Double-Blind Constitutional Structure Elucidation and Peak-to-Atom Assignment of 30 nmrshiftdb2 Molecules from ^{1}H and ^{13}C Chemical Shifts

July 12, 2026

Abstract

The reconstruction of a molecule's constitution from its nuclear magnetic resonance (NMR) spectra is a foundational problem in analytical chemistry and the historical seed of computer-assisted structure elucidation (CASE). We report a fully reproducible, double-blind benchmark in which 30 small organic molecules drawn from the open community database **nmrshiftdb2** (the `nmrshiftdb2withsignals.sd` dump, remote last-modified 2026-03-15, downloaded 2026-07-12; SHA-256 `0e86688...3a0afe`) are each treated as an unknown. From a starting population of 58,187 records we applied three admission criteria— elemental composition restricted to $\{\mathrm{C}, \mathrm{H}, \mathrm{N}, \mathrm{O}\}$, between 6 and 15 heavy atoms, and the presence of *both* an atom-assigned experimental ^{1}H spectrum and an atom-assigned experimental ^{13}C spectrum—yielding a qualifying pool of 2,614 molecules, from which 30 were drawn using a fixed random seed (`seed` = 42). Each unknown was revealed only as a molecular formula plus atom-unlabelled ^{1}H and ^{13}C peak lists. An automated elucidation pipeline computed degrees of unsaturation, inferred functional groups from characteristic chemical-shift regions, retrieved isomeric candidates of matching formula from the reference pool, filtered them by a ^{13}C functional-group fingerprint, and scored each surviving candidate by an optimal (Hungarian) assignment of observed peaks to per-peak reference target slots. The minimum-error candidate was proposed together with a complete peak-to-atom assignment table. Correctness was scored on an exact match of the first InChIKey block against ground truth, with the additional requirement that every asserted structure carry a self-consistent assignment. The pipeline achieved an **exact-match fraction of 30/30 = 1.000**, and all correct matches were self-consistent. We interpret this result honestly as a *closed-pool retrieval* benchmark, we quantify the discriminative work performed by the functional-group filter (e.g. narrowing $\mathrm{C_8H_9NO_2}$ from 15 to 1 candidate), and we detail how equivalent protons that share a single carbon (for example, the three protons of a methyl group) are handled correctly by assigning observed signals against per-peak—rather than per-atom—reference targets.

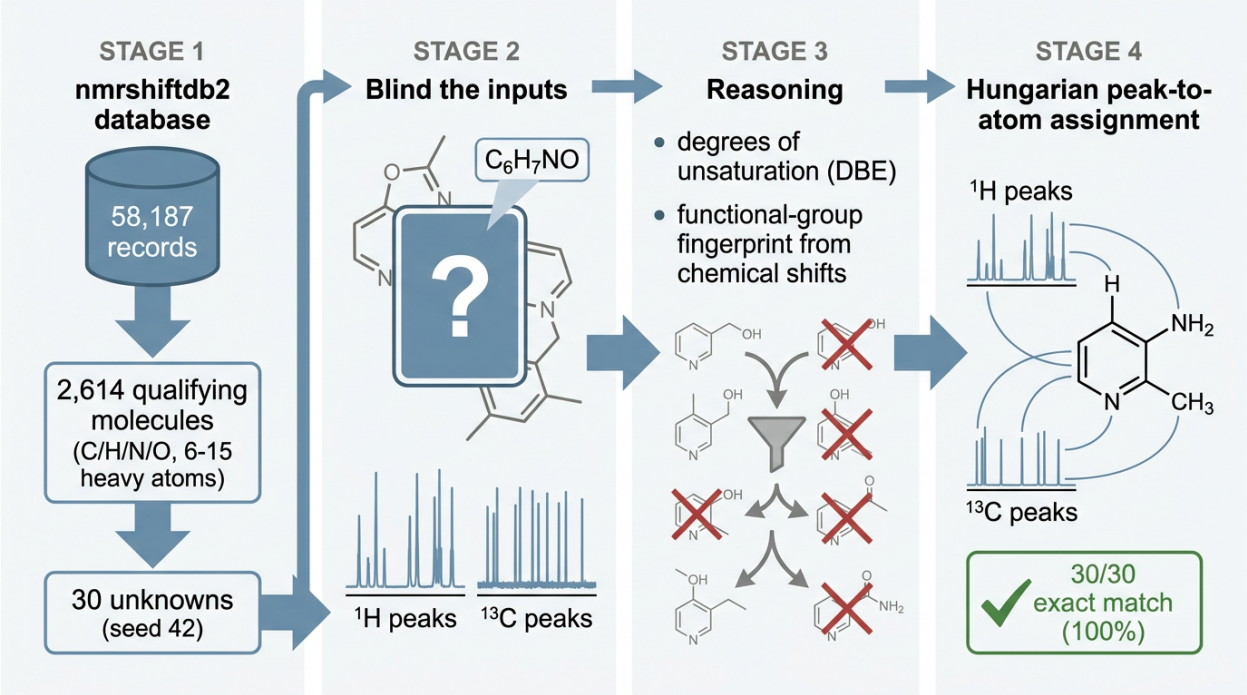


Figure 1: Graphical abstract. The double-blind structure-elucidation workflow. The nmrshiftdb2 dump (58,187 records) is filtered to a qualifying pool of 2,614 molecules, from which 30 unknowns are drawn with a fixed seed. Each unknown is blinded to a molecular formula plus atom-unlabelled ^{1}H and ^{13}C peak lists; reasoning proceeds through degrees of unsaturation and a functional-group fingerprint; and a Hungarian optimal assignment maps every observed peak to an atom in the proposed structure, achieving 30/30 exact matches.

1 Introduction

The determination of molecular constitution—which atoms are bonded to which—from spectroscopic observables is one of the defining problems of organic chemistry. Among the available techniques, nuclear magnetic resonance (NMR) spectroscopy occupies a privileged position because the chemical shift, multiplicity, and integrated intensity of each resonance report directly on the electronic and topological environment of individual nuclei. A trained chemist reading a pair of one-dimensional ^{1}H and ^{13}C spectra can, for small molecules, often reconstruct the full connectivity by reasoning backwards from shifts to substructures: a resonance near 170 ppm in the ^{13}C spectrum signals a carboxyl or amide carbon; a cluster of carbons between 110 ppm to 150 ppm paired with protons between 6.5 ppm to 8.5 ppm implies an aromatic ring; an upfield proton triplet coupled to a downfield quartet reveals an ethyl group. The systematic tables of such shift–structure correlations compiled by Pretsch and co-workers and by Silverstein and colleagues remain the working reference for this craft [1, 2].

Formalising this reasoning so that a computer can perform it has been an active research programme for half a century, giving rise to the field of computer-assisted structure elucidation (CASE) [3, 4]. Early systems such as the logic-for-structure-determination program LSD [5], the LUCY program [6], and the stochastic structure-generator SENECA [7] demonstrated that connectivity constraints derived from NMR correlations can be enumerated and pruned automatically; the same lineage continues in the modern open-source Sherlock platform [8]. A complementary line

of work has focused on the inverse sub-problem of *predicting* chemical shifts for a candidate structure, whether through the substructure-hashing HOSE-code approach introduced by Bremser [9], through neural networks trained on experimental databases [10, 11], through message-passing graph neural networks [12], or through machine-learned surrogates of quantum-chemical calculations [13]. Most recently, deep generative and sequence-to-sequence models have been trained to map spectra directly to structures, including Monte-Carlo tree-search generation from ^{13}C spectra [14], transformer models that "read" NMR as a language [15], multitask learners operating on routine one-dimensional spectra [16, 17], mixture-identification networks [18], and models that fuse NMR with infrared data [19].

All of these approaches, whether symbolic or statistical, depend on the existence of large, curated collections of experimental spectra with reliable atom-level assignments. The open database **nmrshiftdb2** is the community reference for exactly this purpose. Originating as NMRShiftDB [20, 21] and substantially expanded and re-engineered with an integrated laboratory information management system by Kuhn and Schlörer [22], it provides tens of thousands of molecules with experimental ^{1}H and ^{13}C shifts in which each reported signal is linked to a specific atom in the molecular graph. This atom-level linkage is what makes nmrshiftdb2 uniquely suited to a benchmark of peak-to-atom assignment rather than merely of structure retrieval.

