NMR-CRAFT: STRUCTURE ELUCIDATION THROUGH BIDIRECTIONAL MODELING AND FRAGMENT RECOMBINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Nuclear magnetic resonance (NMR) spectroscopy is fundamental to molecular structure determination, yet recovering complete structures from experimental spectra remains challenging, especially for unseen molecular scaffolds. Motivated by recent advances in large language models (LLMs), we first show that adapting pretrained LLMs to NMR data yields strong structure elucidation performance. A closer analysis of prediction errors reveals two complementary patterns: the correct structure might be generated but not ranked first, while incorrect predictions also preserve substantial parts of the correct structure. These observations suggest that better structure elucidation requires not only stronger generation, but also the ability to reuse partial hypotheses and verify candidate structures. We therefore introduce NMR-LM, a unified language model trained in both directions: spectrum-to-structure prediction generates structural hypotheses, while structure-to-spectrum modeling evaluates their consistency with the observed spectrum. Building on these capabilities, we develop NMR-CRAFT, an efficient test-time procedure that recombines useful fragments from generated and retrieved candidates into new hypotheses and ranks them by spectral likelihood. NMR-LM achieves state-of-the-art structure elucidation performance on NMRGym and NMRSpec, with competitive spectral prediction. NMR-CRAFT further improves Top-1 accuracy by 120% over direct generation on scaffold-split NMRGym, with a speedup exceeding $15\times$ over the reference search. These results highlight our method's potential as a practical tool for NMR structure elucidation.

1 INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy is a central tool for molecular characterization (Silverstein & Bassler, 1962). Two key computational tasks support this process: (a) *structure elucidation* (inverse), which infers molecular structures from spectra, and (b) *spectral prediction* (forward), which predicts spectra from molecular structures. Accurate spectral predictions support structure elucidation by assessing candidate consistency with observed spectra (Guan et al., 2021; Xu et al., 2025). In practice, structure elucidation remains challenging, as it requires recovering global molecular connectivity from local and often ambiguous spectral information (Hu et al., 2024). This challenge is particularly pronounced for unseen molecular scaffolds, as highlighted by evaluations on NMRGym (Fang et al., 2026b).

Recent neural methods have advanced direct spectrum-to-structure prediction, including task-specific Transformers (Hu et al., 2024; Xue et al., 2025; Yang et al., 2026b), pretrained language models (Su et al., 2025), and unified spectroscopic models such as SpecMol (Li et al., 2026). Motivated by the emerging success of pretrained language models, we first examine how far a general-purpose pretrained LLM can go when it is directly finetuned on experimental NMR data. Surprisingly, simple supervised fine-tuning already makes the pretrained LLM competitive with specialized architectures designed specifically for NMR structure elucidation. This strong baseline motivates us to examine its remaining errors more closely.

Our analysis reveals two complementary patterns. First, the correct structure may already appear among the generated candidates but fail to rank first. Second, even incorrect predictions often pre-

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.

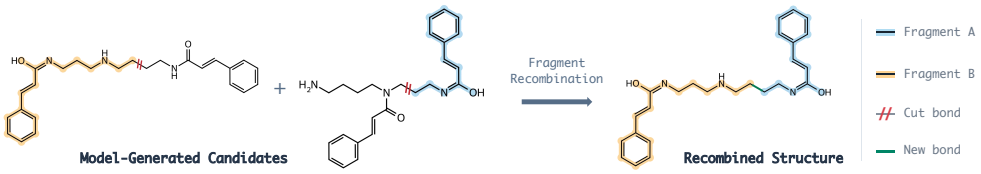

Figure 1: Reusing substructures from molecular hypotheses. Fragments highlighted in different colors are correct substructures extracted from distinct *incorrect* model-generated candidates and recombined to recover the *ground-truth structure* on the right. Marked bonds indicate the cuts and the newly formed connection.

serve substantial parts of the target structure, with different candidates containing complementary substructures (Figure 1). These observations suggest that direct generation leaves useful structural information underexploited, motivating a framework that can both reuse partial hypotheses and verify candidate structures. Related ideas appear in NMR-Solver (Jin et al., 2026) and NMRAgent (Fang et al., 2026a), which use spectral evidence to guide fragment-based refinement. However, recovering unseen scaffolds still remains difficult, and iterative refinement can be costly, leaving room for a more accurate and efficient test-time refinement framework.

These observations suggest that imperfect predictions need not be discarded: useful substructures can be recombined into new candidates, while the reverse structure-to-spectrum mapping can assess their consistency with the observed spectra. Motivated by this cycle-consistency intuition, we develop NMR-LM, a 4B-parameter language model jointly trained for structure elucidation and spectral prediction, and further refined with direct preference optimization (DPO) (Rafailov et al., 2023) to strengthen candidate generation. NMR-LM thus provides both reusable structural hypotheses and a spectral-consistency signal for candidate selection.

Building on NMR-LM, we develop NMR-CRAFT, an efficient test-time refinement framework that combines constrained fragment recombination with spectral reranking. Model generation and database retrieval first form an initial candidate pool, whose fragments are recombined under chemical constraints. The resulting candidates are shortlisted based on source confidence and recurrence, and finally reranked by the likelihood of the observed spectra under NMR-LM.

Our experiments show that NMR-LM achieves state-of-the-art structure elucidation among the evaluated generative models on NMRGym and NMRSpec. On NMRGym, it ranks first in two spectral prediction metrics and second in the other two. DPO further improves candidate generation over supervised fine-tuning while preserving spectral prediction performance. On scaffold-split NMRGym, the full framework achieves approximately $2.2\times$ the formula-conditioned Top-1 accuracy of direct generation, while providing a $15\times$ speedup over NMR-Solver (Jin et al., 2026) under the same protocol. Together, these results demonstrate the strength of the bidirectional model and the value of efficiently reusing and verifying imperfect structural hypotheses.

Our contributions are summarized as follows: (1) We develop NMR-LM, a 4B-parameter language model jointly trained for structure elucidation and spectral prediction, with preference optimization further improving structure generation while retaining spectral assessment capability. (2) We introduce NMR-CRAFT, an efficient test-time refinement framework that reuses imperfect molecular hypotheses through constrained fragment recombination and reranks the resulting candidates using spectral likelihood from the same model. (3) We demonstrate that NMR-LM achieves state-of-the-art structure elucidation performance among the evaluated generative models on NMRGym and NMRSpec, with competitive spectral prediction capability. On scaffold-split NMRGym, NMR-CRAFT reaches approximately $2.2\times$ the formula-conditioned Top-1 accuracy of direct generation, with a speedup exceeding $15\times$ over NMR-Solver under the same evaluation protocol.

2 Related Work

Learning spectrum-structure relationships. Neural NMR structure elucidation includes conditional generation (Alberts et al., 2023; Yao et al., 2023) and Transformer models such as NMR2Struct (Hu et al., 2024), NMRMind (Xue et al., 2025), NMRTrans (Yang et al., 2026b), and image-based NMRViT (Han et al., 2026). DiffNMR (Yang et al., 2026c) and ChefNMR (Xiong et al., 2026) use diffusion to generate molecular graphs and atomic structures, respectively. UltraNMR (Yang et al., 2026a) explores large-scale spectral representation learning, while Spec-

traLLM (Su et al., 2025) predicts structures from multiple spectroscopic modalities. Graph-based and geometry-aware models predict chemical shifts (Jonas & Kuhn, 2019; Kwon et al., 2020), including GT-NMR (Chen et al., 2024) and NMRNet (Xu et al., 2025). SpecMol (Li et al., 2026) unifies bidirectional spectrum–structure generation within a language model, while NMRPeak (Xu et al., 2026) integrates spectral prediction, molecular retrieval, and structure generation. Anthropic (Anthropic, 2026) evaluates general-purpose Claude models for NMR chemical-shift prediction and spectrum-based structure elucidation. NMRGym (Fang et al., 2026b) provides a standardized experimental-NMR benchmark with scaffold-based splits.

Search and verification for structure elucidation. DP4-AI (Howarth et al., 2020) assesses structural hypotheses by matching calculated and experimental NMR data. CReSS (Yang et al., 2021) supports cross-modal retrieval, while spec2struct (Mirza & Jablonka, 2024) performs embedding-guided molecular optimization. SECS (Mirza et al., 2026) combines contrastive spectrum–structure alignment with retrieval-seeded evolutionary search. NMR-Solver (Jin et al., 2026) combines database retrieval, fragment recombination, and staged spectral evaluation using NMRNet. NMRAgent (Fang et al., 2026a) orchestrates retrieval, molecular generation, spectral prediction, and peak–atom verification with formula-aware structural refinement. Building on these directions, NMR-CRAFT integrates constrained fragment recombination into a budgeted search, using the same language model for candidate generation and spectral-likelihood verification.

3 METHOD

3.1 PROBLEM DEFINITION

Let M denote a molecule represented by its canonical SMILES, F its molecular formula, and $S = (S_H, S_C)$ its experimental ^{1}H and ^{13}C NMR spectra. We serialize S into a text sequence denoted by $\tau(S)$. For example, the spectra of acetonitrile ($M =$ `CC#N`, $F = \mathrm{C_2H_3N}$) are serialized as "`<13C_NMR> 116.43 1.89 <1H_NMR> 2.10 2.10 2.10`" for NMRGym, with chemical shifts in ppm. We study two complementary tasks: (a) *structure elucidation* (inverse), which infers molecular structures from spectra; and (b) *spectral prediction* (forward), which predicts spectra from molecular structures. We model both directions with shared parameters θ:

$$p_\theta^{\mathrm{inv}}(M \mid S, c), \qquad p_\theta^{\mathrm{fwd}}(S \mid M), \qquad c \in \{F, \emptyset\}. \tag{1}$$

For the inverse task, we consider *formula-conditioned* ($c = F$) and *formula-free* ($c = \emptyset$) settings. The forward task conditions only on the molecular structure.

