hu *et al.* and NMRTrans invert one-dimensional NMR with multitask or set-transformer decoders (Hu et al., 2024; Yang et al., 2026); Alberts and coworkers train a transformer-based architecture on NIST experimental absorbance and report end-to-end top-k SMILES (Alberts et al., 2024; 2025). Each of these models is trained on the representation native to its own corpus, such as a NIST infrared trace or a one-dimensional NMR list, and emits candidate structures as molecular graphs or SMILES. Those training inputs are not the heterogeneous printed IR+^{1}H+^{13}C peak lists that IRSpectra-Bench supplies. Training and scoring on the native format does not make the model a solver for the lists used here. Because they are typically scored by a single top-1 or top-k number, a failure of candidate generation is indistinguishable from a failure of ranking: both look like a miss. Only NMRTrans, among these models, is run on the roster below, and only as an NMR-only baseline: formula and printed ^{1}H/^{13}C, IR unused (Appendix A.13). That score is not an IR+^{1}H+^{13}C measurement. A top-k on NIST absorbance or on a simulated trace does not say whether the true constitution was proposed from a printed three-list experimental section.

Recent language-model systems occupy three operating regimes, and the regimes differ in what they can repair. Tool-using laboratory systems such as ChemCrow (M. Bran et al., 2024) and Coscientist (Boiko et al., 2023) call instruments, search tools or reaction planners; they are not instrumented to say whether an elucidation failure is a missing candidate or a mis-ranked one. Spectrum-native systems take a spectrum or peak list and re-rank a **proposed** pool: tree-search over MolPuzzle (Zhuang et al., 2025), SpecMol (Shen et al., 2025), Priessner *et al.* (Priessner et al., 2026), NMR-Solver (Jin et al., 2026) and NMRAgent (Fang et al., 2026a). Priessner already notes that a verifier cannot recover a structure that never entered the pool; we measure that failure on literature peak lists at n=500. SpectraLLM (Su et al., 2026) is a fine-tuned elucidator trained on simulated multi-spectral corpora, a different object from printed/reported experimental peak lists.

IR-Agent (Noh et al., 2026) is the closest concurrent infrared elucidator and the cleanest illustration of why the stage split matters. It is a tool-using multi-agent system on NIST experimental absorbance: a translator proposes a SMILES pool, a table-interpretation expert assigns peaks to functional groups, a retriever expert pulls structurally similar library spectra, and a structure-elucidation expert re-ranks the proposed pool. The architecture is extensible: new chemical information can be appended to prompts, and it reports top-k SMILES accuracy on a single-instrument experimental library. The structure-elucidation expert, like any verifier, can change which candidate is first; it cannot put a missing constitution into the pool. Table and retrieval features likewise condition ranking of structures that were already proposed. IRSpectra-Bench asks a different measurement on a different object: literature-reported IR+^{1}H+^{13}C peak lists plus molecular formula $f_{\mathcal{M}}$, one out-of-the-box closed-book model, and explicit generation recall versus verification precision | recall (Eqs. 3). There is no IR-Agent or SpectraLLM row on this roster. This distinction makes the benchmark a diagnostic of where an elucidation pipeline loses the structure, rather than a leaderboard comparison across incompatible inputs. The title therefore emphasizes literature-reported peak lists: the input format determines whether proposal and verification can be separated (see Section 2).

A third regime compares architectures on a shared suite. MolQuest (Han et al., 2026), NMR-Gym (Fang et al., 2026b) and Espejo Morales *et al.* (Espejo Morales et al., 2026) (education $\rightarrow$ industrial: 80.9% $\rightarrow$ 20.6%) ask which search or tooling stack wins. Those suites still need a stage contract if a win is to be attributed to a better proposer rather than a better ranker.

Heterogeneous peak lists and CASE Most learned suites use simulated spectra or single-instrument libraries such as NIST infrared or nmrshiftdb2-style collections (National Institute of Standards and Technology; Kuhn & Schlörer, 2015; Alberts et al., 2025). Literature peak lists are author-transcribed, often incomplete (missing weak bands or quaternary ^{13}C), referenced to local solvents and spectrometers, and stripped of the lineshape and two-dimensional correlation evidence that CASE engines exploit (Elyashberg et al., 2012; Buevich & Elyashberg, 2016). Matching predicted to observed shifts — DP4 probability (Smith & Goodman, 2010), HOSE-code lookup (Bremser, 1978), CASE enumeration (Elyashberg et al., 2012) — is a verification problem on a generated isomer list. CASE enumerators have exhaustive recall by construction, subject to the formula and spectral constraints; published LLM elucidation numbers rarely state whether the true structure was proposed at all. SpecX (Xiang et al., 2026) bounds a related multi-modal spectroscopy task on simulated traces.

NMRTrans (Yang et al., 2026) is run here on formula and printed ^{1}H/^{13}C lists, with IR unused, because the released model accepts that printed peak-list interface. The roster scores are 79/500 top-

1 and 102/500 recall, with 9/500 gold keys in its training split (see Appendix A.13). Those scores are the trained baseline obtained on this interface. It is not a full IR+^{1}H+^{13}C peer. IR-Agent, CASE, Spectro, NMIRacle, the Alberts infrared transformers and SpectraLLM were not run, because API access, licence, and input do not match this printed three-list harness.

3 METHODOLOGY

IRSpectra-Bench is a frozen roster, the InChIKey-14 contract, and the three rates in Eq. 1 (Appendix A.4). Section 2 is the list of systems this section does not score.

3.1 PROBLEM SETUP

The molecular elucidation problem provides a molecular formula $f_{\mathcal{M}}$ (as from high-resolution mass spectrometry) together with three author-transcribed lists: IR band positions $I \subset \mathbb{R}_+$ in cm^{-1}, ^{1}H multiplets H (shifts, integrals, and couplings as printed), and ^{13}C shifts $C \subset \mathbb{R}$ in ppm; see Listing 1 for an example. The target is the covalent constitution $y^\star$ of the unknown, represented as a SMILES string and scored at the first fourteen characters of the RDKit InChIKey (Eq. 2). Stereo, tautomer and salt-form distinctions that live in the remaining InChIKey characters are held out of the headline and reported on a validate-clean subset (Appendix A.5). No name, SMILES, scaffold, or retrieved spectrum is part of the input; the solver never sees the source paper.

A solver returns a ranked pool $\mathcal{C} = (c_1, \ldots, c_k)$ with $k \leq 3$. Generation recall, top-1 and self-ranking precision | recall are the evaluation protocol of this paper, defined for the frozen n=500 IRSpectra-Bench roster under that candidate budget. Generation recall is the fraction of the roster for which $y^\star$ appears anywhere in $\mathcal{C}$. Top-1 is the fraction for which c_1 matches $y^\star$. Self-ranking precision | recall is the fraction of recalled problems on which the solver's own order places $y^\star$ first. An independent forward-verification ranker (Eq. 4) may reorder the same pool; it cannot add a missing constitution. The n=500 forward-verify diagnostic then partitions the same pools into verified, misranked and never-proposed cells (Appendix A.8); that split is not the generation wall. Because top-1 factorises as recall times precision | recall, a better ranker cannot raise recall, and the two rates must not be differenced.

Cohort and inputs Each of the 500 problems is exactly the textual form reported in open-access papers: formula plus IR band list plus ^{1}H/^{13}C shift lists. The inputs are author-transcribed IR band and chemical shift lists, not absorbance traces or FIDs, and therefore a different object from digitised-spectrum benchmarks (Espejo Morales et al., 2026). That is also the form a language model reads when it is asked to work from the literature rather than from an instrument file.

The evaluation set is a frozen roster of 500 question IDs released with the code. Problems are drawn from the structure-complete split of (Anonymous, 2026) (`irexp_resolved`). The question IDs are a pre-specified list, not an outcome-selected subset (Appendix A.3.3). Histograms of molecular weight, ring count, C–F bonds, and nitrogen atoms for the validate-clean subset (n=295: locked cohort plus clean expansion) of (Anonymous, 2026) data are in Appendix A.5. On that plate, the median molecular weight is 306 $\mathrm{g{\cdot}mol^{-1}}$ and the median ring count is two rings. The thinking-tier expand-500 cut is not in the plate. Listing 1 is the solver-facing contract: formula plus printed peak lists; gold constitution is held out and scored offline.

Difficulty and scoring Difficulty is declared *a priori* by RDKit ring analysis, before any solver is run: a problem is *simple* if and only if it has at most two rings, no fused, spiro, or bridgehead system, and at most 22 heavy atoms; otherwise it is *complex* (248/252 on n=500). The released roster is **near-50/50 by design** (248/252), against an eligible corpus that is 17.5% simple / 82.5% complex. Table 1 therefore puts the balanced-roster top-1 next to the corpus-reweighted correction on a validate-clean subset. The reweighted rows are not an n=500 interval, and they are the number to quote for an arbitrary literature draw.