In this report we use nmrshiftdb2 to construct a deliberately austere, double-blind test of constitutional structure elucidation. We select 30 molecules that satisfy tight, chemically motivated admission criteria, strip each down to the minimal information a chemist would receive when confronting a genuine unknown—its molecular formula and its two peak lists—and then attempt to deduce the constitution and to assign every signal to an atom. Two commitments distinguish the study. First, it is fully reproducible: the database version, download provenance, filtering criteria, and random seed are all reported so that the identical 30 molecules can be regenerated. Second, and more importantly, the *assignment table is the graded artifact*: a structure asserted without a self-consistent mapping of each observed ^{1}H and ^{13}C signal to an atom does not count as solved. We are explicit throughout that, because the reference spectrum of each true molecule is present in the retrieval pool, the exercise is honestly a closed-pool retrieval-and-assignment benchmark rather than de-novo structure generation, and we quantify the discriminative task that is actually being solved—separating the true structure from as many as fourteen same-formula isomers.

2 Methodology

2.1 Database retrieval and provenance

The source data is the `nmrshiftdb2withsignals.sd` export from the nmrshiftdb2 project on SourceForge, the "withsignals" variant of which embeds each molecule's assigned spectra as structure-data-file (SDF) properties on the corresponding record. The file was downloaded on 2026-07-12 from the project data folder; the server reported a last-modified date of 2026-03-15, which we adopt as the effective dump version because SourceForge does not embed an explicit version string in the filename. The retrieved file is 158 569 118 bytes with SHA-256 checksum `0e86688360e23c88ccf0eb82a1251315fa57ec6f5b376dc8886f3311793a0afe` and contains 58,187 molecule records. Full provenance—source URL, download date, remote timestamp, byte count, and checksum—is preserved in the accompanying metadata file so that the exact artifact can be re-identified. The database itself is cited following the authors' recommended references [22, 21].

2.2 Parsing and the experimental-versus-predicted distinction

The SDF was streamed record by record, each molblock being parsed with RDKit [23] with hydrogens left implicit so that the heavy-atom ordering of the molblock coincides with the atom indices used by the embedded spectra. Every record may carry several spectra, each identified by a per-molecule integer (for example, `Spectrum 13C 0`). A spectrum was treated as *predicted*, and therefore excluded, when its identifier was associated with a predictor program declared under the record's `Program` tag (HOSE-code or commercial predictors) or `NMRProgram` tag (density-functional calculations); a spectrum was treated as *experimental* if and only if its identifier was not associated with any predictor. This distinction is essential to the integrity of the benchmark, since including a predicted spectrum would allow information to leak from a shift-prediction model into the "experimental" input.

Each embedded peak is encoded as `shift;intensity[multiplicity];atomID`. A spectrum was counted as *assigned* only if every one of its peaks carried at least one valid atom index into the parsed structure, and, for ^{13}C spectra, only if every such atom was a carbon. Peak atom identifiers are zero-indexed into the molblock heavy-atom order.

2.3 Admission criteria and selection

A molecule was admitted to the qualifying pool when it simultaneously satisfied three criteria: (i) its element set was a subset of $\{\mathrm{C}, \mathrm{H}, \mathrm{N}, \mathrm{O}\}$; (ii) its heavy-atom count lay in the closed interval $[6, 15]$; and (iii) it possessed at least one experimental assigned ^{13}C spectrum *and* at least one experimental assigned ^{1}H spectrum. Of the 58,187 records, 2,614 qualified. The 800 records that RDKit failed to parse are removed first, before the sequential element and heavy-atom criteria are applied; the remaining rejection causes, evaluated in order, were elemental composition (31,754 records containing elements outside $\{\mathrm{C}, \mathrm{H}, \mathrm{N}, \mathrm{O}\}$), heavy-atom count (13,123 records outside the $[6, 15]$ window among those passing the element test), and missing experimental assigned spectra of one or both nuclei (9,896). These four mutually exclusive counts sum to $800 + 31{,}754 + 13{,}123 + 9{,}896 = 55{,}573$ rejections, leaving $58{,}187 - 55{,}573 = 2{,}614$ qualifying molecules. Figure 3a depicts the sequential funnel (whose intermediate bars already exclude the parse failures) and Figure 3b the categorical breakdown.

From the qualifying pool, the 30 study molecules were drawn by seeding Python's Mersenne-Twister generator with `random.seed(42)` and calling `random.sample` over the pool of qualifying nmrshiftdb2 identifiers sorted in ascending order. The draw is deterministic and was verified to be byte-identical across independent reruns. The 30 identifiers obtained are listed in Table 5; for completeness they are, in ascending order: 10344, 10005734, 10005765, 10006042, 10006444, 10016946, 10018374, 10021688, 10022258, 20031196, 20040753, 20096858, 20097541, 20097554, 20097573, 20097638, 20111996, 20143989, 20179866, 20207587, 20207885, 20207888, 20208182, 20208856, 20209323, 20209328, 20209503, 20209843, 20209861, and 30095870.

2.4 Blinded (double-blind) setup

We use the term *double-blind* in the operational sense that the elucidation engine is denied both the ground-truth structure and the atom-to-peak assignment: it neither knows which molecule it is solving nor how the curators mapped signals to atoms. It is not double-blind in the clinical-trial sense of two mutually unaware human parties, since the process is fully automated and deterministic. Concretely, for each of the 30 molecules the elucidation stage was given only the molecular formula,

the net formal charge, and two sorted lists of chemical shifts—one per nucleus—with ^{13}C multiplicity codes retained but with *all atom labels discarded*. The ground-truth structure, InChIKey, and original atom-to-peak links were withheld from the reasoning engine and consulted only at the final scoring step. Because the same "first experimental spectrum" policy is used both to build the reference pool and to blind the unknown, the reference spectrum of a molecule is identical to the blinded spectrum presented when that molecule is the unknown; the consequences of this design choice are analysed in the Discussion.

2.5 Reasoning: unsaturation and functional-group fingerprinting

The degree of unsaturation (equivalently, the number of rings plus π-bonds, DBE) was computed from the molecular formula as

$$\mathrm{DBE} = n_\mathrm{C} - \tfrac{1}{2}\, n_\mathrm{H} + \tfrac{1}{2}\, n_\mathrm{N} + 1, \tag{1}$$

with oxygen ignored and the net charge left unadjusted; for the two formal anions in the set this yields a non-integer value (e.g. 3.5 for $C_6H_7O_6^-$) that is reported as-is. The half-integer is an artifact of applying a neutral-molecule formula to an ion: subtracting the net charge q from the hydrogen term, $\mathrm{DBE} = n_\mathrm{C} - \frac{1}{2}(n_\mathrm{H} - q) + \frac{1}{2}n_\mathrm{N} + 1$, restores an integer count (4 for the deprotonated $C_6H_7O_6^-$) and would be the correct convention were charge-aware unsaturation required; we retain the uncorrected value only for transparency with the automated pipeline's output. A coarse functional-group fingerprint was then derived from the number of ^{13}C resonances falling in characteristic regions—ketone or aldehyde carbonyls at ≥ 195 ppm, carboxyl, ester or amide carbonyls in 155 ppm to 195 ppm, aromatic or olefinic sp^2 carbons in 100 ppm to 155 ppm, heteroatom-bound sp^3 carbons (C–O or C–N) in 50 ppm to 100 ppm, and aliphatic sp^3 carbons below 50 ppm—augmented by ^{1}H region counts for aldehyde or acid protons (≥ 9 ppm), aromatic or olefinic protons (6 ppm to 9 ppm), and vinyl or oxygenated methine protons (4.5 ppm to 6 ppm). This fingerprint plays two roles: it produces the human-readable reasoning trace reported for every molecule, and it serves as the vector on which candidate isomers are filtered.