3.2 MODEL TRAINING

We jointly fine-tune Qwen3.5-4B (Qwen Team, 2026) on both task directions using task-specific instructions, training a separate model for each dataset. We randomly reserve 5,000 molecules from each dataset's training set $\mathcal{D}_{\mathrm{Full}}$ as a pool $\mathcal{D}_{\mathrm{pool}}$ for preference data construction, and use the remaining data $\mathcal{D}_{\mathrm{SFT}} = \mathcal{D}_{\mathrm{Full}} \setminus \mathcal{D}_{\mathrm{pool}}$ for SFT.

Bidirectional supervised fine-tuning. Using the prompt templates in Appendix B.1, we construct prompt–target pairs $\mathcal{P}_{\mathrm{SFT}}$ from $\mathcal{D}_{\mathrm{SFT}}$ for both directions. Structure elucidation maps serialized spectra and an optional molecular formula to canonical SMILES, while spectral prediction maps canonical SMILES to serialized spectra. We minimize the negative log-likelihood over target tokens:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{P}_{\mathrm{SFT}}}\left[\sum_{t=1}^{|y|} \log p_\theta(y_t \mid x, y_{<t})\right]. \tag{2}$$

These two mappings support candidate generation and spectral assessment at test time.

Offline preference pair construction. For each structure-elucidation prompt x constructed from a molecule in $\mathcal{D}_{\mathrm{pool}}$, we sample $K = 16$ candidates from the SFT model. We retain RDKit-valid candidates with the correct molecular formula, deduplicate them by canonical SMILES, and remove ground-truth matches, yielding a filtered set $\mathcal{C}^-(x)$. We use the ground-truth canonical SMILES as the preferred response and select the candidate with the highest Morgan fingerprint Tanimoto similarity as the rejected response:

$$y^+ = y_{\mathrm{GT}}, \qquad y^- = \mathrm{argmax}_{y\in\mathcal{C}^-(x)} \mathrm{Tanimoto}_{\mathrm{Morgan}}(y, y_{\mathrm{GT}}). \tag{3}$$

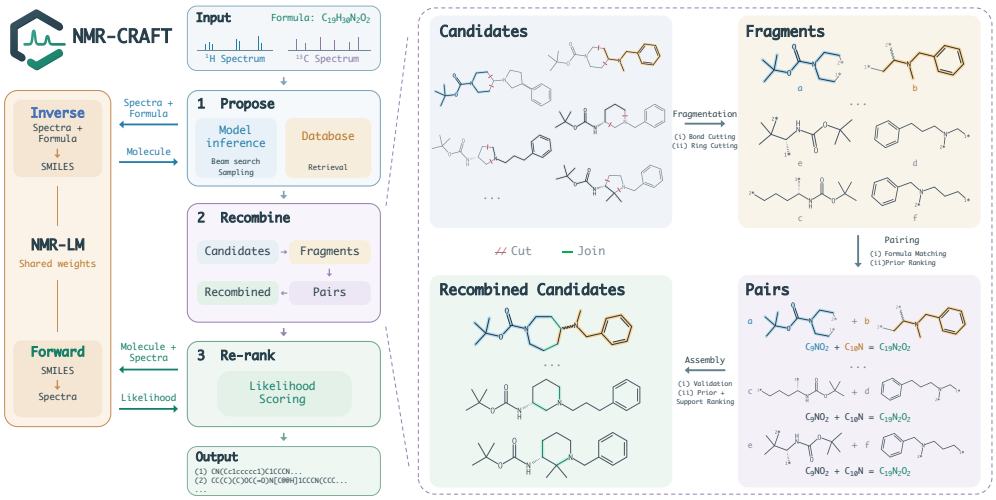

Figure 2: Overview of NMR-CRAFT. (1) The *Propose* stage combines database retrieval with beam search and sampling from NMR-LM. (2) The *Recombine* stage constructs new candidates through fragmentation, constrained pairing, and assembly, followed by validation and preliminary ranking. The expanded panel illustrates these operations; red marks indicate bond cuts and green bonds indicate new connections. (3) The *Re-rank* stage uses spectral likelihood scores from the same NMR-LM model to refine the candidate ranking.

where y_{GT} denotes the canonical SMILES of the ground-truth molecule for prompt x. We discard prompts with no eligible negative and collect the resulting triples (x, y^+, y^-) into $\mathcal{D}_{\mathrm{DPO}}$.

Preference post-training. We initialize the policy from the SFT model and apply DPO (Rafailov et al., 2023) to the preference triples $(x, y^+, y^-) \in \mathcal{D}_{\mathrm{DPO}}$, using a frozen copy of the SFT model as the reference policy p_{ref}. For each triple, we define the relative preference score

$$\Delta_\theta = \log \frac{p_\theta(y^+ \mid x)}{p_{\mathrm{ref}}(y^+ \mid x)} - \log \frac{p_\theta(y^- \mid x)}{p_{\mathrm{ref}}(y^- \mid x)}. \tag{4}$$

We optimize the DPO objective with an auxiliary SFT loss on preferred responses:

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\mathbb{E}_{\mathcal{D}_{\mathrm{DPO}}}\left[\log \sigma(\beta \Delta_\theta)\right] + \lambda_{\mathrm{SFT}} \mathcal{L}^+_{\mathrm{SFT}}(\theta), \tag{5}$$

where the expectation is over preference triples, σ is the sigmoid function, and β controls the preference scale. The preference score uses full-sequence log-probabilities, while $\mathcal{L}^+_{\mathrm{SFT}}$ is the target-token SFT loss on preferred responses, weighted by λ_{SFT}. SFT and DPO hyperparameters are provided in Appendices B.2 and B.3, respectively.

3.3 Test-Time Search

As illustrated in Figure 2, NMR-CRAFT takes experimental spectra S and a molecular formula F as input and proceeds through three stages: candidate proposal, constrained fragment recombination, and spectral reranking. NMR-LM supports both candidate generation and spectrum-based assessment. We use the spectral database and retrieval procedure of NMR-Solver (Jin et al., 2026) and draw inspiration from its fragment recombination strategy.

(1) Propose. We generate molecular candidates through beam search and stochastic sampling from $p^{\mathrm{inv}}_\theta(M \mid S, F)$. These proposals are supplemented by spectral retrieval from SimNMR-PubChem, a database of approximately 106 million molecules. We form the combined candidate pool as

$$\mathcal{P} = \mathrm{Unique}(\mathcal{P}_{\mathrm{beam}} \cup \mathcal{P}_{\mathrm{sample}} \cup \mathcal{P}_{\mathrm{db}}), \tag{6}$$

where Unique canonicalizes and deduplicates molecules. We retain up to 32, 32, and 50 candidates from beam search ($\mathcal{P}_{\mathrm{beam}}$), stochastic sampling ($\mathcal{P}_{\mathrm{sample}}$), and database retrieval ($\mathcal{P}_{\mathrm{db}}$), respectively. Sampled candidates are ranked in the beam-and-sample pool by descending mean log-likelihood under p^{inv}_θ. We assign each candidate $P \in \mathcal{P}$ a score $\rho(P)$ based on its source memberships and ranks for subsequent prioritization and fragment recombination. Details are in Appendix B.5.

(2) Recombine. We first decompose each candidate in $\mathcal{P}$ through non-ring bond cutting and ring opening, retaining attachment sites. For each fragment A, its *parent set* $\mathcal{D}(A)$ contains all candidates in $\mathcal{P}$ for which A appears as a fragment. Let $\mathbf{h}(F)$ denote the heavy-element count vector of the target formula F, and define $\mathbf{h}(A)$ analogously for fragment A, excluding attachment dummy atoms. A pair (A, B) is admissible if the following conditions hold for at least one parent assignment $\pi(A) \in \mathcal{D}(A)$ and $\pi(B) \in \mathcal{D}(B)$:

$$\mathbf{h}(A) + \mathbf{h}(B) = \mathbf{h}(F), \qquad \mathrm{compatible}(A, B) = 1, \qquad \pi(A) \neq \pi(B), \tag{7}$$

where $\mathrm{compatible}(A, B)$ indicates that A and B have compatible attachment types and can be connected through their cut sites. We score each admissible pair by maximizing $\rho(\pi(A)) + \rho(\pi(B))$ over valid assignments to distinct parents (Appendix B.5). We retain up to B_{pair} highest-scoring pairs for assembly with RDKit. Only products that pass sanitization and structural-validity checks and exactly match F, including hydrogen counts, are retained. Canonicalization and deduplication yield the recombined candidate pool $\mathcal{C}_{\mathrm{rec}}$.

For each candidate $M \in \mathcal{C}_{\mathrm{rec}}$, we compute two signals from the selected fragment pairs that generate it: (1) *prior*, the maximum of the pair scores defined above; and (2) *support*, the number of distinct pairs. Let $r_{\mathrm{prior}}(M)$ and $r_{\mathrm{support}}(M)$ denote the descending ranks of these two quantities among candidates in $\mathcal{C}_{\mathrm{rec}}$. We combine these ranks using reciprocal-rank fusion:

$$a(M) = \frac{1}{\kappa + r_{\mathrm{prior}}(M)} + \frac{1}{\kappa + r_{\mathrm{support}}(M)}, \tag{8}$$

where ranks start at one and $\kappa > 0$ controls rank smoothing.

The final pool $\mathcal{C}_{\mathrm{cand}}$ contains up to B_{cand} distinct candidates, prioritizing those from $\mathcal{P}$ that exactly match F and then adding recombined candidates in descending order of $a(M)$. This inexpensive preselection yields a compact candidate pool for subsequent spectral-likelihood reranking by NMR-LM. In our experiments, $B_{\mathrm{pair}} = 1{,}000$ and $B_{\mathrm{cand}} = 100$.