A candidate $\hat{y}$ is correct at constitution *iif* the first 14 characters of the RDKit InChIKey match

$$s(\hat{y}, y^\star) \;=\; \mathbf{1}\big[\mathrm{IK}_{14}(\hat{y}) = \mathrm{IK}_{14}(y^\star)\big]. \tag{2}$$

Those 14 characters are the connectivity layer; stereo-sensitive InChIKey is reported on the validate-clean subset (Appendix A.5). External submissions deposit up to three ranked SMILES per `qid`

(docs/LEADERBOARD.md; scripts/score_submission.py). With a ranked pool $\mathcal{C}_i = (c_{i1}, \ldots, c_{ik})$, $k \leq 3$,

$$\text{top-1} = \frac{1}{n}\sum_{i=1}^{n} s(c_{i1}, y_i^\star),\ \text{prec.} \mid \text{recall} = \frac{\sum_i s(c_{i1}, y_i^\star)}{\sum_i \mathbf{1}[y_i^\star \in \mathcal{C}_i]},\ \text{recall@}k = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[\exists\, j \leq k : s(c_{ij}, y_i^\star) = 1]\,. \tag{3}$$

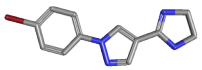
Top-1 and recall@k share the denominator n; precision | recall is defined only on the recalled subset. Eq. 1 is exactly top-1 $=$ (recall@k) $\times$ (prec.|recall) when both stages use the same ranked list. Confidence intervals, where reported, are bootstrap 95%; paired model differences use McNemar with Holm correction. The n=500 headline does not carry confidence intervals.

4 Experimental setup

The experiments run the problem of Section 3.1 under a fixed closed-book harness: one out-of-the-box model family on the frozen roster, a declared candidate budget, and an optional second ranker that cannot enlarge the pool. What follows specifies the payload, the solver, the scoring contract, and the probes that later sections use to interpret the generation wall. For further details regarding close-book harness see Section A.15 in the SI.

Blind payload and solver. The solver sees four JSON fields per qid: formula (as from HRMS), ir_bands_cm-1, h_nmr, c_nmr — peak lists as printed (Listing 1). No name, SMILES, scaffold, atom-type hint, or retrieved spectrum enters the prompt; the source paper is not fetched; ground truth is scored offline. The headline solver is Claude Opus 5 via a consumer subscription, closed-book, with one sub-agent per batch and an RDKit formula check that rejects candidates whose elemental composition disagrees with f. The pooled n=500 roster mixes earlier no-thinking runs with the thinking-tier expand-500 cut (200 of 500) that completes the cohort. Both parts use the same InChIKey-14 contract, with no fine-tuning and no paid API. **Frozen:** question ids, gold InChIKey-14 keys, prediction JSONL, and the scorers. Those regenerate 227/500 and 249/500. **Not frozen:** model snapshot, temperature, seed, and the batch wrapper; only the four-field payload was kept. Early sampler seeds were not stored, so the roster is reusable and not re-drawable (Section 6).


```
"id": "ILJNJJNKEOAREX-UHFFFAOYSA-N",
"smiles": "Brc1ccc(-n2cc(C3=NCCN3)cn2)cc1",
"inchikey": "ILJNJJNKEOAREX-UHFFFAOYSA-N",
"formula": "C12H11BrN4",
"ir_bands_cm-1": [3144.0, 3110.0, 3066.0, 2931.0,
                  2863.0, 1632.0, 1589.0, 1565.0, 1485.0],
"ir_source": "experimental",
"source_doi": "PMC:12566713", "pmcid": "PMC12566713",
"h_nmr": "8.83 (s, 1H), 8.04 (s, 1H), 7.81 (d, J = 8.8 Hz, 2H), 7.72 (d,
    J = 8.8 Hz, 2H), 3.56 (s, 4H)",
"c_nmr": "157.5, 140.3, 138.4, 132.4, 127.5, 120.4, 119.0, 116.4, 49.1",
```

Listing 1: One IRexp JSON record (commercial DoR; ILJNJJNKEOAREX) (Anonymous, 2026).

Scaffold splitting. A Bemis–Murcko scaffold split is the appropriate hold-out when a model is trained on molecules, and generalization is measured on unseen scaffolds. That design does not apply to the headline solver. Claude Opus 5 is an out-of-the-box closed-book model; every compound on the frozen roster is an evaluation item, and the solver is not trained on a split of IRSpectra-Bench. Memorisation of a chemical neighbourhood, or of a particular paper, is examined with the contamination probes: formula only against the full lists on the nested control, and accuracy against source-paper year (Section 5.2; Figure 2; Appendix A.9.5, A.9.6). Murcko-scaffold comparisons on the validate-clean subset describe how close a prediction is to the reference; they are not a data-split protocol.

Scoring contract and candidate budget. A deposit is JSONL of {qid, candidates: [SMILES,...]} under a stated budget. The scorer reports the three Section 3 metrics at InChIKey-connectivity, with bootstrap 95% CIs on arms that carry them. Self-ranking and forward-verified re-ranking must be labelled separately: they share a pool but not a ranker, so their precision | recall figures are not interchangeable. Headline protocol: at most three ranked SMILES per qid.

On the 60-compound arm, Claude Opus 5 returned a mean of 2.20 candidates versus 3.00 for the other vendors, so recall rankings across vendors are approximate. **Generate-wide** (that arm only) runs ten independent solvers with up to six SMILES each, then pools and re-ranks; that budget is not the n=500 reporting contract.

Forward-verification and probes. The second ranker follows generate → forward-predict ^{13}C (blind to the observed list) → re-rank by symmetric chamfer on ^{13}C peak sets $P, Q \subset \mathbb{R}$

$$d(P,Q) \;=\; \frac{1}{|P|}\sum_{p\in P}\min_{q\in Q}|p-q| \;+\; \frac{1}{|Q|}\sum_{q\in Q}\min_{p\in P}|q-p| \tag{4}$$

(Section 5.4). Chamfer is assignment-free: it remains defined when $|P| \neq |Q|$, which is the usual case for incomplete printed ^{13}C lists. This ranker uses the same candidate pool and therefore cannot raise generation recall. Controls, reported in order below, are a nested formula-only ablation and recency (Section 5.2); four-vendor replication on a 60-compound arm (Section 5.3); generate-wide sampling; HOSE lookup and GNN ^{13}C verifiers; and a small IRexp-fine-tuned generator probe that is excluded from headline claims.

5 RESULTS

The measurements below are one closed-book harness on the frozen n=500 roster, then the 60-compound controls (Sections 3–4). The generation wall is the lead result. Forward verification is the secondary drop on the same pools. Systems named in Section 2 do not appear in these tables. NMRTrans is scored in Appendix A.13 as an NMR-only baseline (formula and printed ^{1}H/^{13}C; IR unused; 79/500 top-1). These tables do not include an IR+^{1}H+^{13}C row.

5.1 HEADLINE PERFORMANCE

Table 1 summarizes the generation cohort of the closed-book harness. Generation recall is 249/500, and self-ranking selects the correct structure in 227/249 recalled cases, yielding an overall top-1 accuracy of 227/500 (45.4%).

Table 1: Closed-book generation (n=500 point estimates). * reweights to the corpus mix (17.5% simple / 82.5% complex), not an n=500 interval.

metric	value
top-1 exact constitution	**45.4%** (227/500)
simple (n=248)	65% (161/248)
complex (n=252)	26% (66/252)
generation recall (top-3)	**49.8%** (249/500)
self-rank prec.|recall	91.2% (227/249)
reweighted top-1*	**26.5%** [21–32]
reweighted recall*	31.3% [25–37]

The resulting generation wall is therefore 227 exact top-1 predictions, 22 candidates that were generated but incorrectly ranked, and 251 structures that were never proposed (Figure 1). Thus, most failures arise during candidate generation rather than self-ranking. The evaluation roster is approximately balanced between simple and complex molecules (248/252), whereas only 17.5% of the eligible corpus is simple. Reweighting to the corpus composition therefore reduces top-1 accuracy to 26.5% [21–32] and recall to 31.3% [25–37]. Forward verification on the same candidate pools achieves 204/500 top-1, 23 fewer correct predictions than self-ranking (Section 5.4); this ranking effect is distinct from the generation wall. Performance also depends strongly on molecular complexity: top-1 accuracy is 65% (161/248) for simple molecules but 26% (66/252) for complex molecules. Once the molecular formula $f_{\mathcal{M}}$ is fixed, fused and polycyclic systems admit substantially more constitutions compatible with the reported spectral peak lists. On the validate-clean subset, scaffold-level recovery is 64% (80% simple, 48% complex), with a mean best Tanimoto similarity of 0.66. Accuracy also decreases with molecular size: 67.9% top-1 for molecules with $\leq$ 15 heavy atoms, 42.8% for 16–25, and 16.7% above 25 (n=53/152/90; Appendix A.5). Among 137 analysable top-1 misses in an instrumented slice, **76.6%** are constitutional isomers of the true structure, whereas only **22.6%** share its Murcko scaffold. The typical failure therefore preserves composition but not the underlying connectivity, rather than representing a small regiochemical perturbation of the correct structure. A battery-electrolyte subset (n=46) is similarly recall-bound, with 26% top-1 accuracy; results from the underpowered four-model Claude ladder are reported in Appendix A.9.10.

5.2 Ablation of the printed lists

Figure 2a is the nested formula-only control on the non-thinking n=60 roster (Appendix A.9.5). Formula-only recovers 3/60 (5%), and all three stay recovered once IR and NMR lists are supplied (full lists 14/60; McNemar p=0.001). On that nested arm, formula+IR stays at the formula-only floor (5%; Appendix A.9.5).

Table 2 is the separate Claude thinking leave-one-modality arm on the same roster (full lists 35/60).

Accuracy is flat in source-paper year on the n=194 slice (r=−0.007; Appendix A.9.6), which bounds verbatim recall but does not exclude contamination (Section 6). Removing ^{1}H drops top-1 from 35/60 to 21/60 (p=0.001) and removing the formula drops it to 25/60 (p=0.006); removing IR or ^{13}C does not (p=1.000), and spectra-only remains well above the nested formula-only floor. The lists are being used; the remaining question is whether the same recall ≪ precision split is a property of one vendor.