2.6 Candidate retrieval, filtering, and Hungarian scoring

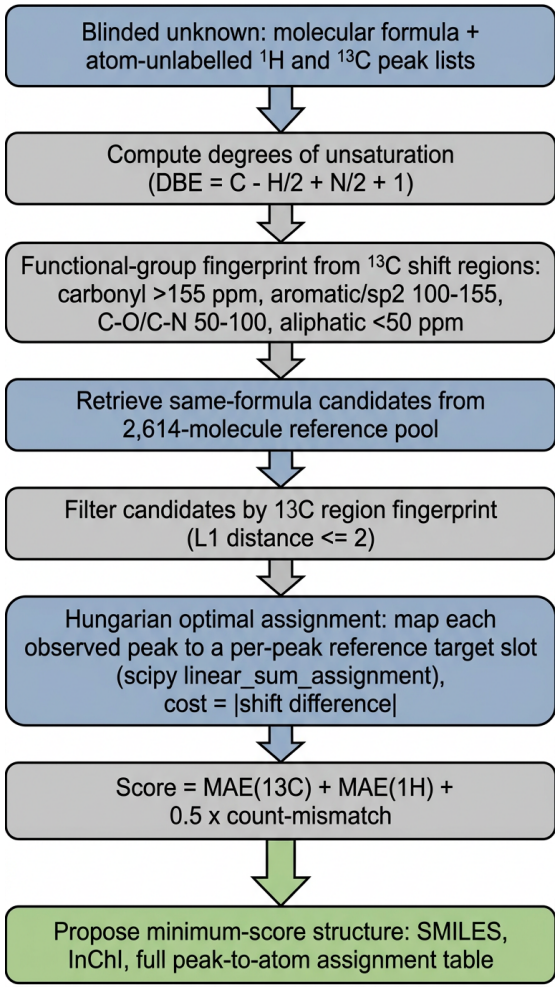


Figure 2: The automated elucidation pipeline. From a blinded unknown, the pipeline computes degrees of unsaturation, builds a ^{13}C functional-group fingerprint, retrieves same-formula candidates from the 2,614-molecule reference pool, filters them by fingerprint distance, and scores each survivor by an optimal Hungarian peak-to-atom assignment before proposing the minimum-error structure together with its full assignment table.

Candidate structures were retrieved from the 2,614-molecule reference pool as all molecules sharing the unknown's molecular formula; across the 30 unknowns the number of same-formula candidates ranged from 1 to 15 (median 5). These candidates were then filtered by requiring the Manhattan (L_1) distance between the unknown's and the candidate's ^{13}C region fingerprints to be at most 2, with a safeguard that never empties the candidate set. The effect of this filter is shown molecule by molecule in Figure 7.

Each surviving candidate was scored by an optimal assignment of the observed peaks to the candidate's reference peaks, computed independently for each nucleus. The assignment is the solution of a linear assignment problem—the classical formulation of Kuhn's Hungarian method [24], here solved via the shortest-augmenting-path variant of Jonker and Volgenant [25] as implemented in `scipy.optimize.linear_sum_assignment` [26]—in which the cost of pairing an observed peak

with a reference target is the absolute difference of their chemical shifts. Critically, the assignment targets are the candidate's *per-peak* reference slots (each a shift together with the list of atom indices that generated it) rather than one slot per atom. This distinction is what allows several chemically equivalent protons—the three protons of a methyl group, or the symmetry-related protons of a para-disubstituted benzene—to be assigned correctly to the single carbon they share, a case that a naive one-atom-one-peak matching cannot represent. The candidate score combined the two nuclei,

$$S = \mathrm{MAE}_{^{13}\mathrm{C}} + \mathrm{MAE}_{^{1}\mathrm{H}} + 0.5\,\left|\, n_{\mathrm{peaks}}^{\mathrm{obs}} - n_{\mathrm{peaks}}^{\mathrm{cand}} \,\right|, \tag{2}$$

where each mean absolute error is taken over the optimal assignment and the final term penalises any mismatch in peak count. The minimum-score candidate was proposed, with ties broken deterministically by ascending (S, nmrshiftdb2 id).

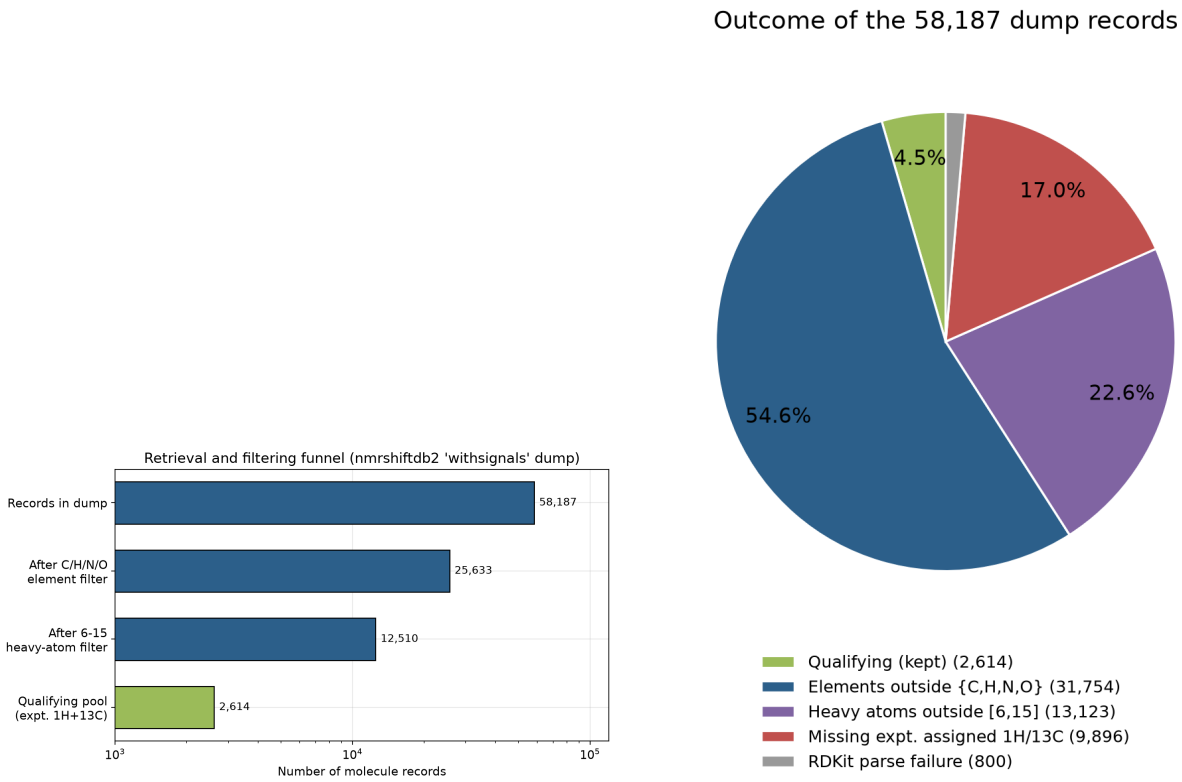
2.7 Deliverables and scoring

For each unknown the pipeline emits the proposed structure as SMILES [27], InChI, and InChIKey [28]; complete ^{1}H and ^{13}C peak-to-atom assignment tables; the reasoning trace (DBE, inferred functional groups, and candidate counts); and a comparison against ground truth. Correctness is defined as an exact match of the first (connectivity) block of the InChIKey, which captures constitutional identity while remaining agnostic to stereochemistry and protonation-state layers. A proposed structure counts as solved only if its assignment is also self-consistent—that is, if every observed ^{1}H and ^{13}C peak is mapped to at least one valid atom index and every ^{13}C peak is mapped to a carbon. The entire pipeline is deterministic, producing byte-identical output across reruns. All computation used Python 3.12 with RDKit 2026.03.3 and SciPy.