(3) Re-rank. For each candidate $M \in \mathcal{C}_{\mathrm{cand}}$, NMR-LM uses its forward model p_θ^{fwd} to score the observed spectral sequence $z = \tau(S)$. The mean conditional log-likelihood is

$$\ell_\theta(M; S) = \frac{1}{|\mathcal{I}_S|} \sum_{t \in \mathcal{I}_S} \log p_\theta^{\mathrm{fwd}}(z_t \mid M, z_{<t}), \tag{9}$$

where $\mathcal{I}_S$ indexes the scored spectral target tokens. This teacher-forced evaluation assesses spectral consistency without generating a spectrum. Final reranking selects the top-k structures using a weighted combination of normalized spectral and preliminary ranking scores (Appendix B.5).

4 Experiments

4.1 Experimental Setup

Datasets and tasks. NMRGym (Fang et al., 2026b) contains approximately 270k molecules and serves as our primary benchmark for both *structure elucidation* and *spectral prediction*. It uses a scaffold-based split, with test scaffolds absent from training. NMRSpec (Yang et al., 2026b) contains approximately 212k molecules and serves as a secondary benchmark for *structure elucidation* only, using a random split. Both benchmarks use an approximately 8:1:1 split into training, validation, and test sets. Models are trained independently on each benchmark's training split.

Retrieval database and protocols. We use SimNMR-PubChem (Jin et al., 2026), containing approximately 106 million PubChem molecules with ^{1}H and ^{13}C spectra predicted by NMRNet (Xu et al., 2025). Following NMR-Solver, we retrieve candidates using approximate nearest-neighbor search over spectral vectors, followed by set-similarity reranking against the observed spectra. To respect NMRGym's scaffold-based split (Fang et al., 2026b), we apply *Exclude-Scaffold* before retrieval using the same Bemis–Murcko scaffold definition, removing all database molecules sharing scaffolds with molecules in the full test set. For randomly split NMRSpec, we apply *Exclude-Molecule*, removing all test-set molecules from the database. Details are provided in Appendix C.1.

Structure elucidation evaluation. We evaluate formula-conditioned and formula-free settings using (1) sequence accuracy (Seq. Acc.), measuring exact matches between predicted and reference

Table 1: Comparison of structure elucidation performance across methods on NMRGym. *Formula Condition* indicates use of the molecular formula. The NMR-CRAFT workflow takes molecular formulas as input. Best and second-best results are marked in **bold** and underline, respectively.

	Top-1 (% ↑)			Top-5 (% ↑)			Top-10 (% ↑)		
Method	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto
NMR-Solver (Search-Only)	6.27	21.75	33.28	13.96	34.78	42.88	17.37	40.13	46.84
+*Formula Condition*	**17.92**	27.64	43.31	**33.97**	41.06	45.70	**36.48**	42.79	46.11
CLAMS	0.00	7.90	0.72	0.00	12.78	2.24	0.00	14.70	3.49
NMRFormer	1.75	22.49	27.87	2.81	29.58	40.94	3.30	32.32	44.85
NMR2Struct	0.24	21.97	35.18	1.05	33.65	43.43	1.85	38.67	47.47
DiffNMR	0.00	17.37	9.05	0.00	25.37	13.33	0.00	28.49	14.67
ChefNMR-S (Finetune)	1.69	11.87	17.17	3.87	23.85	33.81	5.04	28.18	39.42
ChefNMR-L (Finetune)	1.93	13.59	20.29	4.36	25.92	37.62	5.66	30.07	42.91
NMR-LM	2.98	34.44	44.83	6.41	43.77	51.72	7.66	46.89	53.89
+*Formula Condition*	3.75	35.45	44.41	7.75	46.28	52.63	9.51	50.12	55.49
NMR-CRAFT (*Formula Condition*)	8.23	**44.81**	**55.31**	15.14	**54.64**	**60.30**	17.34	**57.92**	**61.83**

Table 2: Comparison of structure elucidation performance across methods on NMRSpec. Best and second-best results are marked in **bold** and underline, respectively.

	Top-1 (% ↑)			Top-5 (% ↑)			Top-10 (% ↑)		
Method	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto
NMR2Struct	3.13	30.54	27.30	4.76	41.18	38.97	5.53	45.69	42.41
NMR-Solver (search-only)	22.35	29.22	34.72	23.81	30.16	35.23	23.83	30.17	35.26
NMRMind	37.33	63.51	72.71	41.96	68.33	77.05	43.33	69.71	78.14
NMRTrans	42.81	71.30	74.62	58.22	81.46	82.52	61.15	84.18	84.41
NMR-LM	47.80	73.20	75.51	63.61	83.59	84.39	66.48	85.77	86.22
NMR-CRAFT	**58.51**	**83.81**	**86.28**	**77.66**	**90.77**	**90.98**	**79.67**	**91.82**	**91.70**

canonical SMILES; (2) token accuracy (Tok. Acc.), measuring position-wise SMILES token agreement; and (3) Tanimoto similarity based on Morgan fingerprints. For $k \in \{1, 5, 10\}$, each metric is maximized independently over the top-k candidates for each molecule and then averaged across the test set. Our full search framework operates in the formula-conditioned setting.

Spectral prediction evaluation. Given molecular structure, we evaluate predicted ^{1}H and ^{13}C spectra separately using two complementary metrics: (1) set similarity (Set Sim.), measuring peak-level agreement through optimal matching while accounting for unequal peak counts; and (2) vector similarity (Vec. Sim.), measuring spectral agreement through cosine similarity between Gaussian-broadened vectors. Scores are averaged across test molecules. See Appendix C.2 for details.

Baselines. For structure elucidation, we compare with Transformer-based spectrum-to-SMILES models, including CLAMS (Tan, 2025), NMRFormer (Alberts et al., 2023), NMR2Struct (Hu et al., 2024), and NMRMind (Xue et al., 2025), as well as NMRTrans (Yang et al., 2026b), which uses permutation-invariant set encoding. Diffusion baselines include DiffNMR (Yang et al., 2026c) for molecular graph generation and fine-tuned ChefNMR (Xiong et al., 2026) for atomic structure generation. We also compare with NMR-Solver (Jin et al., 2026) on both benchmarks in its search-only configuration, which retrieves and ranks structures from a spectral database.

For spectral prediction, we compare with the commercial software Mnova (Claridge, 2009) and three geometry-aware neural predictors: CASCADE (Guan et al., 2021), a 3D graph neural network with DFT-trained and experimentally adapted variants; DetaNet (Zou et al., 2023), an equivariant tensor network; and NMRNet (Xu et al., 2025), which uses pretrained 3D molecular representations.

4.2 Main Results

Tables 1 and 2 compare NMR-LM and the NMR-CRAFT framework with existing methods on NMRGym and NMRSpec. Across both benchmarks, NMR-LM outperforms all evaluated generative baselines in sequence accuracy, token accuracy, and Morgan similarity at all three Top-k levels. Its advantage on NMRGym holds without molecular formulas, while its formula-conditioned variant achieves higher sequence accuracy. On NMRSpec, NMR-LM achieves 47.80% Top-1 sequence accuracy versus 42.81% for NMRTrans. Building on these candidates, NMR-CRAFT improves every reported metric on both benchmarks. On scaffold-split NMRGym, it achieves approximately 2.2× the formula-conditioned Top-1 accuracy of direct generation, reaching 8.23%; on NMRSpec, Top-1 accuracy reaches 58.51%, a gain of 10.71 points over direct generation. The framework also improves Top-5 and Top-10 accuracy, along with Morgan similarity, across both benchmarks.

Table 3: Spectral prediction performance on NMRGym. * denotes closed-source commercial software. Best and second-best results are **bolded** and underlined, respectively.

Method	^{1}H NMR (% ↑)		^{13}C NMR (% ↑)	
	Set Sim.	Vec. Sim.	Set Sim.	Vec. Sim.
Mnova*	87.90	84.43	72.68	**89.48**
Cascade *(DFTNN)*	68.99	86.92	44.74	33.03
Cascade *(ExpNN-DFT)*	–	–	14.66	12.32
Cascade *(ExpNN-FF)*	–	–	3.30	5.97
DetaNet	68.71	86.32	77.84	64.17
NMRNet	70.46	87.85	**90.80**	84.92
NMR-LM	**88.14**	**91.21**	90.67	86.36

Table 4: Effect of candidate refinement on NMRGym. Formula and Likelihood denote formula filtering and spectral-likelihood reranking, applied to the same 32 beam candidates, while NMR-LM baseline uses beam width 10.

Method	Seq. Acc. (% ↑)		
	Top-1	Top-5	Top-10
NMR-LM	3.75	7.75	9.51
+ Formula	5.13	9.95	11.37
+ Likelihood	4.84	9.05	10.59
+ Formula + Likelihood	5.36	9.42	10.76
NMR-CRAFT	**8.23**	**15.14**	**17.34**

The results also reveal useful information in imperfect predictions. The gap between Top-1 and Top-10 accuracy shows that correct structures are often generated but not ranked first. On NMRGym, NMR-Solver achieves higher exact-match accuracy in its search-only setting, benefiting from the broad coverage of SimNMR-PubChem, which contains 74.9% of NMRGym and 33.2% of NMRSpec test molecules. Nevertheless, NMR-LM achieves higher token accuracy and Morgan similarity, suggesting that its predictions retain useful structural information even when exact recovery fails. Unlike the search-only NMR-Solver baseline, NMR-CRAFT excludes test scaffolds on NMRGym and test molecules on NMRSpec during retrieval. Together, these findings highlight the value of substructure reuse and spectral verification for more accurate structure elucidation.

For spectral prediction, NMR-LM ranks first on both ^{1}H metrics and second on both ^{13}C metrics (Table 3). In particular, its ^{1}H vector similarity reaches 91.21%, compared with 87.85% for the strongest baseline. These results demonstrate competitive spectral prediction alongside strong structure generation within a single model. This forward capability provides the basis for assessing candidate structures through the conditional likelihood of the observed spectra in NMR-CRAFT.