Table 2: Leave-one-modality (Claude thinking, n=60). p: McNemar vs full lists.

condition	top-1	top-3	p
$f_{\mathcal{M}}$ + IR+^{1}H + ^{13}C	35/60 (58%)	38/60 (63%)	—
$f_{\mathcal{M}}$ + ^{1}H + ^{13}C	34/60 (57%)	41/60 (68%)	1.000
$f_{\mathcal{M}}$ + IR +^{13}C	21/60 (35%)	27/60 (45%)	0.001
$f_{\mathcal{M}}$ + IR + ^{1}H	34/60 (57%)	37/60 (62%)	1.000
IR+^{1}H + ^{13}C	25/60 (42%)	26/60 (43%)	0.006

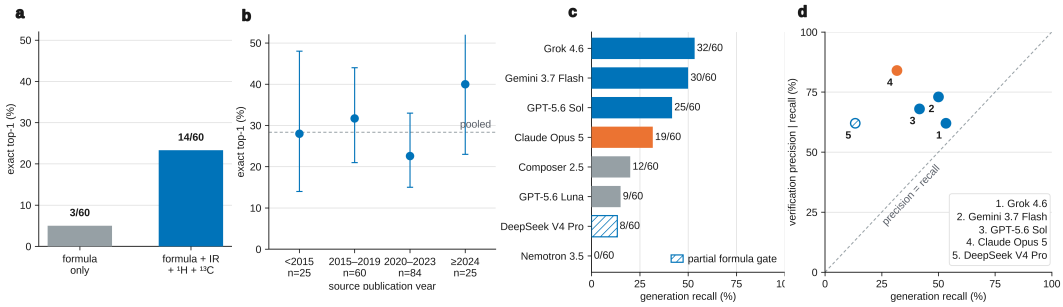


Figure 2: (a) Nested formula-only control on the non-thinking n=60 roster (not Table 2): formula-only 3/60 versus full lists 14/60. (b) Top-1 by source-paper year (n=194). (c) Vendor generation recall (n=60). (d) Chamfer precision against recall (n=60).

5.3 Cross-vendor replication

The 60-compound arm repeats the closed-book payload on three further vendor families. It does not repeat the n=500 headline or the thinking leave-one arm (Table 2): Claude Opus 5's self-rank top-1 on this nested arm is 14/60 (23%), not 227/500 (45.4%). Grok 4.6, Gemini 3.7 Flash and GPT-5.6 Sol saw the same four fields and the same candidate contract. Forward-verify precision | recall exceeds generation recall in every row of Table 3, on different denominators (Eq. 3). The precision column is chamfer re-ranking of each solver's pool, the same ranker as Section 5.4, not the solver's own order. Claude Opus 5's self-rank top-1 on this arm remains 14/60; the 16/19 cell does not multiply with 19/60 to that top-1. Bootstrap intervals on the paired forward-verify-minus-recall contrast separate for Claude Opus 5 (+52.5 points), GPT-5.6 Sol (+26.3) and Gemini (+23.3); Grok's contrast (+9.2) is directional. Candidate budgets differ (Claude Opus 5 mean 2.20 versus 3.00 for the others), so recall rankings among vendors are approximate. A clean-clone control for Grok bounds key leakage (McNemar p=0.39).

5.4 Forward verification

Forward verification asks whether an independent ^{13}C ranker can repair the misses that remain after the solver has proposed a pool. When the true structure is among the candidates on the 60-compound arm, the chamfer ranker selects it in 16/19 cases (84%). On this nested arm the margin over self-ranking (14/60 → 16/60 top-1) is small and unresolved; we do not claim an accuracy advance on that arm. On n=500, forward-verify top-1 is 204/500, a drop of 23 from self-rank 227/500. The

Table 3: Cross-vendor n=60. Recall is pool membership; prec.|rec. and multi-cand. are chamfer re-rank, not self-rank.

model	recall	prec.|rec.	multi-cand.
Claude Opus 5	19/60 = 32%	16/19 = 84%	10/13 = 77%
Grok 4.6	32/60 = 53%	20/32 = 62%	20/32 = 62%
Gemini 3.7 Flash	30/60 = 50%	22/30 = 73%	22/30 = 73%
GPT-5.6 Sol	25/60 = 42%	17/25 = 68%	16/24 = 67%

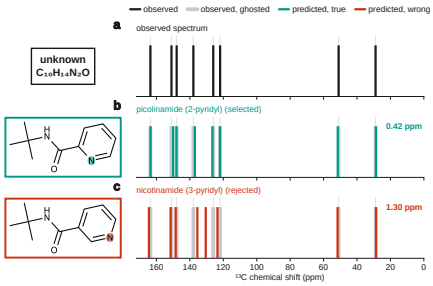


Figure 3: v3-R25: self-rank picks 3-pyridyl; forward ^{13}C picks 2-pyridyl (0.42 vs 1.30 ppm).

forward-verify diagnostic — not the generation wall — is 204 verified / 45 misranked / 251 never proposed (Appendix A.8). No re-ranking of an already-proposed pool repairs the 251 compounds where the true structure never entered the list.

Worked cases. Three deposits instantiate these modes (Appendix A.9; true and top-1 structures in Figure S3). **main-R06** ($C_{29}H_{30}FNO_2$): true acridinedione never enters the 3-candidate pool (Tanimoto-0.90 regioisomer); verifier unused. **v3-R25** ($C_{10}H_{14}N_2O$): verify-save — self-rank prefers 3-pyridyl, forward ^{13}C recovers 2-pyridyl (chamfer 0.42 vs 1.30 ppm; Figure 3). **v3-R26** ($C_{14}H_{11}N_3O_3$): verify-fail — true 3-nitrophenyl isomer in the pool but 2-nitro ranked first (chamfer 1.35 vs 1.36 ppm).

Generate-wide. If proposal is the expensive stage, enlarging the pool should lift recall more than swapping the ranker. Ten solver agents, up to six candidates each, pooled and re-ranked on the 60-compound arm, move recall 32% → **42%** and top-1 23% → **30%** (McNemar p=0.34); verification precision falls 84% → 72% as noisier candidates enter. The training-free ceiling remains recall-limited (the 23% → 27% → 30% ladder is Figure S2).

Non-LLM verifiers. The same candidate sets can be re-ranked without a language model. A HOSE-style lookup ties self-ranking; a modest GNN reaches 59/65 (91%) versus the LLM verifier's 58/65 (89%), a directional gap. Deranging observed ^{13}C pairings collapses precision (p=0.001), so the ranker is using the printed shifts rather than a composition prior. A small IRexp-fine-tuned generator pooled with Claude lifts recall on an instrumented slice 33.5% → 54.1% and top-1 28.4% → 35.1% (McNemar p=0.015). The recall limit is therefore a property of **training-free elicitation**: once a proposer is allowed to train on the same peak-list resource, recall moves.

5.5 Literature comparison

Published top-1 $= a$ and top-k $= b$ imply recall $\geq b$ and conditional precision $\leq a/b$ (`scripts/literature_decomposition.py`). The rows in Appendix A.11.3 (this work, NMR-Solver, NMRAgent, Espejo, Alberts, SpecX) are separate published settings. They are not scores on this roster, and they are not one collapse onto literature peak lists. The n=500 precision entry in that table is self-ranking (227/249); the forward-verify entry is an instrumented protocol slice.

6 Discussion

The reusable objects are the frozen n=500 deposits and the InChIKey-14 scorer. On those deposits one closed-book harness self-ranks a proposed pool in 227/249 cases and fails to propose the true constitution in 251 of 500 (generation wall 227/22/251; Figure 1). Forward verification is the secondary drop, 227/500 → 204/500 (Appendix A.8). Table 1 places corpus-reweighted 26.5% beside 45.4%. They are the score of one closed-book harness. They are not a score for IR-Agent or CASE, which remain unscored. NMRTrans is a separate trained NMR-only baseline on the same roster (79/500 top-1, IR unused; Appendix A.13), not a full multi-modal peer.

Reporting contract. External work should deposit ranked SMILES and report (i) top-1 constitution, (ii) generation recall , (iii) verification precision | recall (Eqs. 3), with a stated candidate budget. Nonsignificant ladder gains (protocol-slice forward-verification vs self-ranking, p=0.55; generate-wide top-1, p=0.34) say where effort should go. They are not accuracy advances. IRexp fine-tuning already lifts recall on an instrumented slice $33.5\% \rightarrow 54.1\%$, which is the direction a proposer-side intervention moves.

Implications for tool-using elucidation. A second ^{13}C ranker does not recover the 251 never proposed. The expensive stage is anything that enlarges $P(\text{true} \in \text{pool})$: broader sampling, formula-constrained decode, isomer enumeration, or library search that can introduce a missing constitution. IR-Agent-style table-interpretation and retrieval experts (Noh et al., 2026) are feature agents for ranking a pool that already exists; they inherit the same limit as the chamfer ranker, and they were not run here. MolQuest- and Espejo-style architecture search (Han et al., 2026; Espejo Morales et al., 2026) still needs the stage contract if a win is to be attributed to proposal rather than re-ranking. On literature peak lists, the binding improvement is a better proposer.