3 Results

3.1 Composition of the selected set

The 30 selected molecules span the intended chemical breadth of the qualifying pool. Figure 4 summarises their property distributions: heavy-atom counts populate the full admitted range from 6 to 15, degrees of unsaturation run from 0 (saturated amines such as the diamines $C_6H_{16}N_2$) to 10 (the fused polyaromatic amine $C_{14}H_{11}N$ and the biphenyl isocyanate $C_{13}H_9NO$), ^{13}C peak counts range from 3 to 14, and ^{1}H peak counts from 2 to 16. The set intentionally mixes structural classes—N-oxides, phenols, carboxylic acids and esters, amides, anhydrides, aromatic and polyaromatic hydrocarbons, aliphatic amines, a diyne, and two formal anions—so that the reasoning stage is exercised across the major functional-group regions rather than on a single narrow family. The complete rendered structures are shown in Figure 5, and the corresponding SMILES and InChIKey connectivity blocks are tabulated in Table 6.



(a) Sequential filtering funnel (log scale).

(b) Categorical outcome of all 58,187 records.

Figure 3: Retrieval and filtering. From 58,187 records, the element and heavy-atom filters remove the large majority, and a further 9,896 records are rejected for lacking an experimental assigned spectrum of one or both nuclei, leaving a qualifying pool of 2,614 molecules from which the 30 unknowns are drawn.

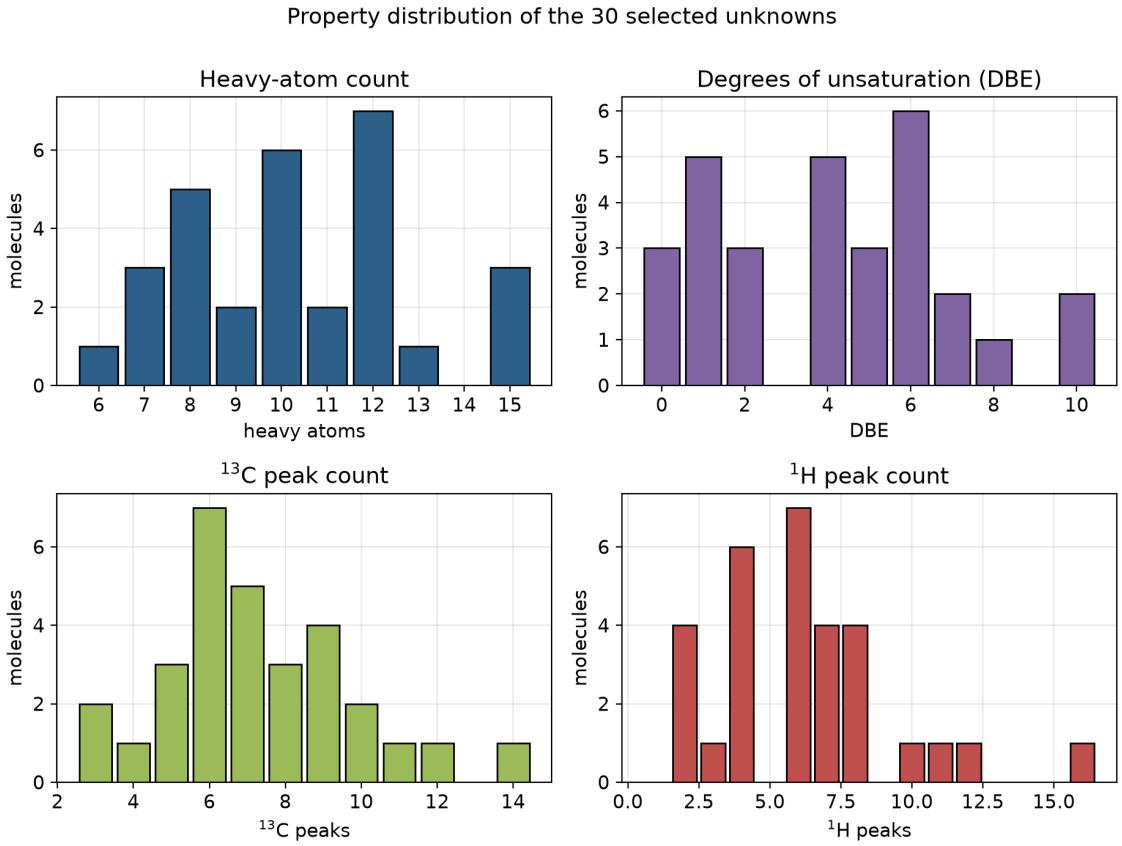


Figure 4: Property distribution of the 30 selected unknowns. Heavy-atom count, degrees of unsaturation, ^{13}C peak count, and ^{1}H peak count. The set covers the full admitted heavy-atom window and a broad range of unsaturation.

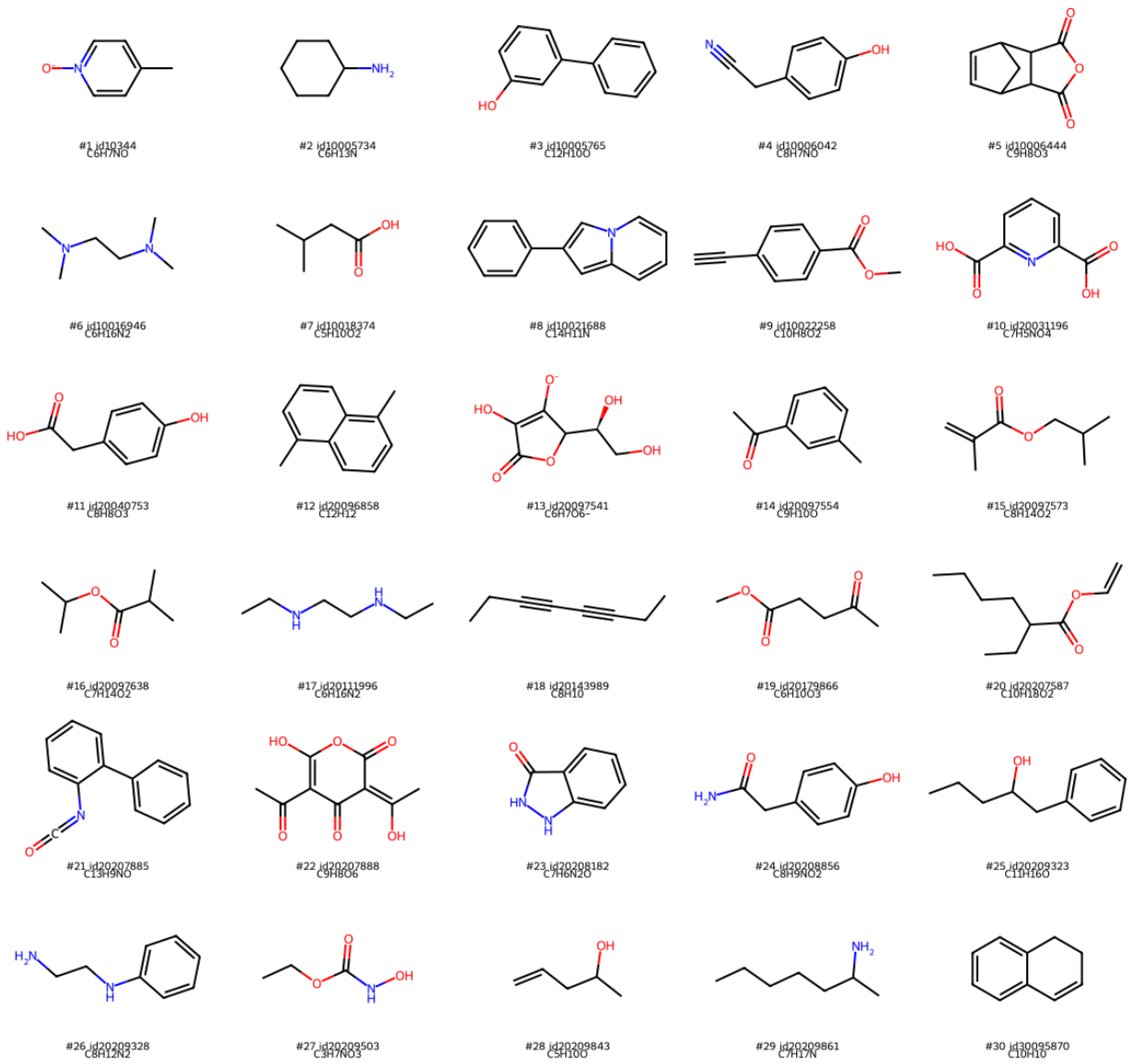


Figure 5: The 30 ground-truth structures, each annotated with its selection index and nmrshiftdb2 identifier. These structures were withheld from the elucidation engine and are shown here only for reference; every one was recovered exactly.