Candidate filtering and reranking. Motivated by the gap between Top-1 and Top-10 accuracy, we investigate how much verification signals can improve candidate selection. Table 4 evaluates formula filtering and spectral-likelihood reranking of 32 beam candidates. Formula filtering yields 5.13% Top-1 sequence accuracy; adding likelihood reranking raises it to 5.36% but slightly lowers Top-5/Top-10 accuracy. The full NMR-CRAFT framework reaches 8.23% Top-1 accuracy and outperforms all beam-based variants at all three Top-k levels. These results highlight the value of verification signals for candidate selection and the additional gains from broader structural search.

Molecular size and training-set similarity. We further examine whether the overall gains extend to larger molecules and those less similar to the training set. Figure 3(a,b) compares Top-1 sequence accuracy on the 21,298 NMRSpec test molecules, grouped separately by heavy-atom count h and training-set similarity s. We use cutoffs of 20 and 30 for h, and 0.5, 0.7, and 0.9 for s, defined as the maximum Morgan Tanimoto similarity to any training molecule. NMR-LM outperforms NMRTrans except in the highest-similarity group, while NMR-CRAFT achieves the highest accuracy in every group. The framework's gains over NMR-LM increase with molecular size, with accuracy rising from 34.0% to 47.0% for $h \geq 31$. For $s < 0.5$, it improves accuracy from 14.0% to 22.6%. These results show that the benefits extend to larger and structurally less familiar molecules. Full results and details are provided in Appendix D.1.

4.3 Model Training Ablations

Model size and initialization. We examine how model size and pretrained initialization affect structure elucidation. Figure 3(c) compares pretrained Qwen3.5 models of 0.8B, 2B, 4B, and 9B parameters with a randomly initialized 4B model on formula-conditioned NMRGym, using the same bidirectional SFT data. At 4B, pretrained initialization yields 7.00% Top-10 sequence accuracy versus 0.42% with random initialization, highlighting the importance of pretraining. Accuracy improves with model size, but scaling from 4B to 9B adds only 0.06 points to Top-10 accuracy. Under this training setup, these results support our choice of a pretrained 4B backbone, which approaches 9B accuracy with fewer than half as many parameters. Full results, loss curves, and implementation details are provided in Appendix D.2.

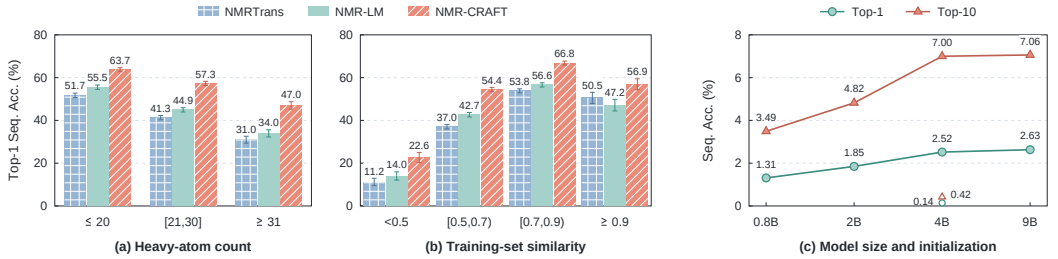

Figure 3: (a,b) Top-1 sequence accuracy on NMRSpec by heavy-atom count and training-set similarity, with 95% bootstrap confidence intervals. (c) Top-1 and Top-10 sequence accuracy across Qwen3.5 model sizes on formula-conditioned NMRGym. Markers denote pretrained (filled) or random (hollow, 4B) initialization.

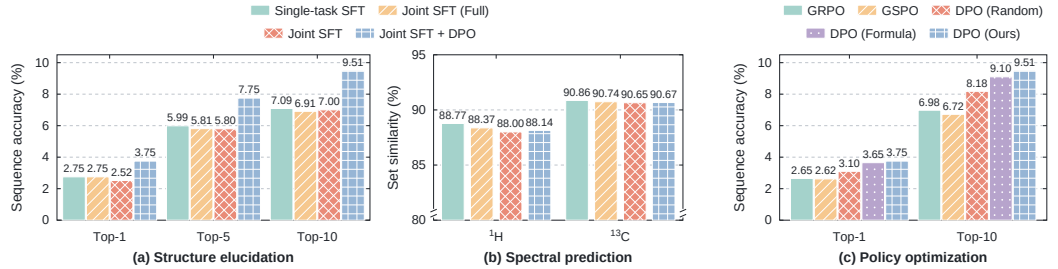

Figure 4: Ablations on NMRGym. (a,b) Effects of SFT configuration and DPO on structure elucidation and spectral prediction. (c) Comparison of policy optimization methods and DPO negative-selection strategies.

Training configuration. We next examine how joint training and preference optimization affect both task directions. Figure 4(a,b) compares four configurations on NMRGym. *Single-task SFT* trains a separate model for each direction, whereas *Joint SFT (Full)* and *Joint SFT* train both directions in one model using $\mathcal{D}_{\text{Full}}$ and $\mathcal{D}_{\text{SFT}}$, respectively. *Joint SFT + DPO* further optimizes the latter model using preference pairs constructed from the 5,000 molecules reserved in $\mathcal{D}_{\text{pool}}$. Joint SFT performs comparably to single-task training in both directions, with little change after reserving the preference pool. DPO increases Top-1 sequence accuracy from 2.52% to 3.75% while maintaining comparable spectral prediction performance. Full results are provided in Appendix D.2.

Policy optimization ablations. Figure 4(c) compares policy optimization methods and DPO negative-selection strategies on formula-conditioned NMRGym. GRPO (Shao et al., 2024) and GSPO (Zheng et al., 2025) yield lower accuracy than DPO in our experiments; implementation details appear in Appendix B.3. With ground-truth positives, we examine two variants: *DPO (Formula)* randomly selects negatives from the formula-matched set $\mathcal{C}^{-}(x)$; *DPO (Random)* additionally removes formula filtering, using the full molecule pool $\mathcal{D}_{\text{pool}}$ and randomly selecting valid, incorrect SFT outputs. *DPO (Ours)* performs best, reaching 9.51% Top-10 sequence accuracy versus 9.10% for DPO (Formula) and 8.18% for DPO (Random). These results support our choice of formula-matched, structurally similar negatives for DPO. See Table 18 for full results.

4.4 Test-Time Search Ablations

Comparison with NMR-Solver. To assess search accuracy and efficiency, we compare NMR-CRAFT with the full NMR-Solver pipeline on 1,000 randomly sampled NMRGym test molecules under formula-conditioned evaluation (Table 5). NMR-Solver is evaluated under unrestricted retrieval, *Exclude-Molecule*, and *Exclude-Scaffold*, with both exclusion protocols defined using the full NMRGym test set. NMR-Solver achieves strong results under unrestricted retrieval, benefiting from SimNMR-PubChem's broad coverage, while accuracy drops sharply when test molecules are excluded. Under the same Exclude-Scaffold protocol, NMR-CRAFT achieves 17.5% Top-10 sequence accuracy versus 5.9% for NMR-Solver, while improving all other reported accuracy and similarity metrics. Its mean inference time is 14.24 seconds per molecule, yielding a speedup of over 15× compared with the fastest evaluated NMR-Solver configuration.

Candidate-source analysis. To understand how candidate sources support structure recovery, we trace recorded recombination routes for 1,345 cases across the full NMRGym test set where the

Table 5: Structure elucidation on 1,000 randomly sampled NMRGym test molecules with molecular formulas provided. Molecule and Scaffold denote the Exclude-Molecule and Exclude-Scaffold retrieval protocols, respectively. Time denotes mean inference time (s/sample).

		Top-1 (% ↑)			Top-5 (% ↑)			Top-10 (% ↑)			
Method	**Exclusion**	Seq. Acc.	Tok. Acc.	Morgan	Seq. Acc.	Tok. Acc.	Morgan	Seq. Acc.	Tok. Acc.	Morgan	Time (↓)
NMR-Solver	-	12.50	41.41	51.40	31.50	51.94	54.76	36.50	54.07	55.91	222.37
NMR-Solver	Molecule	2.90	28.46	36.67	5.80	34.64	39.58	7.20	35.99	40.37	253.58
NMR-Solver	Scaffold	2.90	17.05	20.41	5.20	20.76	22.59	5.90	21.77	23.03	260.34
NMR-CRAFT	Scaffold	8.10	45.10	56.12	14.80	54.97	60.72	17.50	58.33	62.50	14.24

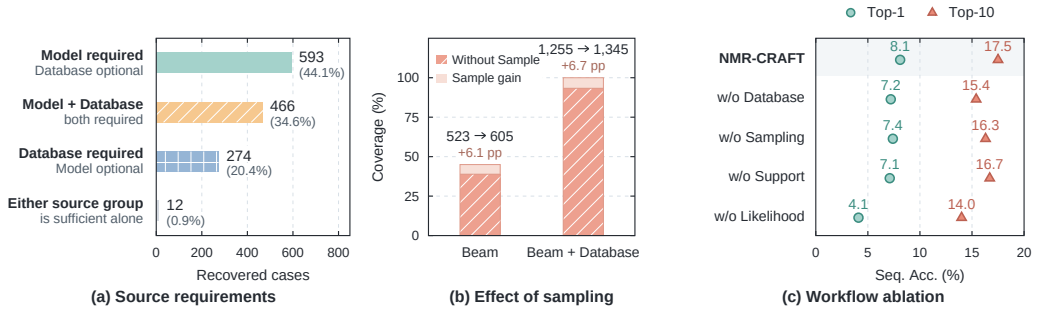

Figure 5: Workflow analysis on NMRGym. (a,b) Source analysis of 1,345 recombination recoveries identified across the full test set: (a) source requirements and (b) coverage without Sample (dark) and additional coverage from Sample (light), with gains in percentage points. (c) Component ablations on a subset of 1,000 formula-conditioned test molecules.

ground truth is absent from Beam Top-32 but recovered in the final Top-10 (Figure 5(a,b)). *Model* combines Beam and Sample, while *Database* denotes retrieved candidates. A source is required if no recorded successful route can be supported without it, considering all donor source memberships. Model-generated donors are required in 78.7% of cases, including 34.6% requiring both Model and Database. Adding Sample to Beam and Database supports successful routes for the remaining 6.7% of cases. These findings support recombining complementary fragments from model-generated and retrieved candidates for structure recovery.