Limitations. What is frozen, and what is not, is stated in Section 4: deposits and the InChIKey-14 scorer regenerate the totals; the consumer subscription has no pinned snapshot, and the n=500 pool mixes no-thinking rows with a thinking-tier expand-500 cut (200 of 500). That mix is not a controlled comparison. Solver transcripts were audited for zero ground-truth access. Contamination is bounded by the nested formula-only control (Figure 2a; $23\% \rightarrow 5\%$ on n=60, full lists $14/60$ to formula-only $3/60$) and flat recency (r=−0.007), and is not excluded. The object is printed band lists with formula supplied as from HRMS. The unscored systems in Section 2 were not run. Forward-verify versus self-rank drops by 23 on n=500 and is unresolved on the protocol slice (p=0.55); generate-wide top-1 is unresolved (p=0.34); the n=24 ladder is underpowered. No expert-chemist audit was run. Leave-one-modality on the paired thinking arm (Table 2) was completed: Both ^{1}H and −formula are significant leave-one-out effects: ^{1}H drops from 35/60 to 21/60 (p=0.001), and removing formula drops to 25/60 (p=0.006); IR and ^{13}C do not (both 56.7%, p=1.000). Extraction fidelity is inherited from IRexp rather than re-audited here. The battery subset is literature electrolyte chemistry, not operando cells; scoring is single-sample; the roster inherits PMC-OA organic bias.

7 Conclusion

IRSpectra-Bench is a frozen roster of literature-reported IR, ^{1}H, and ^{13}C peak lists, read by one closed-book harness and scored with an InChIKey-14 contract. Each problem is the textual form of an experimental section: a molecular formula together with author-transcribed band positions and shift tables. The solver does not see the source paper, and the gold constitution is scored offline. The fourteen-character key fixes the covalent constitution that counts as correct, so the same scorer can replay a later deposit. The roster, the contract, and the deposits are a diagnosis of this harness.

On that harness, the failure that binds is generation recall. Top-1 factorises into two stages with different denominators: whether the true constitution enters the proposed pool, and whether, given that it did, the ranker places it first. A ranker cannot insert a structure that never entered the pool. Forward verification of a finished pool is a ranking diagnostic in the same sense. It can change which proposed structure is placed first. It does not restore a constitution that was never proposed. Generation on the balanced roster is 45.4% top-1, and most of what that figure misses was never proposed. The change that would move the headline is a better proposer.

The balanced roster and the corpus reweighting are separate scores. The roster is balanced simple against complex by design, and on that roster the harness scores 45.4% top-1. Reweighted to the eligible corpus, top-1 on a validate-clean subset is 26.5%, the figure for an arbitrary draw from the literature. The first is the score of the frozen roster. The second is the score of the same harness under the corpus mix. A reported top-1 has to say which of the two was computed.

The deposits and the InChIKey-14 scorer are the diagnostic for follow-on work. Under that contract, a true constitution that never enters the pool is a proposer failure, and one that enters but is not placed first is a ranker failure. A later deposit can be read as one or the other without a new scorer. We release both. The claim stops at this harness.

AI USE STATEMENT

Frontier LLMs were used as **experimental solvers** under the consumer harness in Sections 4–5 (Claude Opus 5 on the headline cohort; three other vendor families on the 60-compound arm), not as a substitute for author analysis, scoring, or claim-checking. Inference used no pinned model snapshot. The $n{=}500$ pool includes the thinking-tier expand-500 cut (Section 4). We did not use generative AI to invent or alter reported metrics. Generative tools assisted manuscript editing and formatting; the authors reviewed the text and take responsibility for all claims, including any AI-assisted wording.

ETHICS STATEMENT

This work evaluates publicly available open-access literature peak lists and commercial LLM APIs under consumer terms. No human-subject data. We do not claim clinical or regulatory readiness for automated elucidation. Licensing of redistributed numeric extractions is documented with the IRexp resource and dataset notices (PMC-OA terms are mixed — not uniformly CC-BY; Chemotion is CC-BY-SA).

REPRODUCIBILITY STATEMENT

Frozen prediction deposits and the scorer regenerate all training-free metrics. Frozen deposits and scorers are in the anonymous review mirror at `https://anonymous.4open.science/r/peaklist-corpus-review-10C4/`.

The InChIKey scorer is scripted. Inference is not bit-reproducible without the recorded temperature, seed, or wrapper. (Section 6). Replay the deposits. A bit-exact regeneration of the subscription run is not available. Fine-tuning against the roster should use `data/train_no_bench.jsonl.gz` so roster InChIKeys are withheld.

REFERENCES

Marvin Alberts, Teodoro Laino, and Alain C. Vaucher. Leveraging infrared spectroscopy for automated structure elucidation. *Commun. Chem.*, 7:268, 2024. doi: 10.1038/s42004-024-01341-w.

Marvin Alberts, Federico Zipoli, and Teodoro Laino. Setting new benchmarks in AI-driven infrared structure elucidation. *Digital Discovery*, 4:1936–1943, 2025. doi: 10.1039/d5dd00131e.

Anonymous. Spectro: a multi-modal approach for molecule elucidation using IR and NMR data. *ChemRxiv*, 2024. doi: 10.26434/chemrxiv-2024-37v2j. Author names omitted for double-blind review.

Anonymous. IRexp: experimental infrared band lists from open literature, 2026. Spectral peak-list resource; anonymised for double-blind review.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Wolfgang Bremser. HOSE — a novel substructure code. *Anal. Chim. Acta*, 103:355–365, 1978. doi: 10.1016/S0003-2670(01)83100-7.

Alexei V. Buevich and Mikhail E. Elyashberg. Synergistic combination of CASE algorithms and DFT chemical shift predictions: A powerful approach for structure elucidation, verification, and revision. *Journal of Natural Products*, 79(12):3105–3116, 2016. doi: 10.1021/acs.jnatprod.6b00799.

Mikhail Elyashberg, Kirill Blinov, Sergey Molodtsov, and Antony Williams. Elucidating 'undecipherable' chemical structures using computer-assisted structure elucidation approaches. *Magn. Reson. Chem.*, 50:22–27, 2012. doi: 10.1002/mrc.2849.

Mikhail E. Elyashberg and Antony J. Williams. *Computer-based structure elucidation from spectral data*. Lecture Notes in Chemistry. Springer Berlin Heidelberg, 2015. doi: 10.1007/978-3-662-46402-1.

Irina Espejo Morales, Damon Hinz, Marvin Alberts, Geraud Krawezik, Haewon Jeong, and Shirley Ho. NMR elucidation as an agentic search problem, not a modeling problem. *arXiv*, 2026. doi: 10.48550/arXiv.2607.19406.

Zheng Fang, Chen Yang, Yusen Tan, Yunpeng Zhao, Fanjie Xu, Hongxin Xiang, Hanyu Sun, Hanyu Gao, Xiaojian Wang, Wenjie Du, Yuqiang Li, and Jun Xia. Towards generalizable and evidential nuclear magnetic resonance-based molecular structure elucidation via large language model agent. *arXiv*, 2026a. doi: 10.48550/arXiv.2606.29776.

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*, 2026b. doi: 10.48550/arXiv.2601.15763.

Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Can LLMs solve molecule puzzles? A multimodal benchmark for molecular structure elucidation. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, pp. 134721–134746, 2024. doi: 10.52202/079017-4281. the benchmark is named MolPuzzle; OpenReview t1mAXb4Cop.

Taolin Han, Shuang Wu, Jinghang Wang, Yuhao Zhou, Renquan Lv, Bing Zhao, and Wei Hu. MolQuest: a benchmark for agentic evaluation of abductive reasoning in chemical structure elucidation. *arXiv*, 2026. doi: 10.48550/arXiv.2603.25253.

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 Cent. Sci.*, 10(11):2162–2170, 2024. doi: 10.1021/acscentsci.4c01132.

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. *Nat. Commun.*, 17:4740, 2026. doi: 10.1038/s41467-026-71315-0.

David Kamber. Making Claude a chemist: how Claude performs on NMR prediction and structure elucidation. Non-peer-reviewed company white paper, Anthropic, 6 2026. URL `https://anthropic.com/research/making-claude-a-chemist`.

Stefan Kuhn and Nils E. Schlörer. Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2 — a free in-house NMR database with integrated LIMS for academic service laboratories. *Magn. Reson. Chem.*, 53:582–589, 2015. doi: 10.1002/mrc.4263.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.*, 6:525–535, 2024. doi: 10.1038/s42256-024-00832-8.

National Institute of Standards and Technology. NIST Chemistry WebBook, SRD 69. URL `https://webbook.nist.gov`.

Heewoong Noh, Namkyeong Lee, Gyoung S. Na, Kibum Kim, and Chanyoung Park. IR-Agent: expert-inspired LLM agents for structure elucidation from infrared spectra. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2508.16112.

Federico Ottomano, Yingzhen Li, and Alex M. Ganose. NMIRacle: multi-modal generative molecular elucidation from IR and NMR spectra. *arXiv*, 2025. doi: 10.48550/arXiv.2512.19733.

Martin Priessner, Richard J. Lewis, Magnus J. Johansson, Jonathan M. Goodman, Jon Paul Janet, and Anna Tomberg. Enhancing molecular structure elucidation with reasoning-capable LLMs. *Digital Discovery*, 5:1237–1251, 2026. doi: 10.1039/D5DD00359H.

Shuaike Shen, Jiaqing Xie, Zhuo Yang, Antong Zhang, Shuzhou Sun, Ben Gao, Tianfan Fu, Biqing Qi, and Yuqiang Li. SpecMol: a spectroscopy-grounded foundation model for multi-task molecular learning. *arXiv*, 2025. doi: 10.48550/arXiv.2509.21861. circulated in v1 as MolSpectLLM.

Steven G. Smith and Jonathan M. Goodman. Assigning stereochemistry to single diastereoisomers by GIAO NMR calculation: the DP4 probability. *J. Am. Chem. Soc.*, 132:12946–12959, 2010. doi: 10.1021/ja105035r.

Yunyue Su, Jiahui Chen, Zao Jiang, Zhenyi Zhong, Liang Wang, Qiang Liu, and Zhaoxiang Zhang. SpectraLLM: uncovering the ability of LLMs for molecular structure elucidation from multi-spectral data. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2508.08441.