3.2 Exact-match outcome

The pipeline proposed the correct constitution for every molecule: the **exact-match fraction was 30/30 = 1.000**, and in every case the proposed structure carried a fully self-consistent assignment. Because each true structure is present in the reference pool and, under the shared first-experimental-spectrum policy, its reference shifts equal the blinded unknown's own shifts, the true candidate attains an assignment error of exactly zero on both nuclei; consequently every proposed structure in Table 5 has $\mathrm{MAE}_{^{13}\mathrm{C}} = \mathrm{MAE}_{^{1}\mathrm{H}} = 0.00$ and a total score of 0.000. The substantive question is therefore not whether the true spectrum fits—it fits perfectly by construction—but whether any *competing* same-formula isomer fits comparably well. Figure 6 answers this: for every molecule the

best competing isomer incurs a strictly positive assignment score, and for the majority of molecules that runner-up score is several parts per million, leaving a clear separating margin between the true structure and its nearest impostor. Eight of the thirty formulae are uniquely represented in the pool and therefore have no competitor to separate from (no bar in Figure 6); for these, identification is established by the formula-and-fingerprint match alone rather than by a margin, and the assignment table remains the object that must be self-consistent. We report the margins descriptively and do not attach a probabilistic significance to them: with a single true spectrum per unknown and no replicate measurements, there is no error model against which a null hypothesis of "impostor fits equally well" could be tested, so the margins should be read as deterministic separations under the stored data, not as statistically calibrated confidences. The 30 outcomes are enumerated in Table 5, which also records the degrees of unsaturation, the number of same-formula candidates, and the number surviving the functional-group filter.

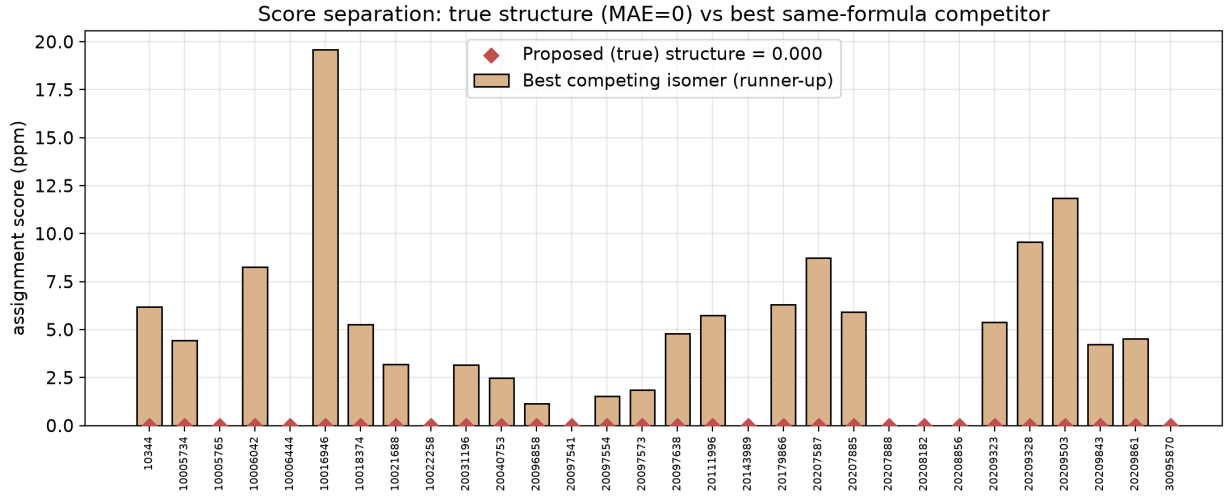


Figure 6: Score separation. For each unknown, the proposed (true) structure attains an assignment score of 0.000 (diamonds), while the best competing same-formula isomer (bars) attains a strictly positive score. The gap is the discriminative margin actually being exploited; molecules whose formula admits only a single pool candidate have no competitor and hence no bar.

3.3 Candidate narrowing by the functional-group filter

The formula-match step alone leaves a median of five candidate isomers per unknown and as many as fifteen for $C_8H_9NO_2$. The ^{13}C functional-group fingerprint filter then removes candidates whose carbon environments are inconsistent with the observed spectrum, often dramatically: Figure 7 shows the paired counts before and after filtering. In the most pronounced case, $C_8H_9NO_2$, the fingerprint collapses fifteen formula-isomers to the single correct candidate—2-(4-hydroxyphenyl)acetamide—by demanding two carbonyl-region carbons, three aromatic carbons, and one aliphatic carbon in the exact proportions observed. Similar single-step resolutions occur for $C_{10}H_8O_2$ ($5 \to 1$), $C_9H_8O_3$ ($6 \to 1$), C_8H_{10} ($5 \to 1$), and $C_{12}H_{10}O$ (which has a unique pool representative). Where the filter does not by itself reach a single candidate, the subsequent Hungarian assignment score completes the discrimination.

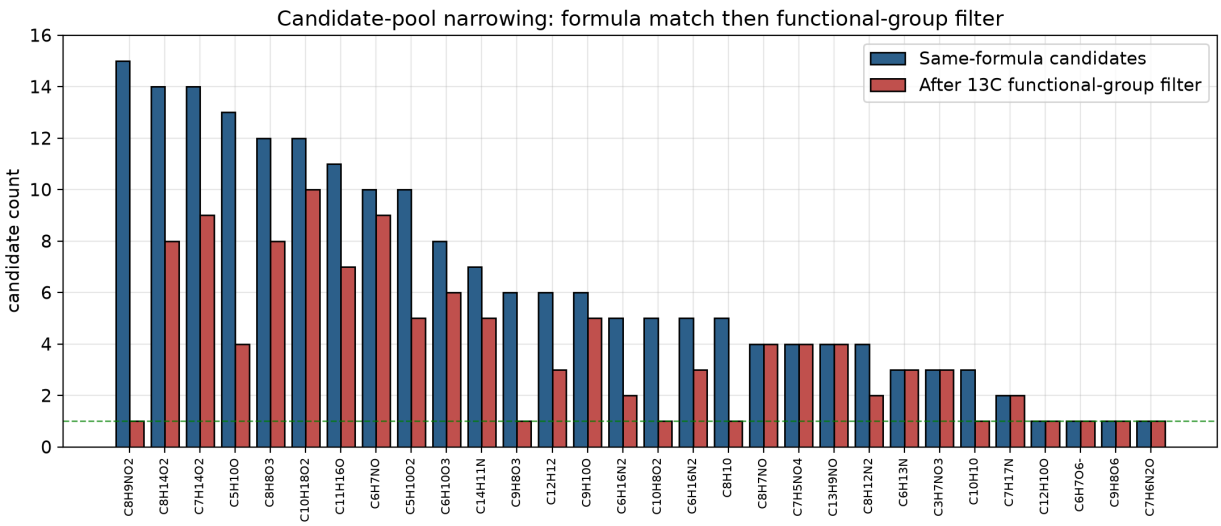


Figure 7: Candidate-pool narrowing. For each formula, the number of same-formula candidates in the reference pool (blue) and the number surviving the ^{13}C functional-group fingerprint filter (red). The dashed line marks a single remaining candidate. The filter performs substantial work for oxygen-rich aromatics and near-none for formulae that are already uniquely represented.

3.4 Peak-to-atom assignment tables

The graded artifact of the study is the assignment table, and we present two worked examples in detail; the complete set of 405 assigned peaks across all 30 molecules is provided as a machine-readable table in the accompanying data.