Workflow ablations. To assess the contribution of individual workflow components, we conduct ablations on the 1,000-molecule NMRGym subset used for the NMR-Solver comparison (Table 5). Each variant removes one component: database retrieval, stochastic sampling, the *support* signal in candidate preselection, or spectral likelihood in final reranking (Figure 5(c)). All four ablations reduce Top-1 and Top-10 sequence accuracy, supporting the use of complementary candidate sources and ranking signals. Removing spectral likelihood causes the largest drop, reducing Top-1 accuracy from 8.1% to 4.1%, highlighting the value of the model's forward capability for verifying candidates against the observed spectra and further supporting our bidirectional design. Additional analysis of search-budget sensitivity is provided in Appendix Figure 9.

5 Conclusions

In this paper, we introduce NMR-LM and NMR-CRAFT for NMR structure elucidation. NMR-LM is a 4B-parameter language model jointly trained for structure elucidation and spectral prediction, with preference optimization further strengthening molecular candidate generation. Building on these capabilities, NMR-CRAFT reuses partial structures from generated and retrieved candidates through constrained fragment recombination, and verifies the resulting hypotheses using spectral likelihoods from NMR-LM. Across NMRGym and NMRSpec, NMR-LM achieves state-of-the-art structure elucidation performance among the evaluated generative models, alongside competitive spectral prediction capability. On scaffold-split NMRGym, NMR-CRAFT achieves about $2.2\times$ the formula-conditioned Top-1 accuracy of direct generation and over $15\times$ speedup compared with NMR-Solver under the same protocol. Together, these results show that imperfect molecular predictions can provide useful structural hypotheses that can be efficiently reused and verified for more accurate structure elucidation.

AI USE STATEMENT

We used large language models from the GPT-5.6 and GPT-6 families for language editing and code maintenance. Language assistance included grammar correction, stylistic refinement, and improvements in the clarity and fluency of author-written English text. Code assistance was limited to reviewing exception handling and cleaning up and organizing author-written code in preparation for public release. Generative AI assistance was limited to the uses described above. The research ideas, dataset construction, methodology, experimental design, and interpretation of results were developed and carried out by the authors.

The authors reviewed all AI-assisted text revisions and code changes. We checked that the text revisions preserved the intended meaning and technical content, and verified that the adopted code changes preserved the intended functionality. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops computational methods for NMR structure elucidation and spectral prediction using existing molecular datasets and spectral databases. It does not involve human participants, animal experiments, or personal data. Our model and search framework are intended to support chemical research. Predicted structures should be treated as hypotheses and independently verified before being used to draw chemical conclusions.

REPRODUCIBILITY STATEMENT

We train NMR-LM from Qwen3.5-4B using LLaMA-Factory for SFT and VeRL for DPO. Prompt templates and training and inference configurations are provided in Appendix B. The stages of NMR-CRAFT are described in Section 3.3, with implementation details in Appendix B.5. Dataset splits, retrieval exclusions, and evaluation metrics are documented in Appendix C. We provide the final NMRGym checkpoint[1] and complete structure elucidation outputs[2] from NMR-CRAFT. Our anonymous repository[3] contains code for data processing, model inference, the search workflow, and evaluation.

REFERENCES

Raymond J Abraham and Mehdi Mobli. An nmr, ir and theoretical investigation of 1h chemical shifts and hydrogen bonding in phenols. *Magnetic Resonance in Chemistry*, 45(10):865–877, 2007.

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

Anthropic. Making Claude a chemist. Anthropic Research Blog, June 2026. URL https://www.anthropic.com/research/making-claude-a-chemist. Published June 5, 2026.

Haochen Chen, Tao Liang, Kai Tan, Anan Wu, and Xin Lu. Gt-nmr: a novel graph transformer-based approach for accurate prediction of nmr chemical shifts. *Journal of Cheminformatics*, 16 (1):132, 2024.

Tim Claridge. Software review of mnova: Nmr data processing, analysis, and prediction software. *Journal of Chemical Information and Modeling*, 49(4):1136–1137, 03 2009. ISSN 1549-9596. doi: 10.1021/ci900090d. URL https://doi.org/10.1021/ci900090d.

Zheng Fang, Chen Yang, Yusen Tan, Yunpeng Zhao, Fanjie Xu, Hongxin Xiang, Hanyu Sun, Hanyu Gao, Xiaojian Wang, Wenjie Du, et al. Towards generalizable and evidential nuclear magnetic

[1] https://huggingface.co/Anonymouszzzz/NMR-LM

[2] https://huggingface.co/datasets/Anonymouszzzz/NMR-CRAFT-results

[3] https://anonymous.4open.science/r/NMR-CRAFT-0049

resonance-based molecular structure elucidation via large language model agent. *arXiv preprint arXiv:2606.29776*, 2026a.

Zheng Fang, Chen Yang, Haitao Yu, Haoming Luo, Haitao He, Jiaqing Xie, Zhuo Yang, Yuqiang Li, and Jun Xia. Nmrgym: A comprehensive benchmark for nuclear magnetic resonance based molecular structure elucidation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 8835–8846, 2026b.

Yanfei Guan, S. V. Shree Sowndarya, Liliana C. Gallegos, Peter C. St. John, and Robert S. Paton. Real-time prediction of 1h and 13c chemical shifts with dft accuracy using a 3d graph neural network. *Chemical Science*, 12(36):12012–12026, 09 2021. ISSN 2041-6520. doi: 10.1039/d1sc03343c. URL https://doi.org/10.1039/d1sc03343c.

Chao Han, Xiaolin Pan, and Yingkai Zhang. End-to-end molecular structure elucidation from multimodal nmr spectra images using vision transformers. *Chemical Science*, 17(26):13072–13082, 2026.

Alexander Howarth, Kristaps Ermanis, and Jonathan M Goodman. Dp4-ai automated nmr data analysis: straight from spectrometer to structure. *Chemical science*, 11(17):4351–4359, 2020.

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, 2024.

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

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

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.

Yuqiang Li, Shuaike Shen, Jiaqing Xie, Zhuo Yang, Antong Zhang, Shuzhou Sun, Ben Gao, Tianfan Fu, and Biqing Qi. Specmol: A unified framework for spectroscopy-grounded molecular modeling and evaluation. *arXiv preprint arXiv:2509.21861*, 2026.

Adrian Mirza and Kevin Maik Jablonka. Elucidating structures from spectra using multimodal embeddings and discrete optimization. *ChemRxiv*, 2024(1125), 2024. doi: 10.26434/chemrxiv-2024-f3b18-v2. URL https://chemrxiv.org/doi/abs/10.26434/chemrxiv-2024-f3b18-v2.

Adrian Mirza, Luc Patiny, and Kevin Maik Jablonka. End-to-end multimodal structure elucidation from raw spectra combining contrastive learning and evolutionary algorithms. *Nature Communications*, 17(1):5013, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Robert M Silverstein and G Clayton Bassler. Spectrometric identification of organic compounds. *Journal of Chemical Education*, 39(11):546, 1962.

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. *arXiv preprint arXiv:2508.08441*, 2025.

Xiaofeng Tan. A transformer based generative chemical language ai model for structural elucidation of organic compounds. *Journal of Cheminformatics*, 17(1), 2025. ISSN 1758-2946. doi: 10.1186/s13321-025-01016-1. URL http://dx.doi.org/10.1186/s13321-025-01016-1.

Patrick H Willoughby, Matthew J Jansma, and Thomas R Hoye. A guide to small-molecule structure assignment through computation of (1h and 13c) nmr chemical shifts. *Nature protocols*, 9(3): 643–660, 2014.

Ziyu Xiong, Yichi Zhang, Foyez Alauddin, Chu Xin Cheng, Joon An, Mohammad Seyedsayamdost, and Ellen Zhong. Atomic diffusion models for small molecule structure elucidation from nmr spectra. *Advances in Neural Information Processing Systems*, 38:115995–116031, 2026.

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

Fanjie Xu, Jinyuan Hu, Jingxiang Zou, Junjie Wang, Boying Huang, Zhifeng Gao, Xiaohong Ji, Zhong-Qun Tian, Fujie Tang, Jun Cheng, et al. Nmrpeak: a ready-to-use intelligent system for molecular structure elucidation enabled by synergistic cross-modal learning. *arXiv preprint arXiv:2602.08752*, 2026.

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

Chen Yang, Zheng Fang, Hanyu Sun, Fanjie Xu, Hongxin Xiang, Hanyu Gao, Xiangxiang Zeng, Yuqiang Li, Xiaojian Wang, and Jun Xia. A large-scale foundation model enables simulation-to-real adaptation for nuclear magnetic resonance-based molecular structure analysis. *arXiv preprint arXiv:2606.20756*, 2026a.

Liujia Yang, Zhuo Yang, Jiaqing Xie, Yubin Wang, Ben Gao, Xingjian Wei, Jiaxing Sun, Jiang Wu, Conghui He, Yuqiang Li, et al. Nmrtrans: Structure elucidation from experimental nmr spectra via set transformers. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 12621–12632, 2026b.

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

Zhuo Yang, Jianfei Song, Minjian Yang, Lin Yao, Jiahua Zhang, Hui Shi, Xiangyang Ji, Yafeng Deng, and Xiaojian Wang. Cross-modal retrieval between 13c nmr spectra and structures for compound identification using deep contrastive learning. *Analytical chemistry*, 93(50):16947–16955, 2021.