Chengrui Xiang, Tengfei Ma, Yujie Chen, Tong Wang, Haowen Chen, and Xiangxiang Zeng. SpecX: a large-scale benchmark for multi-modal spectroscopy and cross-paradigm evaluation. *arXiv*, 2026. doi: 10.48550/arXiv.2605.18791.

Liujia Yang, Zhuo Yang, Jiaqing Xie, Yubin Wang, Ben Gao, Xingjian Wei, Jiaxing Sun, Jiang Wu, Conghui He, Yuqiang Li, and Qinying Gu. NMRTrans: structure elucidation from experimental NMR spectra via set transformers. In *Proc. 32nd ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD '26), Vol. 2*, pp. 12621–12632, 2026. doi: 10.1145/3770855.3818935. arXiv:2602.10158.

Xiang Zhuang, Bin Wu, Jiyu Cui, Kehua Feng, Xiaotong Li, Huabin Xing, Keyan Ding, Qiang Zhang, and Huajun Chen. Boosting LLM's molecular structure elucidation with knowledge-enhanced tree-search reasoning. In *Proc. 63rd Annu. Meet. Assoc. Comput. Linguist. (Vol. 1: Long Papers)*, pp. 22561–22576, 2025. doi: 10.18653/v1/2025.acl-long.1100. URL https://aclanthology.org/2025.acl-long.1100/.

A APPENDIX

This appendix records the blind protocol, scoring identities, scale tables, worked cases, and reproducibility notes. Headline generation on the frozen $n{=}500$ roster is $227/500$ (45.4%) top-1 and $249/500$ (49.8%) recall; self-ranking precision | recall is $227/249$ (91.2%). Independent forward verification of the same pools is §A.8. Confidence intervals are not reported for $n{=}500$. The $n{=}519$ pooled print ($243/519$) is not the headline.

A.1 CROSS-REFERENCE TO THE MAIN TEXT

Table S1: Main-text objects and the appendix sections that discuss them. Figures 1–3 and Listing 1 are not repeated. Figures S1–S3 appear only here.

object	role in this appendix
Figure 1	$n{=}500$ generation: $249/500$ recall, $227/249$ self-rank, 45.4% top-1; wall $227/22/251$.
Listing 1	IRexp record (ILJNJJNKEOAREX; CID 57398578; $C_{12}H_{11}BrN_4$). SMILES and InChIKey are resource fields, not solver payload.
Figure 2	(a) Nested formula-only, non-thinking $n{=}60$ (§A.9.5); (b) source-paper year, $n{=}194$ (§A.9.6); (c)–(d) vendor recall and chamfer, $n{=}60$.
leave-one-modality	Thinking arm, full lists 35/60 (Table 2). Nested full lists are 14/60 (§A.9.5).
Figure 3	v3-R25 verify-save (§A.9).
§A.8	Forward-verify diagnostic on the same pools, $204/45/251$ (table in that section).
Figure S1	Validate-clean chemistry ($n{=}295$; median MW 306 $\mathrm{g{\cdot}mol^{-1}}$; median 2 rings). The thinking-tier expand-500 cut is separate.
Figure S2	Inference ladder on the 60-compound arm ($23\% \to 27\% \to 30\%$).
Figure S3	30 deposited cases: true structure beside the top-1 prediction (exact top-1, in pool but not first, never proposed).
Table S9	Worked cases (§A.9).
validate-clean extras	Scaffold, Tanimoto, stereo, size bins, and corpus-reweight (§A.5).
literature rows	Top-1 / top-k comparison (§A.11.3); separate published settings.

Figure 1, Figure 2, Figure 3, and Listing 1 are in the main text. Main-text Eqs. 1, 2, 3, and 4 (factorisation, InChIKey-14, the three rates, and ^{13}C chamfer) are restated in §A.4 with the same meaning. The pooled estimator lives only here.

A.2 APPENDIX FIGURES

Figure S1 plots molecular weight, ring count, C–F bonds, and nitrogen count on the validate-clean subset (§A.5). Figure S2 is the inference ladder on the 60-compound arm (§A.9.5–§A.9.10). Figure S3 places each deposited true structure beside the top-1 prediction for 30 representative cases of exact top-1, in the pool but not first, and never proposed. Structures and Tanimoto scores are taken only from released deposits; no structures are invented for illustration.

A.3 BLIND PROTOCOL

A.3.1 SOLVER CONTRACT

Each IRSpectra-Bench item is a four-field JSON object. The solver sees *only*

$$\{\texttt{qid, formula, ir_bands_cm-1, h_nmr, c_nmr}\}.$$

No name, SMILES, InChIKey, scaffold, atom-type hint, retrieved spectrum, licence stamp, or PMC id enters the prompt (Listing 1, adapted from one IRexp record). Gold constitution is scored offline by the InChIKey-14 indicator (§A.4).

Difficulty is declared *a priori* by RDKit ring analysis: *simple* iff $\leq$2 rings, no fused/spiro/bridgehead system, $\leq$22 heavy atoms; else *complex*. The $n{=}500$ headline is near-50/50 (248/252); the eligible

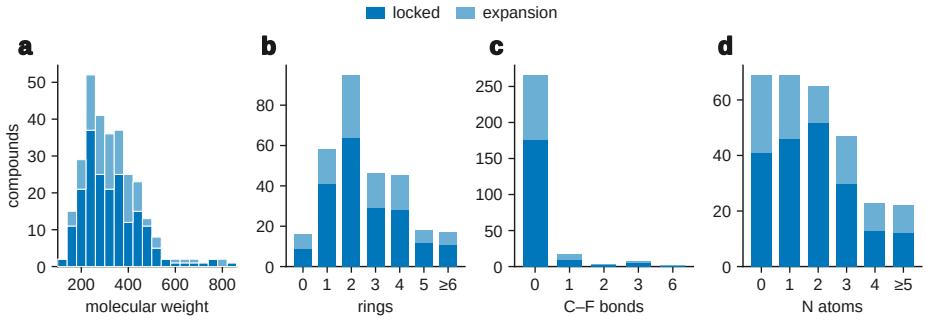


Figure S1: Molecular weight, ring count, C–F bonds, and nitrogen atoms on the validate-clean subset (n=295: locked cohort plus clean expansion; median MW 306 $\mathrm{g{\cdot}mol^{-1}}$; median 2 rings). The thinking-tier expand-500 cut is separate from this subset.

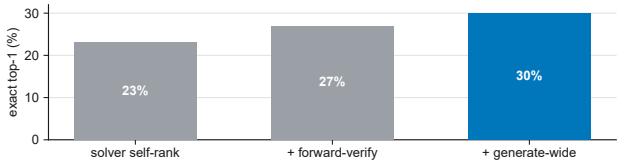


Figure S2: Inference ladder on the 60-compound arm: 23% $\rightarrow$ 27% $\rightarrow$ 30% (generate-wide recall 32% $\rightarrow$ 42%).

IRexp corpus is 17.5% simple / 82.5% complex. Corpus-reweighted extras on a validate-clean subset are in §A.5 and are not n=500 headline numbers.

A.3.2 Frozen roster

The code release records the frozen roster and the expand-500 campaign.

- n=500 roster: frozen `qid` lists in the code release. Sampler seeds for early draws were not captured; the released roster is fixed and reusable but not exactly re-drawable.
- expand-500 campaign: 115/115 simple/complex; seed 2026500; 0 collisions with any prior round. Six ^{13}C-overread flags known *before* any solve: R26, R31, R102, R105, R107, R138.

A.3.3 Stopping and pooling

- The n=500 roster is a pre-specified frozen `qid` list, not an outcome-selected subset.
- Gold keys for the expand-500 arm were used only for local scoring and are not in the review copy.
- Official expand-500 chamfer is 103/230 verify vs 129/230 self (§A.8); confidence intervals were not computed on that arm.

A.3.4 Allowed tools

RDKit formula / parse check only. No web search. No retrieved spectrum. The pooled n=500 headline mixes no-thinking runs with the thinking-tier expand-500 cut (200 of 500). The two harnesses are not interchangeable; scoring is the same InChIKey-14 contract.

A deposit is JSONL of {`qid, candidates: [SMILES,...]`} under a stated budget. Headline budget: $\leq$ 3 ranked SMILES per `qid`. Self-ranking and forward-verified re-ranking are reported separately.

A.4 Scoring identities

The identities in this section restate main-text Eqs. 1, 2, 3, and 4. The pool estimator (Eq. S7) is not in the main text.