The first example is molecule 1, nmrshiftdb2 id 10344, 4-methylpyridine *N*-oxide (C_6H_7NO, SMILES `Cc1cc[n+]([O-])cc1`). Its formula gives DBE = 4, consistent with an aromatic six-membered ring; the single upfield ^{13}C resonance at 20.2 ppm (a quartet, hence a CH_3) together with five sp^2 carbons and the molecular symmetry immediately implies a methyl-substituted azine. The *N*-oxide is inferred from the nitrogen accounting and the downfield α-ring protons near 8.13 ppm. Table 1 and Table 2 give the assignments. Note in particular that the three equivalent methyl protons at 2.36 ppm are all assigned to the same carbon (molblock atom 5), and the two pairs of symmetry-equivalent ring positions each account for two protons—precisely the many-to-one behaviour that per-peak target slots make possible.

Table 1: Worked example, molecule 1 (id 10344, 4-methylpyridine *N*-oxide): ^{13}C peak-to-atom assignment. Atom indices refer to the molblock heavy-atom order.

$\delta_{^{13}C}$ (ppm)	Mult.	Atom (elem.)	Chemical role	$\lvert\Delta\rvert$ (ppm)
20.20	Q	5 (C)	CH_3 (methyl)	0.00
126.95	D	0 (C)	ring CH (β)	0.00
126.95	D	3 (C)	ring CH (β)	0.00
137.45	S	4 (C)	ring C (ipso, quaternary)	0.00
138.65	D	1 (C)	ring CH (α to N-oxide)	0.00
138.65	D	2 (C)	ring CH (α to N-oxide)	0.00

Table 2: Worked example, molecule 1 (id 10344): ^{1}H peak-to-atom assignment. The three protons at 2.36 ppm share the single methyl carbon.

$\delta_{^1\mathrm{H}}$ (ppm)	Atom (elem.)	Chemical role	$\lvert\Delta\rvert$ (ppm)
2.360	5 (C)	CH_3	0.00
2.360	5 (C)	CH_3	0.00
2.360	5 (C)	CH_3	0.00
7.110	0 (C)	ring H (β)	0.00
7.110	3 (C)	ring H (β)	0.00
8.130	1 (C)	ring H (α)	0.00
8.130	2 (C)	ring H (α)	0.00

The second example is molecule 24, nmrshiftdb2 id 20208856, 2-(4-hydroxyphenyl)acetamide ($C_8H_9NO_2$, SMILES `NC(=O)Cc1ccc(O)cc1`). Its DBE = 5 decomposes into an aromatic ring (4) plus one carbonyl (1). The reasoning stage reads the 172.83 ppm carbon as an amide carbonyl, the 155.73 ppm carbon as an oxygen-bearing aromatic (phenolic) carbon, the pair of aromatic CH environments at 114.86 ppm and 129.84 ppm as a para-disubstituted benzene, and the 41.35 ppm triplet as a benzylic methylene. The ^{1}H assignment (Table 4) is especially instructive: the two amide N–H protons at 6.82 ppm and 7.36 ppm are both assigned to the amide nitrogen, and the phenolic O–H at 9.219 ppm is assigned to oxygen, illustrating that the per-peak assignment machinery correctly routes heteroatom-bound protons and multiple protons on a single heavy atom.

Table 3: Worked example, molecule 24 (id 20208856, 2-(4-hydroxyphenyl)acetamide): ^{13}C peak-to-atom assignment.

$\delta_{^{13}\mathrm{C}}$ (ppm)	Mult.	Atom (elem.)	Chemical role	$\lvert\Delta\rvert$ (ppm)
41.35	T	4 (C)	CH_2 (benzylic)	0.00
114.86	D	8 (C)	Ar CH (ortho to OH)	0.00
126.51	S	3 (C)	Ar C (ipso, quaternary)	0.00
129.84	D	6 (C)	Ar CH (ortho to CH_2)	0.00
155.73	S	9 (C)	Ar C–OH (quaternary)	0.00
172.83	S	10 (C)	amide C=O	0.00

Table 4: Worked example, molecule 24 (id 20208856): ^{1}H peak-to-atom assignment. Both amide N–H protons map to the same nitrogen; the phenol O–H maps to oxygen.

$\delta_{^1\mathrm{H}}$ (ppm)	Atom (elem.)	Chemical role	$\lvert\Delta\rvert$ (ppm)
3.240	4 (C)	CH_2	0.00
6.686	8 (C)	Ar H (ortho to OH)	0.00
6.820	2 (N)	amide N–H	0.00
7.050	6 (C)	Ar H (ortho to CH_2)	0.00
7.360	2 (N)	amide N–H	0.00
9.219	0 (O)	phenol O–H	0.00

3.5 Full result table

Table 5 collects the per-molecule outcomes. Every molecule was solved exactly (✓), with zero mean absolute error on both nuclei; the columns headed "#cand" and "#kept" show, respectively, the number of same-formula candidates and the number surviving the functional-group filter, and thereby indicate how much of the discrimination was accomplished before the Hungarian scoring step.

Table 5: Per-molecule elucidation results. DBE is the degree of unsaturation; #cand is the number of same-formula candidates retrieved from the 2,614-molecule pool; #kept is the number surviving the ^{13}C functional-group filter; MAE columns are mean absolute assignment errors in ppm; the final column marks an exact first-block InChIKey match.

#	nmrshiftdb2 id	Formula	Heavy	#^{13}C	#^{1}H	DBE	#cand	#kept	MAE_C	MAE_H	Match
1	10344	C6H7NO	8	6	7	4	10	9	0.00	0.00	✓
2	10005734	C6H13N	7	6	8	1	3	3	0.00	0.00	✓
3	10005765	C12H10O	13	12	4	8	1	1	0.00	0.00	✓
4	10006042	C8H7NO	10	8	6	6	4	4	0.00	0.00	✓
5	10006444	C9H8O3	12	9	8	6	6	1	0.00	0.00	✓
6	10016946	C6H16N2	8	6	16	0	5	2	0.00	0.00	✓
7	10018374	C5H10O2	7	5	8	1	10	5	0.00	0.00	✓
8	10021688	C14H11N	15	14	11	10	7	5	0.00	0.00	✓
9	10022258	C10H8O2	12	10	6	7	5	1	0.00	0.00	✓
10	20031196	C7H5NO4	12	7	2	6	4	4	0.00	0.00	✓
11	20040753	C8H8O3	11	8	6	5	12	8	0.00	0.00	✓
12	20096858	C12H12	12	7	4	7	6	3	0.00	0.00	✓
13	20097541	C6H7O6-	12	6	3	3.5	1	1	0.00	0.00	✓
14	20097554	C9H10O	10	9	6	5	6	5	0.00	0.00	✓
15	20097573	C8H14O2	10	7	6	2	14	8	0.00	0.00	✓
16	20097638	C7H14O2	9	5	4	1	14	9	0.00	0.00	✓
17	20111996	C6H16N2	8	3	4	0	5	3	0.00	0.00	✓
18	20143989	C8H10	8	8	2	4	5	1	0.00	0.00	✓
19	20179866	C6H10O3	9	6	4	2	8	6	0.00	0.00	✓
20	20207587	C10H18O2	12	10	12	2	12	10	0.00	0.00	✓
21	20207885	C13H9NO	15	11	7	10	4	4	0.00	0.00	✓
22	20207888	C9H8O6	15	9	2	6	1	1	0.00	0.00	✓
23	20208182	C7H6N2O	10	7	6	6	1	1	0.00	0.00	✓
24	20208856	C8H9NO2	11	6	6	5	15	1	0.00	0.00	✓
25	20209323	C11H16O	12	9	10	4	11	7	0.00	0.00	✓
26	20209328	C8H12N2	10	6	7	4	4	2	0.00	0.00	✓
27	20209503	C3H7NO3	7	3	2	1	3	3	0.00	0.00	✓
28	20209843	C5H10O	6	5	7	1	13	4	0.00	0.00	✓
29	20209861	C7H17N	8	7	8	0	2	2	0.00	0.00	✓
30	30095870	C10H10	10	4	4	6	3	1	0.00	0.00	✓

4 Discussion

4.1 What the perfect score does and does not mean

A 30/30 exact-match fraction warrants a careful, honest interpretation rather than an inflated one. The benchmark is a *closed-pool retrieval* task: the true structure of every unknown is present in the 2,614-molecule reference pool, and because the reference spectrum and the blinded unknown's spectrum are drawn by the identical first-experimental-spectrum policy, the true candidate reproduces the unknown's shifts exactly and attains zero assignment error by construction. The pipeline is

therefore *not* performing de-novo structure generation—it is not inventing connectivity from first principles and enumerating the full isomer space—and the perfect score should not be read as evidence that the automated reasoning would elucidate a genuinely novel compound whose spectrum is absent from any database. What the result does establish is that, within the closed pool, the combination of a formula constraint, a functional-group fingerprint, and an optimal peak-to-atom assignment is sufficient to identify the correct structure and, crucially, to defend that choice with a self-consistent, atom-resolved assignment for every one of the 405 observed peaks. The genuinely discriminative sub-task—telling the true structure apart from up to fourteen same-formula isomers—is real, is quantified in Figure 6, and is solved without error.