Lin Yao, Minjian Yang, Jianfei Song, Zhuo Yang, Hanyu Sun, Hui Shi, Xue Liu, Xiangyang Ji, Yafeng Deng, and Xiaojian Wang. Conditional molecular generation net enables automated structure elucidation based on 13c nmr spectra and prior knowledge. *Analytical Chemistry*, 95(12):5393–5401, 03 2023. ISSN 0003-2700. doi: 10.1021/acs.analchem.2c05817. URL https://doi.org/10.1021/acs.analchem.2c05817.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

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

A Limitations and Future Work

While NMR-CRAFT substantially improves formula-conditioned structure elucidation, its current workflow does not yet support formula-free search. Adapting the workflow to jointly explore molecular formulas and structures, and training dedicated models for spectrum-guided structural refinement, are promising directions for this extension. Secondly, while our approach achieves strong results using one-dimensional ^{1}H and ^{13}C NMR, two-dimensional NMR (e.g., HSQC and HMBC), infrared (IR) spectroscopy, and tandem mass spectrometry (MS/MS) provide valuable complementary evidence about molecular connectivity, functional groups, and substructures. Incorporating these modalities into bidirectional modeling and structural search could broaden the framework's applications and support joint reasoning across spectral modalities.

B Implementation Details

B.1 Prompt Templates

We present the chemical-shift serialization $\tau(S)$ and prompt templates used for bidirectional training. For chemical-shift lists, we retain one ^{13}C shift per distinct carbon environment and repeat each ^{1}H shift according to the hydrogen count indicated by its integral. Both lists are sorted by descending chemical shift (ppm). Fields in braces are placeholders replaced by sample-specific values. Structure elucidation targets are canonical SMILES strings.

For NMRGym, the `<FORMULA>` line is included only in the formula-conditioned setting.

Structure elucidation on NMRGym

```
Predict the canonical SMILES from the NMR chemical-shift sets.

<FORMULA> {FORMULA}
<13C_NMR> {13C_SHIFTS}
<1H_NMR> {1H_SHIFTS}

Return only the SMILES.
```

For NMRSpec, the input includes the molecular formula and retains additional attributes of each ^{1}H signal.

Structure elucidation on NMRSpec

```
Predict the canonical SMILES from the NMR data.

<FORMULA> {FORMULA}
<13C_NMR> {13C_SHIFTS}
<1H_NMR> {1H_RICH}

Return only the SMILES.
```

This rich format {1H_RICH} is used only for NMRSpec structure elucidation; spectral prediction and likelihood scoring use the chemical-shift serialization $\tau(S)$ for both datasets. Each entry in {1H_RICH} follows the format below. Entries are separated by semicolons; coupling constants are comma-separated, with `none` indicating no reported coupling.

```
{shift}|{linewidth}|{multiplicity}|{integral}|J={couplings}
```

Both datasets share the following spectral prediction prompt and target format. For NMRSpec, each target ^{1}H shift is repeated according to its integral (e.g., a 6H signal contributes six identical shifts).

Spectral prediction on NMRGym and NMRSpec

Input

```
Predict the NMR chemical-shift sets from the canonical SMILES.
```

```
<SMILES> {SMILES}

Return only the compact block with <13C_NMR> and <1H_NMR>.
```

Target

```
<13C_NMR> {13C_SHIFTS}
<1H_NMR> {1H_SHIFTS}
```

B.2 Supervised Fine-Tuning

We fine-tune Qwen3.5-4B jointly on structure elucidation and spectral prediction, training a separate model on each dataset. Table 6 summarizes the training configuration.

Table 6: SFT hyperparameters for NMR-LM.

Hyperparameter	**Value**
Base model	Qwen3.5-4B
Number of GPUs	8 × A100 80GB
Precision	BF16
Attention implementation	FlashAttention-2
Maximum sequence length	2048
Per-device batch size	2
Gradient accumulation steps	8
Effective batch size	128
Learning rate	1×10^{-5}
Weight decay	0.1
Learning-rate schedule	Cosine decay
Warmup ratio	0.03
Gradient checkpointing	Enabled
Training epochs	10

B.3 Policy Optimization

DPO Data construction We sample 16 candidates per prompt from the SFT checkpoint at step 25,000, using temperature 0.8 and top-$p = 0.95$. On NMRGym, 3,970 of the 5,000 reserved prompts yield eligible preference pairs, while the remaining 1,030 are discarded. The selected negatives have a mean Morgan fingerprint Tanimoto similarity of 0.877 to the ground-truth molecules. The full-data configuration uses all 3,970 pairs for training, whereas the audit configuration uses 3,770 training pairs and 200 validation pairs with no prompt overlap.

DPO training. We apply offline DPO to structure elucidation, using ground-truth canonical SMILES as preferred responses and the frozen SFT model as the reference policy. We use an auxiliary SFT loss on preferred responses, with hyperparameters summarized in Table 7.

GRPO and GSPO. GRPO and GSPO use the same training data, model initialization, rollout configuration, optimization hyperparameters, and reward design. The two methods differ only in the policy-loss formulation. GRPO applies the standard token-level group-relative policy loss, whereas GSPO applies a sequence-level policy loss to the same group-relative advantages. Both methods use 16 stochastic rollouts per prompt with a temperature of 0.8 and top-p value of 0.95.

For both methods, the reward is computed from the generated SMILES and the ground-truth structure. Invalid SMILES receive a reward of -1, while valid candidates are scored using a weighted combination of canonical exact match, Morgan fingerprint Tanimoto similarity, and dense molecular-formula similarity:

$$R = 0.50\, R_{\text{exact}} + 0.35\, R_{\text{Morgan}} + 0.15\, R_{\text{formula}}.$$

Table 7: DPO hyperparameters for NMR-LM.

Hyperparameter	**Value**
Number of GPUs	$8\times$ A100 80GB
Precision	BF16
Attention implementation	SDPA
Distributed strategy	FSDP2
Maximum sequence length	2048
Micro batch size per GPU	1
Effective batch size	8
Gradient accumulation steps	1
Learning rate	2×10^{-7}
DPO β	0.1
DPO loss	Sigmoid
Chosen SFT loss coefficient	0.05
Training epochs	2

Table 8: Shared training hyperparameters for GRPO and GSPO.

Hyperparameter	**Value**
Training data	5,000 prompts
Number of GPUs	$8\times$ A100 80GB
Precision	BF16
Attention implementation	SDPA
Distributed strategy	FSDP
Maximum sequence length	2048
Micro batch size per GPU	1
Train batch size	8
PPO mini-batch size	8
Rollout group size	16
Rollout temperature	0.8
Rollout top-p	0.95
Maximum response length	256
Learning rate	1×10^{-6}
PPO update epochs	1
Training epochs	5
KL loss coefficient	0.005
Clip ratio	0.2 / 0.2

B.4 INFERENCE

We evaluate direct generation using the settings in Table 9. Structure elucidation is evaluated on both NMRGym and NMRSpec at $k \in \{1, 5, 10\}$, while spectral prediction is evaluated only on NMRGym.

Table 9: Direct-generation settings for structure elucidation on both NMRGym and NMRSpec benchmarks and spectral prediction on NMRGym.

Parameter	**Structure elucidation**	**Spectral prediction**
Decoding	Deterministic beam search	Greedy decoding
Beams	10	1
Candidates per molecule	10	1
Max. input tokens	2048	2048
Max. new tokens	512	2048
Precision	BF16	BF16

B.5 TEST-TIME SEARCH

Hardware and inference configuration. We run test-time search on a server with four NVIDIA A100 80GB GPUs. Structure elucidation and spectral prediction are deployed as persistent services on separate GPUs. The structure-elucidation service uses a request batch size of one; beam search with 32 beams and stochastic sampling each generate 32 candidates per input. The spectral-prediction service evaluates likelihoods in batches of 16 candidates, including the final partial batch.

Parent scores. For each canonical parent molecule $P \in \mathcal{P}$, let $\mathcal{S}(P) \subseteq \{\mathrm{beam}, \mathrm{sample}, \mathrm{db}\}$ denote its sources, where db refers to database retrieval. For each $s \in \mathcal{S}(P)$, let $r_s(P)$ denote the rank assigned to P for that source. We define the parent score as

$$\rho(P) = \max_{s \in \mathcal{S}(P)} \frac{w_s}{r_s(P) + \delta}, \tag{10}$$

with $(w_{\mathrm{beam}}, w_{\mathrm{sample}}, w_{\mathrm{db}}) = (1.00, 0.65, 0.35)$ and $\delta = 4$. All ranks start at one. Beam and database ranks follow their respective output orderings. For sampled candidates, $r_{\mathrm{sample}}(P)$ is the rank in the combined beam-and-sample pool, ordered by descending mean inverse log-likelihood $T_P^{-1} \log p_\theta^{\mathrm{inv}}(P \mid S, F)$, where T_P counts SMILES target tokens.

Fragment pairing and candidate preselection. Heavy-element counts are computed over $\{\mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{S}, \mathrm{P}, \mathrm{F}, \mathrm{Cl}, \mathrm{Br}, \mathrm{I}\}$, excluding hydrogen and attachment dummy atoms. Pair lookup enforces Equation 7, with attachment compatibility requiring complementary labels and matching cut arity.

Using the parent sets defined in the main text, we assign each admissible fragment pair (A, B) the score

$$\rho_{\mathrm{pair}}(A, B) = \max_{\substack{\pi(A) \in \mathcal{D}(A),\, \pi(B) \in \mathcal{D}(B) \\ \pi(A) \neq \pi(B)}} \big[\rho(\pi(A)) + \rho(\pi(B))\big]. \tag{11}$$

We select up to $B_{\mathrm{pair}} = 1{,}000$ pairs by descending ρ_{pair} to form $\mathcal{C}_{\mathrm{pair}}$. Assembled products must pass RDKit sanitization and structural-validity checks and exactly match F, including hydrogen counts. Canonicalization and deduplication yield $\mathcal{C}_{\mathrm{rec}}$.