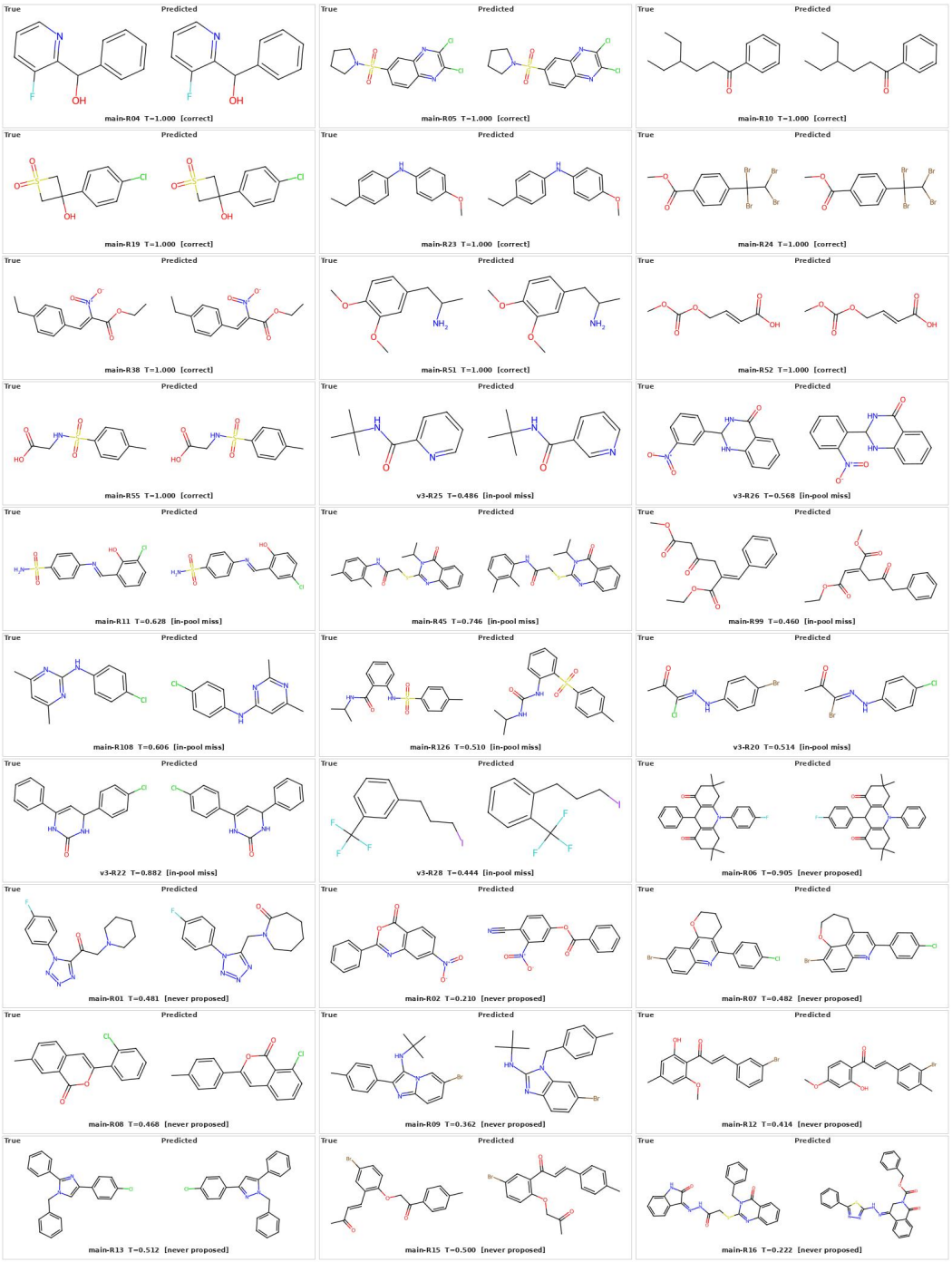


Figure S3: True vs top-1 predicted structures for 30 representative deposited IRSpectra-Bench cases (10 correct top-1, 10 in-pool miss, 10 never-proposed). Labels under each pair give qid, Morgan r=2 Tanimoto, and outcome.

A candidate $\hat{y}$ matches gold $y^\star$ at constitution iff

$$s(\hat{y}, y^\star) = \mathbf{1}\big[\mathrm{IK}_{14}(\hat{y}) = \mathrm{IK}_{14}(y^\star)\big]. \tag{S1}$$

IK_{14} is the first 14 characters of the RDKit InChIKey (connectivity layer). Full stereo-sensitive InChIKey is reported only on validate-clean n=295 (§A.5).

With ranked pool $\mathcal{C}_i = (c_{i1}, \ldots, c_{ik})$, $k \le 3$,

$$\text{top-1} = \frac{1}{n}\sum_i s(c_{i1}, y_i^\star), \tag{S2}$$

$$\text{recall@}k = \frac{1}{n}\sum_i \mathbf{1}\big[\exists\, j \le k :\ s(c_{ij}, y_i^\star) = 1\big], \tag{S3}$$

$$\text{prec.|recall} = \frac{\sum_i s(c_{i1}, y_i^\star)}{\sum_i \mathbf{1}[y_i^\star \in \mathcal{C}_i]}. \tag{S4}$$

These are the three reportable numbers. They factorise as

$$\text{top-1} = (\text{recall@}k) \times (\text{prec.|recall}) \tag{S5}$$

when the same pool is used for both stages. The rates have different denominators and must not be differenced. Self-ranking prec.|recall on n=500 is $227/249$ (91.2%), which is not forward-verify precision.

Forward-verify re-ranks by symmetric chamfer on ^{13}C peak sets $P, Q \subset \mathbb{R}$

$$d(P,Q) = \frac{1}{|P|}\sum_{p\in P}\min_{q\in Q}|p-q| + \frac{1}{|Q|}\sum_{q\in Q}\min_{p\in P}|q-p|, \tag{S6}$$

where $\hat{P}(c)$ is predicted *blind* to the observed list P_{obs}. An instrumented diagnostic on a protocol slice is $58/65$ (§A.8). The n=500 forward-verify diagnostic is $204/45/251$ (§A.8). The expand-500 official arm score is $103/230$ verify vs $129/230$ self (§A.8).

A pool estimator on an index set $\mathcal{S}$ is

$$\hat{p}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} s(c_{i1}, y_i^\star). \tag{S7}$$

The paper headline uses $\mathcal{S}$ of size 500 (frozen `qid` list). The n=519 `score_pooled.py` print ($243/519$) is not used.

Confidence intervals on earlier validate-clean subsets are bootstrap 95% (`score_main.boot`, seed 0, 2000 resamples). Paired model differences use McNemar with Holm correction. Confidence intervals are not reported for n=500 or expand-500.

A.5 VALIDATE-CLEAN EXTRAS

Cohort chemistry is Figure S1. On the pre-expansion validate-clean subset, ^{13}C-overread flags (R12, R22, R25, R82, R91) were marked before the subset totals. They are not the six expand-500 flags listed above.

Table S2: Scaffold and stereo metrics on the validate-clean subset (n=295).

metric	overall	simple	complex
scaffold-level (best Tanimoto ≥ 0.45)	64%	80%	48%
mean best Tanimoto	0.66	0.79	0.53
full InChIKey (stereo) top-1	93/295 (31.5%) [26–37]	—	—
full InChIKey recall	108/295 (36.6%) [31–42]	—	—
corpus-reweighted top-1	**26.5%** [21–32]	—	—
corpus-reweighted recall	**31.3%** [25–37]	—	—

On an instrumented slice, of 137 analysable top-1 misses, 76.6% are constitutional isomers of the true structure; 22.6% share the true Murcko scaffold.

Table S3: Size bins show top-1 and recall on validate-clean n=295.

heavy atoms	n	top-1	recall
≤15	53	36/53 (67.9%)	39/53 (73.6%)
16–25	152	65/152 (42.8%)	75/152 (49.3%)
>25	90	15/90 (16.7%)	16/90 (17.8%)

Table S4: expand-500 scores show generation under `score2` (InChIKey-14; confidence intervals not reported).

set	n	top-1	recall@3
Claude Opus 5, all deposited	230	129/230 (56.1%)	138/230 (60.0%)
simple	115	82/115 (71%)	87/115 (76%)
complex	115	47/115 (41%)	51/115 (44%)
validate-clean	224	127/224 (56.7%)	135/224 (60.3%)
simple	110	81/110 (74%)	85/110 (77%)
complex	114	46/114 (40%)	50/114 (44%)

A.6 EXPAND-500 GENERATION SCORES

Gold keys for this arm were used only for local scoring and are not in the review copy.

Mean best Tanimoto (`score2` footer, all 230): simple 0.863; complex 0.645. Flagged qids (in 230, out of 224): R26, R31, R102, R105, R107, R138. Deposits: 230/230; `predictions2.jsonl` 230 lines, 690 candidates. Unique (formula, IR, ^{13}C) vs `irexp_resolved.jsonl.gz`: 230/230; 0 prior-round InChIKey-14 collisions. This arm used a thinking-tier consumer protocol. 200 of the validate-clean cut is inside the n=500 headline; the all-230 table is not interchangeable with the no-thinking rows.

Verbatim `score2` footer:

```
overall recovered (top-3): 138/230 (60%)
top-1 exact: 129/230 (56%)
  simple  : recovered 87/115 (76%)
            top1 82/115 (71%)  meanBestTani 0.863
  complex : recovered 51/115 (44%)
            top1 47/115 (41%)  meanBestTani 0.645
```

A.7 HEADLINE GENERATION SCORES

The estimator is Eq. S7. The paper headline is n=500 ($227/500$ top-1, $249/500$ recall). Roster: frozen `qid` lists in the code release. The headline pools two consumer protocols. They are not a controlled comparison.

Table S5: Generation on the two protocols inside the n=500 headline (no confidence interval). $118 + 109 = 227$ and $133 + 116 = 249$.

set	n	top-1	recall@3
earlier rows, no-thinking	300	118/300 (39.3%)	133/300 (44.3%)
thinking-tier expand-500 cut	200	109/200 (54.5%)	116/200 (58.0%)
headline pool	**500**	**227/500 (45.4%)**	**249/500 (49.8%)**

Strata on the n=500 headline (same InChIKey-14 rule): simple 161/248 (65%) top-1 / 172/248 (69%) recall; complex 66/252 (26%) / 77/252 (31%). Self-rank precision | recall: $227/249$ (91.2%). The n=500 forward-verify diagnostic is $204/45/251$ (§A.8).