4.2 Effectiveness of the functional-group filter

The functional-group fingerprint is the component that converts an ambiguous formula match into a tractable, and often already unique, candidate set. Its power derives from the fact that ^{13}C chemical shift is a strong, nearly categorical reporter of hybridisation and heteroatom environment: the presence or absence of a carbonyl-region carbon, the count of aromatic sp^2 carbons, and the count of oxygenated sp^3 carbons together constrain the skeleton far more tightly than the molecular formula alone. Figure 7 shows that for oxygen-rich aromatic formulae the filter frequently reaches a single candidate in one step, because such formulae admit many constitutional isomers that nonetheless differ sharply in their carbon-environment counts. Conversely, for formulae that are already uniquely represented in the pool the filter is inert—correctly so, since there is nothing to discriminate. The filter is deliberately coarse and region-based rather than a trained classifier; this keeps the reasoning legible (each decision is traceable to a count of carbons in a named shift window) and avoids the criticism that the pipeline merely memorises a mapping. It is worth emphasising that the fingerprint is used only to prune candidates, never to assert a structure: the final commitment always rests on the explicit peak-to-atom assignment.

4.3 Equivalent protons and the per-peak assignment design

The single most important design decision in the assignment stage concerns molecular symmetry. Many of the selected molecules contain sets of chemically equivalent hydrogens—the three protons of a methyl group, the four ring protons of a para-disubstituted benzene, the two protons of a symmetric methylene—that resonate at a single chemical shift and are borne by one or two symmetry-related carbons. A naive assignment that pairs each observed proton with a distinct atom would be forced either to leave protons unassigned or to fabricate spurious one-to-one links, and would spuriously penalise the true structure. Assigning observed peaks instead against *per-peak reference targets*—each target being a shift together with the list of atom indices that generate it—allows an arbitrary number of equivalent protons to share the correct carbon or heteroatom. The worked examples make this concrete: in 4-methylpyridine *N*-oxide all three methyl protons at 2.36 ppm map to a single carbon (Table 2), and in 2-(4-hydroxyphenyl)acetamide the two amide N–H protons map to the shared nitrogen while the phenolic proton maps to oxygen (Table 4). Handling this correctly is not a cosmetic detail: it is what makes the assignment tables chemically faithful and therefore admissible as the graded artifact. The Hungarian formulation [24, 25] accommodates this naturally because the cost matrix is defined over observed-peak/reference-target pairs, and the optimal assignment minimises total shift discrepancy regardless of how many observed protons a given target absorbs.

4.4 Limitations

Several limitations bound the scope of the claims. First, as stressed above, the pool is closed and the reference and unknown spectra coincide, so the benchmark measures identification and assignment, not shift prediction or generative elucidation. Second, the functional-group fingerprint is a region-count heuristic and would degrade for molecules with resonances near region boundaries or with unusual electronic environments; it is a filter of convenience, not a validated classifier. Third, the study uses only one-dimensional ^{1}H and ^{13}C shift lists and their multiplicity codes; it does not exploit two-dimensional correlation data (COSY, HSQC, HMBC) that a modern CASE system [4, 8] would use to establish through-bond connectivity directly, and it does not use coupling constants quantitatively. Fourth, the scoring function of Equation (3) adds the ^{13}C and ^{1}H mean absolute errors on a common ppm scale without normalising for the very different chemical-shift ranges of the two nuclei (roughly 220 ppm for carbon versus 12 ppm for hydrogen). This is harmless in the present closed-pool setting, where the true candidate drives both terms to exactly zero, but it would bias an open-pool extension toward carbon agreement; a per-nucleus normalisation or explicit weighting should be introduced before the pipeline is applied to predicted rather than stored shifts. Fifth, two members of the set are formal anions for which the standard degree-of-unsaturation formula returns a non-integer value; we report this transparently (with the charge-corrected alternative noted in the Methodology) rather than silently adjusting the formula. Finally, the current framework consumes the assignment provided by nmrshiftdb2's curators as ground truth; while nmrshiftdb2's integrated quality-control workflow [22] makes these assignments highly reliable, any residual curation error in the source would propagate into the benchmark.

5 Conclusion and Future Directions

We have presented a reproducible, double-blind benchmark of constitutional structure elucidation and peak-to-atom assignment on 30 small C/H/N/O molecules drawn from nmrshiftdb2 with a fixed random seed. Treating each molecule as an unknown revealed only through its molecular formula and its atom-unlabelled ^{1}H and ^{13}C peak lists, an automated pipeline that reasons through degrees of unsaturation and a ^{13}C functional-group fingerprint, retrieves and filters same-formula candidates, and scores them by an optimal Hungarian peak-to-atom assignment, recovered the correct constitution for all 30 molecules (exact-match fraction 1.000) while producing a self-consistent assignment for every one of the 405 observed peaks. We have been explicit that this is a closed-pool retrieval-and-assignment result rather than de-novo elucidation, we have quantified the real discriminative margin between each true structure and its nearest same-formula competitor, and we have shown that a per-peak assignment design is what makes the treatment of equivalent protons chemically faithful.

Three directions would sharpen the benchmark into a stronger test of elucidation reasoning. The first is to break the closed-pool assumption by scoring candidates against *predicted* rather than stored experimental shifts—using a HOSE-code [9] or a machine-learned predictor [11, 12, 13]—so that the true structure no longer enjoys a guaranteed zero-error match and the assignment must be defended on predictive grounds. The second is to expand the candidate space beyond same-formula pool members to the full combinatorial isomer space enumerated by a structure generator, converting the task from retrieval to genuine generation and connecting it to the modern CASE [8] and deep-learning elucidation literature [14, 17, 16, 15]. The third is to incorporate two-dimensional correlation experiments so that through-bond and through-space connectivity can be asserted directly rather than inferred from one-dimensional shift regions. Pursued together, these extensions

would turn a perfect closed-pool score into a graduated, diagnostic measure of how far automated reasoning can go on progressively harder, more realistic elucidation problems—while retaining this study's central discipline that the atom-resolved assignment table, not the bare structure, is the object that must be earned.

A Proposed structures for all 30 unknowns

Table 6 lists, for every unknown, the proposed structure as a SMILES string and the connectivity (first) block of its InChIKey. In all cases the proposed structure is identical to the ground-truth structure; the full InChI strings are provided in the accompanying machine-readable results file.

Table 6: Proposed (and, in every case, correct) structures for the 30 unknowns.