For each $M \in \mathcal{C}_{\mathrm{rec}}$, let $\mathcal{R}(M) \subseteq \mathcal{C}_{\mathrm{pair}}$ be the set of all selected pairs that successfully produce M after these checks, with molecular identity determined by canonical SMILES. Its prior and support are

$$\rho_{\mathrm{rec}}(M) = \max_{(A,B) \in \mathcal{R}(M)} \rho_{\mathrm{pair}}(A, B), \qquad u(M) = |\mathcal{R}(M)|. \tag{12}$$

Support counts each successful selected pair once for each canonical product and is computed only after assembly. For each $M \in \mathcal{C}_{\mathrm{rec}}$, let $r_{\mathrm{prior}}(M)$ and $r_{\mathrm{support}}(M)$ denote its ranks by descending

prior $\rho_{\mathrm{rec}}(M)$ and support $u(M)$, respectively:

$$r_{\mathrm{prior}}(M) = 1 + \sum_{M' \in \mathcal{C}_{\mathrm{rec}}} \mathbb{I}[\rho_{\mathrm{rec}}(M') > \rho_{\mathrm{rec}}(M)]\,,$$
$$r_{\mathrm{support}}(M) = 1 + \sum_{M' \in \mathcal{C}_{\mathrm{rec}}} \mathbb{I}[u(M') > u(M)]\,, \tag{13}$$

where $\mathbb{I}[\cdot]$ is the indicator function. Ranks start at one, and tied scores receive the same rank. These ranks are fused using Equation 8 with $\kappa = 60$.

The preselected pool $\mathcal{C}_{\mathrm{cand}}$ retains formula-matching parents from $\mathcal{P}$ first, then fills the remaining slots with recombined candidates in descending order of $a(M)$, up to $B_{\mathrm{cand}} = 100$ distinct molecules.

Spectral-likelihood reranking. Let $N = |\mathcal{C}_{\mathrm{cand}}|$ and let $r_{\mathrm{pre}}(M) \in \{1, \ldots, N\}$ denote each candidate's position in the preselected list. We normalize this rank and standardize the mean spectral log-likelihood from Equation 9:

$$q_{\mathrm{pre}}(M) = 1 - \frac{r_{\mathrm{pre}}(M)}{N}, \qquad \widetilde{\ell}_\theta(M; S) = \frac{\ell_\theta(M; S) - \mu_\ell}{\max(\sigma_\ell, 10^{-9})}, \tag{14}$$

where μ_ℓ and σ_ℓ are the mean and population standard deviation of $\ell_\theta(M; S)$ across $\mathcal{C}_{\mathrm{cand}}$ for the current input. The final score is

$$s_{\mathrm{final}}(M) = (1 - \lambda_{\mathrm{rank}}) q_{\mathrm{pre}}(M) + \lambda_{\mathrm{rank}} \widetilde{\ell}_\theta(M; S). \tag{15}$$

We set $\lambda_{\mathrm{rank}} = 0.32$ and return the top-k structures by descending $s_{\mathrm{final}}(M)$, with $k = 10$.

Fragment reuse and spectral reassessment. NMR-Solver (Jin et al., 2026) accelerates fragment-pair screening by inheriting atomic chemical shifts from parent molecules under constraints that preserve local bonding environments, followed by fresh spectral prediction for surviving candidates. Although useful for efficient screening, preserving local connectivity alone does not guarantee the transferability of chemical shifts, which also depend on molecular conformation, stereochemistry, and non-bonded interactions (Guan et al., 2021; Willoughby et al., 2014; Abraham & Mobli, 2007). Changes in these factors after recombination may therefore affect the accuracy of inherited shifts and the resulting candidate selection. NMR-CRAFT instead preselects candidates using parent-derived priors and recombination support, then assesses complete structures through observed-spectrum likelihoods from its bidirectional model. This reuses structural fragments without transferring parent-molecule chemical shifts.

B.6 NMR-Solver Reproduction

Implementation. We reproduce NMR-Solver (Jin et al., 2026) using its official implementation, released SimNMR-PubChem database, and five-fold NMRNet ensemble. We use the same server as for NMR-CRAFT, with the NMRNet ensemble running on a single NVIDIA A100 80GB GPU. We retain the original retrieval, spectral matching, and fragment-based molecular optimization (FB-MO) procedures. Inputs comprise paired experimental ^{1}H and ^{13}C chemical-shift lists from NMRGym, with allowed element types supplied as constraints. Hydrogen multiplicity constraints are disabled, exchangeable hydrogen signals are included, and stereochemical annotations are retained throughout retrieval and optimization. Database exclusions follow the shared protocols in Appendix C.1, with masks applied during FAISS retrieval. Table 10 summarizes the configuration.

Evaluation protocol. Primary results use the complete candidate pool saved after the second FB-MO iteration, or the last available iteration if optimization terminates early. We filter this pool by exact molecular formula while preserving the original spectral matching-score order, then compute Top-k metrics as defined in Section 4.1. Formula filtering is applied only during output evaluation and does not modify the internal FB-MO search.

C Experiment Details.

C.1 Dataset and Retrieval Statistics

Dataset statistics. NMRGym and NMRSpec use scaffold-based and random splits, respectively. Table 11 summarizes the numbers of molecular records in their training, validation, and test sets.

Table 10: Configuration used for NMR-Solver reproduction.

Parameter	Value
Initial retrieved candidates	1,000
Maximum FB-MO iterations	2
Molecular pool size	1,000
Maximum retained fragment pairs	200,000
Generated molecules per iteration	1,000
Fragment retrieval top-K	1,000
Final candidate budget	1,000
^{1}H matching bandwidth	1 ppm
^{13}C matching bandwidth	10 ppm

Table 11: Number of molecular records in each benchmark split.

Split	NMRGym	NMRSpec
Train	215,861	169,863
Validation	27,084	21,279
Test	27,054	21,298
Total	269,999	212,440

Database exclusions. Table 12 reports the number of molecules excluded from SimNMR-PubChem under each retrieval protocol, using the full test set of each benchmark. Our default protocols are *Exclude-Scaffold* for NMRGym and *Exclude-Molecule* for NMRSpec; the NMRGym *Exclude-Molecule* setting is included for additional comparisons.

Table 12: Number of database molecules excluded under each retrieval protocol.

Dataset	Exclusion protocol	Molecules excluded
NMRGym	Exclude-Molecule	20,265
NMRGym	Exclude-Scaffold	1,226,050
NMRSpec	Exclude-Molecule	7,068

C.2 Evaluation Details

Structure elucidation. Sequence accuracy compares RDKit-canonicalized SMILES. Token accuracy uses a fixed SMILES tokenizer, treating bracketed atoms, `Br`, `Cl`, and multi-digit ring labels as individual tokens. Given the predicted token sequence P and reference token sequence G, token accuracy is defined as:

$$\mathrm{TokAcc}(P, G) = \frac{1}{|G|} \sum_{t=1}^{\min(|P|,|G|)} \mathbf{1}[P_t = G_t]. \tag{16}$$

Missing predicted positions count as mismatches; tokens beyond the reference length are ignored. Invalid predictions receive zero token accuracy. Tanimoto similarity is computed using RDKit Morgan fingerprints with radius 2, 2048 bits, and chirality disabled. For direct generation, we use deterministic beam search with 10 beams and return 10 candidates; the full framework uses its final search ranking.

Set similarity. Let $S = \{s_i\}_{i=1}^{n}$ and $\widehat{S} = \{\widehat{s}_j\}_{j=1}^{m}$ denote the reference and predicted peak lists. We define Gaussian peak affinity as

$$a_{ij} = \exp\left(-\frac{(s_i - \widehat{s}_j)^2}{2\sigma_{\mathrm{set}}^2}\right). \tag{17}$$

Hungarian matching maximizes the total affinity, with zero-affinity padding for unmatched peaks. Given the resulting matched pairs $\mathcal{M}^\star$,

$$\mathrm{SetSim}(S, \widehat{S}) = \frac{\sum_{(i,j)\in\mathcal{M}^\star} a_{ij}}{\sqrt{nm}}. \tag{18}$$

This normalization penalizes both missing and excess peaks.

Vector similarity. On a common grid $\{x_r\}_{r=1}^{128}$, each peak list is represented by a Gaussian-broadened vector:

$$v_r(S) = \sum_{i=1}^{n} \exp\left(-\frac{(x_r - s_i)^2}{2\sigma_{\mathrm{vec}}^2}\right), \qquad \mathrm{VecSim}(S, \widehat{S}) = \frac{\mathbf{v}(S)^\top \mathbf{v}(\widehat{S})}{\|\mathbf{v}(S)\|_2 \|\mathbf{v}(\widehat{S})\|_2}. \tag{19}$$

We use 128 equally spaced grid points over $[-1, 15]$ ppm for ^{1}H and $[-10, 230]$ ppm for ^{13}C. The bandwidths $(\sigma_{\mathrm{set}}, \sigma_{\mathrm{vec}})$ are $(1.0, 0.3)$ ppm and $(10.0, 2.0)$ ppm, respectively. Missing or unparsable predictions receive zero similarity for nonempty reference spectra.

C.3 Comparison with LLM-based Models

SpectraLLM (Su et al., 2025) and SpecMol (Li et al., 2026) are closely related to NMR-LM, but direct comparisons are limited by input settings and model availability. SpectraLLM supports individual ^{1}H and ^{13}C spectra, whereas its reported joint NMR setting additionally includes 2D HSQC. It reports neither a paired-1D-only configuration nor formula-conditioned training and evaluation, so matching our input settings would require additional adaptation. SpecMol supports both structure elucidation and spectral prediction, but its public release currently provides evaluation and preprocessing code without model checkpoints. We therefore do not include these models in our quantitative comparisons.

D Supplementary Results

D.1 Main Results

Additional Fingerprint Similarity Results. We supplement the Morgan fingerprint results in Table 1 with Tanimoto similarities computed using topological torsion and atom-pair fingerprints. Table 13 reports both at Top-1, Top-5, and Top-10, providing complementary measures of structural similarity.