A.8 FORWARD-VERIFY DIAGNOSTIC

The table in this section is the decomposition of those pools. Protocol-slice forward verification and the expand-500 chamfer scores follow. Validate-clean scaffold, size, and stereo extras are in §A.5.

Classification: **verified** = recalled and chamfer-ranked first; **misranked** = recalled, distractor first; **never-proposed** = true structure not in the candidate set. The forward-verify diagnostic is the n=500 row. The three arm rows sum to that diagnostic.

Table S6: Forward-verify split on n=500: $204/45/251$ (no confidence interval). Arm rows sum to that split.

arm	n	verified	misranked	never-proposed
locked protocol slice	194	58	7	129
Claude Opus 5 expansion	106	56	12	38
expand-500 cut	200	90	26	84
n=500 **diagnostic**	**500**	**204**	**45**	**251**

Recalled $= 204 + 45 = 249/500$, matching generation recall. Figure 1 reports the generation wall ($249/500$ recall $\times$ $227/249$ self-rank $= 45.4\%$ top-1; wall $227/22/251$).

A.8.1 Protocol-slice diagnostic

Source: `data/diagnosis.json` / `scripts/forward_verify_all.py`. The rows below are the protocol slice, separate from the n=500 split ($204/45/251$).

Table S7: Protocol-slice forward verification (separate from the n=500 split $204/45/251$).

	60-compound arm	protocol slice
generation recall	19/60 (32%)	65/194 (33.5%)
top-1, solver self-ranking	14/60 (23%)	55/194 (28.4%)
top-1, forward-verified re-ranking	16/60 (27%)	58/194 (29.9%)
precision | recall — fverify	16/19 (84%)	58/65 (89.2%)
multi-candidate only — fverify	10/13 (77%)	30/37 (81%)
derangement floor on n=37	—	54.0%; fverify +27.1 points
McNemar vs self-rank	—	p=0.55 (7 gained, 4 lost)

Self-ranking $227/249$ on n=500 is generation, not a substitute for $204/45/251$. The $14/60$ self-rank is the nested n=60 arm (Figure 2a), not thinking leave-one full lists (Table 2, $35/60$).

A.8.2 Official expand-500 score

Blind ^{13}C campaign over 681 unique candidate SMILES (`f1–f41` all present). Verify does not beat self-rank. Headline generation remains n=500 ($227/500$, $249/500$). This arm score does not replace the n=500 forward-verify table.

Table S8: Official expand-500 chamfer scores.

set	n	recall	self top-1	verify top-1	verify | recall
all deposited	230	138/230 (60.0%)	129/230 (56.1%)	103/230 (44.8%)	103/138 (74.6%)
validate-clean	224	135/224 (60.3%)	127/224 (56.7%)	103/224 (46.0%)	103/135 (76.3%)

Coverage: 41/41 batches; 681/681 unique SMILES; 230/230 qids fully covered. Multi-candidate only: all-230 self 128/137, verify 102/137; clean-224 self 126/134, verify 102/134. `forward_verify_all.py` was not run (no expand-500 `diagnosis.json`).

HOSE-style lookup on the protocol-slice candidate sets ties self-ranking; a modest GNN reaches 59/65 (91%) vs the LLM verifier's 58/65 (89%). Deranging observed ^{13}C pairings collapses precision (p=0.001).

A.9 Worked cases

The three failure modes below are taken from released candidate files. Later panels reuse already-reported numbers. Expansion qids reuse R22/R25. The verify-save drawing is Figure 3.

Table S9: Worked cases show three failure modes from released candidate files (^{13}C chamfers in ppm; Tanimoto Morgan r=2).

mode	id / formula	what failed
Proposal miss	main-R06 / $C_{29}H_{30}FNO_2$	True acridinedione never in the 3-candidate pool (Tanimoto-0.90 regioisomer: 4-fluorophenyl / phenyl swapped). IR 1641, 1510 cm^{-1}; ^{13}C 195.7. Verifier unused.
Verify-save	v3-R25 / $C_{10}H_{14}N_2O$	True 2-pyridyl picolinamide at self-rank 3; self-rank 1 is 3-pyridyl. Forward-verify recovers (chamfer 0.42 vs 1.30 ppm). IR 1679 cm^{-1}. Figure 3.
Verify-fail	v3-R26 / $C_{14}H_{11}N_3O_3$	True 3-nitrophenyl dihydroquinazolinone in pool (self-rank 2); 2-nitro ranked first. Chamfer 1.35 vs 1.36 ppm. IR 1652, 1530, 1352 cm^{-1}.

A.9.1 Proposal miss: main-R06 / $C_{29}H_{30}FNO_2$

True acridinedione never enters the 3-candidate pool (Tanimoto-0.90 regioisomer: 4-fluorophenyl / phenyl swapped). IR 1641, 1510 cm^{-1}; ^{13}C 195.7. Verifier unused. Re-ranking cannot recover a structure that was never proposed (Eq. S5).

A.9.2 Verify-save: v3-R25 / $C_{10}H_{14}N_2O$

True 2-pyridyl picolinamide at self-rank 3; self-rank 1 is 3-pyridyl. Forward-verify recovers (chamfer 0.42 vs 1.30 ppm, Eq. S6). IR 1679 cm^{-1}. The drawing is Figure 3.

A.9.3 Verify-fail: v3-R26 / $C_{14}H_{11}N_3O_3$

The true structure is the 3-nitrophenyl isomer, in the pool at self-rank 2. The 2-nitro isomer is ranked first (chamfer 1.35 vs 1.36 ppm). IR 1652, 1530, 1352 cm^{-1}. Ranking can fail when predicted ^{13}C lists are nearly tied (McNemar p=0.55; 7 gained, 4 lost on the protocol slice).

A.9.4 Example payload

Listing 1 is the IRexp JSON record ($C_{12}H_{11}BrN_4$; PubChem CID 57398578). The InChIKey is `ILJNJJNKEOAREX-UHFFFAOYSA-N`. SMILES, InChIKey, and licence fields are resource metadata. Bench solvers see only formula and printed peak lists. Gold constitution is scored only by Eq. S1.

A.9.5 Formula-only control (n=60)

Nested contamination control on the non-thinking n=60 roster (Figure 2a). The thinking leave-one-modality arm is a separate count (Table 2, full lists 35/60). Formula only: 3/60 (5%) top-1 and 3/60 (5%) recall. Formula + IR stays at that formula-only floor (3/60 top-1, 5%). Formula + IR + ^{1}H + ^{13}C: 14/60 (23%) / 19/60 (32%). Outcomes perfectly nested (McNemar p=0.001).

A.9.6 Source-paper year (n=194)

Accuracy is flat in source-paper year on the n=194 slice (r=−0.007; Figure 2b). The correlation bounds verbatim recall of specific experimental sections. It does not exclude contamination.

A.9.7 Generate-wide (60-compound arm)

Ten independent solvers, up to six candidates each, pooled then re-ranked: recall 32% → 42%, top-1 23% → 30% (McNemar p=0.34); verification precision falls 84% → 72%.

A.9.8 IRexp-fine-tuned generator (locked slice)

Pooled with Claude Opus 5: recall 33.5% → 54.1%, top-1 28.4% → 35.1% (McNemar p=0.015). Training on the same peak-list resource moves $P(\text{true} \in \text{pool})$.

A.9.9 BATTERY-ELECTROLYTE SUBSET

n=46 literature electrolyte chemistry (not operando): 26% top-1, recall-bound.

A.9.10 FOUR-MODEL CLAUDE LADDER (n=24)

A Haiku, Sonnet, Claude Opus 5, and Fable comparison on n=24 is underpowered for adjacent ranks and is not a headline result.

A.10 PROMPTS AND DEPOSIT SCHEMA

A.10.1 HEADLINE PROMPT

The headline Claude wrapper was not captured (main text, Experimental setup and Limitations). Only the per-compound four-field payload is released. The skeleton below is the contract, not a recovered verbatim system prompt.

```
You are given only:
  formula          (as from HRMS)
  ir_bands_cm-1    (author-transcribed experimental list)
  h_nmr, c_nmr     (printed shift lists)
Return <= 3 ranked SMILES.
No name, no SMILES hint, no scaffold, no retrieved spectrum.
Check molecular formula with RDKit before emitting a candidate.
```

The solver-facing fields of Listing 1 are `formula`, `ir_bands_cm-1`, `h_nmr`, and `c_nmr`. SMILES, InChIKey, and licence fields in that record are not part of the prompt.

A.10.2 FORWARD-VERIFY PROMPT

For each candidate SMILES, predict a ^{13}C peak list without seeing the observed list. Re-rank by chamfer distance (Eq. S6). The instruction dispatched to the forward-prediction step was not captured; only the candidate SMILES and the resulting numeric lists are on disk for the protocol slice.

A.10.3 CROSS-VENDOR HARNESS

Verbatim prompts in the code release are the cross-vendor harness prompts, not the missing headline Claude wrapper. External submissions must document model version, prompt, and tool access (`docs/LEADERBOARD.md`).