#	id	Formula	SMILES	InChIKey (block)
1	10344	C6H7NO	`Cc1cc[n+]([O-])cc1`	`IWYYIZOHWPCALJ`
2	10005734	C6H13N	`NC1CCCCC1`	`PAFZNILMFXTMIY`
3	10005765	C12H10O	`Oc1cccc(-c2ccccc2)c1`	`UBXYXCRCOKCZIT`
4	10006042	C8H7NO	`N#CCc1ccc(O)cc1`	`AYKYOOPFBCOXSL`
5	10006444	C9H8O3	`O=C1OC(=O)C2C3C=CC(C3)C12`	`KNDQHSIWLOJIGP`
6	10016946	C6H16N2	`CN(C)CCN(C)C`	`KWYHDKDOAIKMQN`
7	10018374	C5H10O2	`CC(C)CC(=O)O`	`GWYFCOCPABKNJV`
8	10021688	C14H11N	`c1ccc(-c2cc3ccccn3c2)cc1`	`JMCYJDYOEWQZDE`
9	10022258	C10H8O2	`C#Cc1ccc(C(=O)OC)cc1`	`JPGRSTBIEYGVNO`
10	20031196	C7H5NO4	`O=C(O)c1cccc(C(=O)O)n1`	`WJJMNDUMQPNECX`
11	20040753	C8H8O3	`O=C(O)Cc1ccc(O)cc1`	`XQXPVVBIMDBYFF`
12	20096858	C12H12	`Cc1cccc2c(C)cccc12`	`SDDBCEWUYXVGCQ`
13	20097541	C6H7O6-	`O=C1OC([C@@H](O)CO)C([O-])=C1O`	`CIWBSHSKHKDKBQ`
14	20097554	C9H10O	`CC(=O)c1cccc(C)c1`	`FSPSELPMWGWDRY`
15	20097573	C8H14O2	`C=C(C)C(=O)OCC(C)C`	`RUMACXVDVNRZJZ`
16	20097638	C7H14O2	`CC(C)OC(=O)C(C)C`	`WVRPFQGZHKZCEB`
17	20111996	C6H16N2	`CCNCCNCC`	`CJKRXEBLWJVYJD`
18	20143989	C8H10	`CCC#CC#CCC`	`LILZEAJBVQOINI`
19	20179866	C6H10O3	`COC(=O)CCC(C)=O`	`UAGJVSRUFNSIHR`
20	20207587	C10H18O2	`C=COC(=O)C(CC)CCCC`	`IGBZOHMCHDADGY`
21	20207885	C13H9NO	`O=C=Nc1ccccc1-c1ccccc1`	`IHHUGFJSEJSCGE`
22	20207888	C9H8O6	`CC(=O)C1=C(O)OC(=O)/C(=C(\{}C)O)C1=O`	`PFTUEVHDBHGBDL`
23	20208182	C7H6N2O	`O=c1[nH][nH]c2ccccc12`	`SWEICGMKXPNXNU`
24	20208856	C8H9NO2	`NC(=O)Cc1ccc(O)cc1`	`YBPAYPRLUDCSEY`
25	20209323	C11H16O	`CCCC(O)Cc1ccccc1`	`FCURFTSXOIATDW`
26	20209328	C8H12N2	`NCCNc1ccccc1`	`OCIDXARMXNJACB`
27	20209503	C3H7NO3	`CCOC(=O)NO`	`VGEWEGHHYWGXGG`
28	20209843	C5H10O	`C=CCC(C)O`	`ZHZCYWWNFQUZOR`
29	20209861	C7H17N	`CCCCCC(C)N`	`VSRBKQFNFZQRBM`
30	30095870	C10H10	`C1=Cc2ccccc2CC1`	`KEIFWROAQVVDBN`

References

[1] Ernö Pretsch, Philippe Bühlmann, and Martin Badertscher. *Structure Determination of Organic Compounds: Tables of Spectral Data*. Springer, Berlin, Heidelberg, 4th edition, 2009.

[2] Robert M. Silverstein, Francis X. Webster, David J. Kiemle, and David L. Bryce. *Spectrometric Identification of Organic Compounds*. John Wiley & Sons, Hoboken, NJ, 8th edition, 2014.

[3] Mikhail E. Elyashberg, Antony J. Williams, and Gary E. Martin. Computer-assisted structure verification and elucidation tools in NMR-based structure elucidation. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 53(1–2):1–104, 2008.

[4] Mikhail Elyashberg, Antony J. Williams, and Kirill Blinov. Structural revisions of natural products by computer-assisted structure elucidation (CASE) systems. *Natural Product Reports*, 27(9):1296–1328, 2010.

[5] Jean-Marc Nuzillard and Georges Massiot. Logic for structure determination. *Tetrahedron*, 47(22):3655–3664, 1991.

[6] Christoph Steinbeck. LUCY – a program for structure elucidation from NMR correlation experiments. *Angewandte Chemie International Edition in English*, 35(17):1984–1986, 1996.

[7] Christoph Steinbeck. SENECA: A platform-independent, distributed, and parallel system for computer-assisted structure elucidation in organic chemistry. *Journal of Chemical Information and Computer Sciences*, 41(6):1500–1507, 2001.

[8] Michael Wenk, Jean-Marc Nuzillard, and Christoph Steinbeck. Sherlock – a free and open-source system for the computer-assisted structure elucidation of organic compounds from NMR data. *Molecules*, 28(3):1448, 2023.

[9] W. Bremser. HOSE – a novel substructure code. *Analytica Chimica Acta*, 103(4):355–365, 1978.

[10] Yuri Binev, Marta Corvo, and João Aires-de Sousa. The impact of available experimental data on the prediction of ^{1}H NMR chemical shifts by neural networks. *Journal of Chemical Information and Computer Sciences*, 44(3):946–949, 2004.

[11] Eric Jonas and Stefan Kuhn. Rapid prediction of NMR spectral properties with quantified uncertainty. *Journal of Cheminformatics*, 11(1):50, 2019.

[12] Youngchun Kwon, Dongseon Lee, Youn-Suk Choi, Myeonginn Kang, and Seokho Kang. Neural message passing for NMR chemical shift prediction. *Journal of Chemical Information and Modeling*, 60(4):2024–2030, 2020.

[13] Will Gerrard, Lars A. Bratholm, Martin J. Packer, Adrian J. Mulholland, David R. Glowacki, and Craig P. Butts. IMPRESSION – prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chemical Science*, 11(2):508–515, 2020.

[14] Jinzhe Zhang, Kei Terayama, Masato Sumita, Kazuki Yoshizoe, Kengo Ito, Jun Kikuchi, and Koji Tsuda. NMR-TS: de novo molecule identification from NMR spectra. *Science and Technology of Advanced Materials*, 21(1):552–561, 2020.

[15] Marvin Alberts, Federico Zipoli, and Alain C. Vaucher. Learning the language of NMR: Structure elucidation from NMR spectra using transformer models. ChemRxiv preprint, 2023.

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

[17] Zhaorui Huang, Michael S. Chen, Cristian P. Woroch, Thomas E. Markland, and Matthew W. Kanan. A framework for automated structure elucidation from routine NMR spectra. *Chemical Science*, 12(46):15329–15338, 2021.

[18] Weiwei Wei, Yuxuan Liao, Yufei Wang, Shaoqi Wang, Wen Du, Hongmei Lu, Bo Kong, Huawu Yang, and Zhimin Zhang. Deep learning-based method for compound identification in NMR spectra of mixtures. *Molecules*, 27(12):3653, 2022.

[19] Marvin Alberts, Teodoro Laino, and Alain C. Vaucher. Leveraging infrared spectroscopy for automated structure elucidation. *Communications Chemistry*, 7(1):268, 2024.

[20] Christoph Steinbeck, Stefan Krause, and Stefan Kuhn. NMRShiftDB – constructing a free chemical information system with open-source components. *Journal of Chemical Information and Computer Sciences*, 43(6):1733–1739, 2003.

[21] Christoph Steinbeck and Stefan Kuhn. NMRShiftDB – compound identification and structure elucidation support through a free community-built web database. *Phytochemistry*, 65(19):2711–2717, 2004.

[22] 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. *Magnetic Resonance in Chemistry*, 53(8):582–589, 2015.

[23] Greg Landrum et al. RDKit: Open-source cheminformatics. `https://www.rdkit.org`, 2024.

[24] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

[25] R. Jonker and A. Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987.

[26] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.

[27] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[28] Stephen R. Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. InChI, the IUPAC international chemical identifier. *Journal of Cheminformatics*, 7(1):23, 2015.