A.10.4 DEPOSIT SCHEMA

```
{"qid": "R01", "candidates": ["SMILES_rank1", "SMILES_rank2", "SMILES_rank3"]}
```

Score locally (code release). Frozen deposits and scorers are in the anonymous review mirror at `https://anonymous.4open.science/r/peaklist-corpus-review-10C4/` (`scripts/` and `SCHEMA.md`):

```
python scripts/score_submission.py
    --predictions my_run.jsonl --name "YourModel-1.0"
python scripts/score_pooled.py --stereo
python scripts/forward_verify_all.py
```

A.11 MODEL CARDS

A.11.1 HEADLINE MODEL CARD

Table S10: Headline solver card. Fields that were not recorded are marked as such.

item	value
headline solver	Claude Opus 5, consumer subscription
thinking tier	mixed: earlier rows none; expand-500 cut (200 of the n=500 headline) thinking-tier
snapshot / seed / temperature / top_p / top_k / max_tokens	not exposed; not recorded
candidate budget	$\leq$ 3 SMILES / qid
tool access	RDKit formula / parse check only; no web search
fine-tuning	none on the headline arm
paid API	no, for the core protocol
hardware	consumer harness; no cluster job for inference
scoring hardware	commodity CPU; RDKit InChIKey-14
cross-vendor arm	n=60; Claude Opus 5, Grok 4.6, Gemini 3.7 Flash, GPT-5.6 Sol
Claude Opus 5 candidates	2.20 mean (others 3.00) on that arm — recall rankings approximate
expand-500 solver	Claude Opus 5 thinking-tier; 200 of this arm is inside the n=500 headline
Incomplete additional-model deposits	not scored; not pooled
scoring scripts	`scripts/score_pooled.py`, `score_main.py`, `benchmark_v2.py score2`

A.11.2 CROSS-VENDOR MODEL CARD (n=60)

Table S11: Cross-vendor scores on n=60. Recall is pool membership. Both precision columns are chamfer forward-verify, not solver self-rank. Claude Opus 5 self-rank top-1 on this nested arm is 14/60 (thinking leave-one full lists are 35/60).

model	generation recall	chamfer prec.|recall	chamfer, multi-cand.
Claude Opus 5	19/60 = 32%	16/19 = 84%	10/13 = 77%
Grok 4.6	32/60 = 53%	20/32 = 62%	20/32 = 62%
Gemini 3.7 Flash	30/60 = 50%	22/30 = 73%	22/30 = 73%
GPT-5.6 Sol	25/60 = 42%	17/25 = 68%	16/24 = 67%

Verification precision exceeds generation recall in every arm. Bootstrap separates the paired gap for Claude Opus 5 (+52.5 points), GPT-5.6 Sol (+26.3) and Gemini (+23.3); Grok's gap (+9.2) is directional. A clean-clone control for Grok bounds key leakage (McNemar p=0.39). Reasoning-effort tier per vendor arm was not recorded.

A.11.3 LITERATURE COMPARISON

Any system reporting top-1 $= a$ and top-$k = b$ implies recall $\geq b$ and conditional precision $\leq a/b$. Script: `scripts/literature_decomposition.py`.

Table S12: Selected published systems, connectivity-scored or as published. Separate settings, not scores on this roster.

system (n)	top-1	top-k (k)	recall	prec.
this work, solver alone (500)	45.4%	49.8% (3)	49.8%	91.2%
this work + forward-verify (slice)	29.9%	33.5% (3)	33.5%	89.2%
NMR-Solver, literature (450)	52.9%	67.3% (10)	$\geq$67.3%	$\leq$78.6%
NMRAgent, literature (450)	61.6%	70.0% (10)	$\geq$70.0%	$\leq$88.0%
Espejo Morales, education (236)	80.9%	90.0% (5)	$\geq$90.0%	$\leq$89.9%
Espejo Morales, industrial (34)	20.6%	29.1% (5)	$\geq$29.1%	$\leq$70.9%
Alberts, 6–13 atoms, NIST IR (3,455)	63.8%	84.0% (10)	84.0%	75.9%
SpecX, scaffold split, simulated (99,439)	29.7%	50.6% (10)	50.6%	58.7%

These rows are separate published settings, not one paired collapse onto literature peak lists. The solver-alone precision column on n=500 is self-ranking (227/249); the forward-verify row is the instrumented protocol slice (58/65).

A.12 IREXP RESOURCE

Peak lists for IRSpectra-Bench are drawn from IRexp (121,233 records; 57,646 structure-linked; 39,118 full IR+^{1}H+^{13}C+structure quadruples; `irexp_resolved`). IRSpectra-Bench is the frozen peak-list roster and the factorised diagnosis (generation wall $227/22/251$; forward-verify diagnostic $204/45/251$). During double-blind review the commercial redistributable pool is available from an anonymised copy (no author-identifying host):

`https://anonymous.4open.science/r/peaklist-corpus-review-10C4/`

Named archival hosting replaces that URL at camera-ready. File schema and aggregate counts ship with the copy (`SCHEMA.md`, `data/`). Spectral validation filters, the prompt-leakage audit and electrolyte SMARTS live in the code release. Fine-tunes must use `data/train_no_bench.jsonl.gz` so benchmark InChIKeys are withheld. One record is shown in Listing 1.

A.13 OFFLINE PEAK-LIST BASELINE (NMRTRANS)

NMRTrans (Yang et al., KDD 2026; arXiv:2602.10158; code commit `9a72756`; checkpoint `little1d/C-H-Formula`, file `nmrtrans-c-h-nmr-formula.ckpt`, 499/499 weights loaded) is a set transformer with a T5-small decoder. It reads a molecular formula and unordered ^{1}H and ^{13}C peak sets. It has no IR channel. This run is offline CPU inference, beam size 3, no sampling, and no paid API. Questions are the frozen n=500 roster (spectro-agent `48fcd30`). Gold is a unique match of formula, IR bands, and the ^{13}C string into `irexp_resolved` (500/500, none ambiguous). The generator did not see SMILES or InChIKeys. Scoring is RDKit InChIKey-14. No confidence intervals.

Printed lists only. A ^{1}H range becomes a midpoint and a half-width; a single shift has half-width 0. Multiplicity is mapped onto the model's split vocabulary; unknown patterns stay at the unknown index (87 peaks). Printed integrals and J values are copied. 76 of 4,298 proton peaks had no printed integral; those values were stored as 0 and were not filled in as 1H. A ^{13}C shift is repeated only when the list prints an nC count; the model treats that repeat as its intensity channel. Nineteen shifts outside $[0, 220]$ ppm were clamped by the published $/220$ normalisation. Nine formulae contain Se or Sn, which the released 12-element formula vector drops. IR bands were not passed in.

Table S13: NMRTrans on the n=500 roster (InChIKey-14, beam 3). No confidence intervals. Held-out drops gold keys that appear in the released NMRTrans train, validation, or test splits.

set	n	top-1	recall (top 3)
all	500	79/500 (15.8%)	102/500 (20.4%)
simple	248	62/248 (25.0%)	79/248 (31.9%)
complex	252	17/252 (6.7%)	23/252 (9.1%)
locked	194	32/194 (16.5%)	39/194 (20.1%)
expansion	106	12/106 (11.3%)	20/106 (18.9%)
expand-500 200-cut	200	35/200 (17.5%)	43/200 (21.5%)
held out of released splits	489	70/489 (14.3%)	92/489 (18.8%)

Nine gold InChIKey-14 values occur in the released training split (8/9 top-1, 9/9 inside the beam) and two in the validation split (1/2 top-1). None of the 500 are in the released test split. Overlap is constitution identity, not a check that the same peak list was in the training file. 86/500 rank-1 strings were not valid RDKit SMILES. Predictions are deposited under `data/baseline_nmrtrans/`. NMR-Solver, SpectraLLM, IR-Agent, and commercial CASE were not rerun.

A.14 REPRODUCIBILITY

Frozen questions, predictions, and scorers ship with the code release (`https://anonymous.4open.science/r/peaklist-corpus-review-10C4/`). The n=500 generation totals $227/500$ and $249/500$ regenerate from the released deposits. The n=500 forward-verify table is §A.8. The protocol-slice diagnostic is regenerated by `python scripts/forward_verify_all.py`.

Expand-500 gold keys are withheld from the review copy. The official expand-500 chamfer score is 103/230 verify vs 129/230 self.

The IRexp bibliographic entry is anonymous in the review copy; the dataset URL in the PDF is the anonymous.4open review copy. Fine-tunes should use `data/train_no_bench.jsonl.gz`.

Inference is not bit-exact: no pinned snapshot, temperature, or seed was recorded. Distributional agreement of frozen deposits is the intended claim. Unrecorded fields: headline wrapper text; model snapshots; decoding parameters; sampler seeds.

On-bench scores for IR-Agent, CASE, SpectraLLM, NMR-Solver, Spectro, NMIRacle and the Alberts transformers were not obtained. Those systems are blocked on input, licence, or compute, and no adapter was built. Formula-only recovery on the nested 60-compound control is 3/60 top-1 and 3/60 recall. NMRTrans (§A.13) is one trained NMR-only baseline on this roster (IR unused; 79/500 top-1), not a full multi-modal peer and not a score for those systems. An expert-chemist audit was not run. Leave-one-modality on the paired thinking n=60 arm is Table 2 (full lists 35/60).

A.15 CLOSED-BOOK HARNESS

The closed-book harness is the operational form of the Blind protocol (§A.3). Each solver call receives only the four-field peak-list payload {`qid, formula, ir_bands_cm-1, h_nmr, c_nmr`}: no compound name, SMILES, InChIKey, scaffold hint, retrieved spectrum, or web search. The only tool is an RDKit formula / parse check before a candidate is emitted.

The n=500 headline uses Claude Opus 5 on a consumer harness. Thinking-tier mix matches the model card (§A.11): earlier rows without thinking; the expand-500 cut (200 of 500) on the thinking tier. Each deposit holds at most three ranked SMILES; scoring is RDKit InChIKey-14 as in §A.4. The cross-vendor n=60 arm reuses the same payload with vendor-specific prompts in the code release.

What was not recorded is unchanged from the model card and Reproducibility (§A.14): headline wrapper text, model snapshots, temperature, and sampler seeds. Frozen deposits are the intended reproducibility claim, not bit-exact re-inference.