Table 13: Comparison of topological torsion and atom-pair fingerprint Tanimoto similarity across methods on NMRGym. *Formula Condition* indicates use of the molecular formula.

	Top-1 (% ↑)		**Top-5 (% ↑)**		**Top-10 (% ↑)**	
Method	Torsion	Atom-Pair	Torsion	Atom-Pair	Torsion	Atom-Pair
NMR-Solver (Search-Only)	34.63	45.65	45.29	55.48	49.50	59.17
+*Formula Condition*	44.00	46.00	46.18	47.71	46.54	48.01
CLAMS	0.22	0.58	0.75	2.10	1.20	3.51
NMRFormer	29.11	33.25	43.52	49.44	47.92	53.98
NMR2Struct	32.79	39.54	42.64	48.26	47.28	52.16
DiffNMR	6.44	20.48	12.22	28.95	14.40	31.02
ChefNMR-S (Finetune)	18.72	26.18	37.36	49.68	43.79	56.40
ChefNMR-L (Finetune)	22.15	30.68	41.57	53.97	47.55	59.72
NMR-LM	47.82	57.17	55.62	64.03	58.04	65.93
+*Formula Condition*	47.79	59.30	57.16	67.57	60.32	69.88
NMR-CRAFT (*Formula Condition*)	58.89	68.91	64.44	73.16	66.14	74.47

Results by molecular size and training-set similarity. We extend Figure 3(a,b) with Top-10 sequence accuracy and Top-1/Top-10 Morgan Tanimoto similarity in Figures 6 and 7. Table 14 reports the complete Top-1/5/10 results for sequence accuracy, token accuracy, and Morgan Tanimoto similarity, together with group sizes. For NMRTrans, these stratified analyses use per-molecule predictions from our reproduction, whereas Table 2 reports the results from the original NMRTrans paper.

Each partition covers the same 21,298 test molecules. NMR-CRAFT achieves the highest score in every group across all reported metrics and values of k.

Heavy-atom count h measures molecular size. Training-set similarity s is the maximum Morgan fingerprint Tanimoto similarity between the ground-truth test molecule and any training molecule; the evaluation Tanimoto metric instead compares a prediction with the ground truth. We use the same intervals and group assignments as in the main text. For each metric, we take the best score among the first k candidates for each molecule and then average over test molecules. Lists shorter than k use all available candidates; empty lists receive zero. All test molecules remain in the denominator.

Error bars in all three figures show pointwise 95% percentile bootstrap confidence intervals from 2,000 resamples of test molecules within each group, using the same resampled indices across methods, metrics, and values of k. Table values are reported on a 0–100 scale, with the best score in each group and column shown in bold.

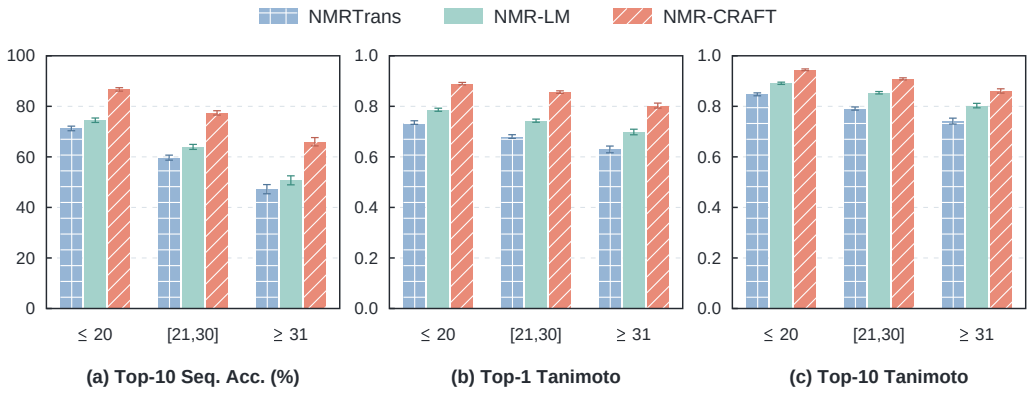

Figure 6: Additional results on NMRSpec by heavy-atom count: (a) Top-10 sequence accuracy, (b) Top-1 and (c) Top-10 Morgan Tanimoto similarity. Error bars show 95% bootstrap confidence intervals.

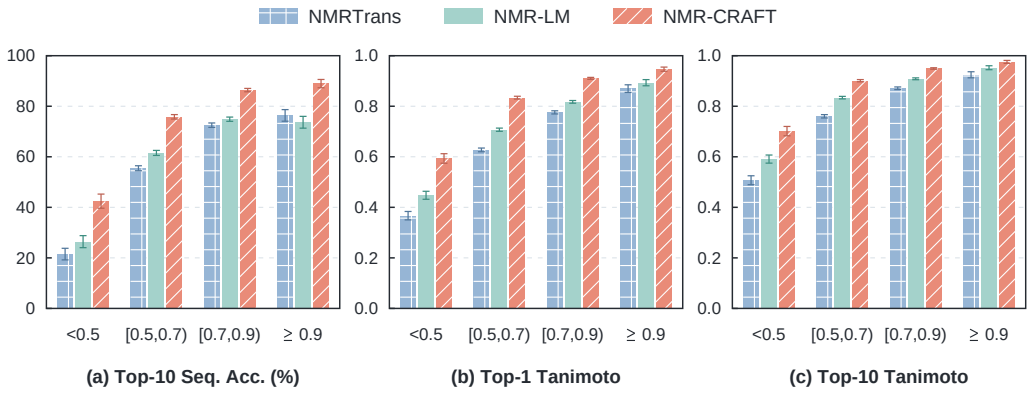

Figure 7: Additional results on NMRSpec by training-set similarity: (a) Top-10 sequence accuracy, (b) Top-1 and (c) Top-10 Morgan Tanimoto similarity. Error bars show 95% bootstrap confidence intervals.

D.2 Ablation Studies

Model size and initialization. Table 15 reports the complete results underlying Figure 3(c) on formula-conditioned NMRGym. All models use the same bidirectional SFT data and are initialized from pretrained Qwen3.5 weights, except for the randomly initialized 4B variant. Figure 8 shows their validation-loss curves over the first 25,000 training steps. The pretrained 4B and 9B models follow similar loss trajectories, while the randomly initialized 4B model exhibits consistently higher

Table 14: Detailed structure elucidation results on NMRSpec, by heavy-atom count or training-set similarity. higher is better. Bold marks the best score within each group and column.

	Top-1 (% ↑)			Top-5 (% ↑)			Top-10 (% ↑)		
Method	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto
Heavy-atom count: $h \leq 20$ *(*$n = 8{,}912$*)*									
NMRTrans	51.66	74.14	73.59	67.95	83.78	82.71	71.25	85.88	84.73
NMR-LM	55.49	78.42	78.65	71.27	88.04	87.21	74.48	90.11	89.12
NMR-CRAFT	**63.70**	**87.36**	**88.97**	**84.38**	**94.13**	**93.76**	**86.68**	**95.05**	**94.45**
Heavy-atom count: $h \in [21, 30]$ *(*$n = 9{,}386$*)*									
NMRTrans	41.30	66.06	68.04	56.72	76.44	77.26	59.66	78.67	79.12
NMR-LM	44.93	71.38	74.32	61.17	82.30	83.54	63.92	84.54	85.37
NMR-CRAFT	**57.28**	**82.97**	**85.63**	**75.55**	**89.91**	**90.14**	**77.38**	**90.99**	**90.87**
Heavy-atom count: $h \geq 31$ *(*$n = 3{,}000$*)*									
NMRTrans	30.97	57.23	62.92	44.43	68.50	72.15	47.20	71.02	74.15
NMR-LM	33.97	63.39	69.88	48.50	74.35	78.68	50.73	76.69	80.30
NMR-CRAFT	**46.97**	**75.88**	**80.32**	**64.30**	**83.54**	**85.35**	**66.00**	**84.83**	**86.12**
Training-set similarity: $s \in [0, 0.5)$ *(*$n = 1{,}198$*)*									
NMRTrans	11.19	39.79	36.74	18.45	52.73	47.60	21.54	56.32	50.81
NMR-LM	14.02	46.30	44.75	22.62	59.40	55.70	26.38	63.40	59.08
NMR-CRAFT	**22.62**	**59.08**	**59.36**	**39.07**	**70.34**	**68.02**	**42.32**	**73.32**	**70.12**
Training-set similarity: $s \in [0.5, 0.7)$ *(*$n = 8{,}926$*)*									
NMRTrans	37.04	62.70	62.73	52.14	73.99	73.60	55.48	76.55	76.04
NMR-LM	42.66	69.67	70.73	58.30	81.06	81.10	61.56	83.60	83.42
NMR-CRAFT	**54.39**	**81.28**	**83.41**	**73.57**	**89.29**	**89.28**	**75.86**	**90.54**	**90.13**
Training-set similarity: $s \in [0.7, 0.9)$ *(*$n = 9{,}866$*)*									
NMRTrans	53.80	74.94	77.58	69.68	84.02	85.61	72.54	85.88	87.14
NMR-LM	56.64	78.54	81.73	72.51	87.78	89.56	74.85	89.50	90.86
NMR-CRAFT	**66.82**	**88.22**	**91.03**	**84.97**	**93.87**	**94.49**	**86.41**	**94.54**	**94.95**
Training-set similarity: $s \in [0.9, 1]$ *(*$n = 1{,}308$*)*									
NMRTrans	50.54	80.87	87.00	73.62	89.48	91.84	76.38	90.78	92.51
NMR-LM	47.17	81.63	89.34	70.26	91.30	94.19	73.70	92.86	95.25
NMR-CRAFT	**56.88**	**90.40**	**94.69**	**85.78**	**96.34**	**97.23**	**89.07**	**96.94**	**97.58**

loss. Although validation loss initially decreases and subsequently increases for all models, final checkpoints are selected based on the corresponding task metrics on the validation set, rather than validation loss.

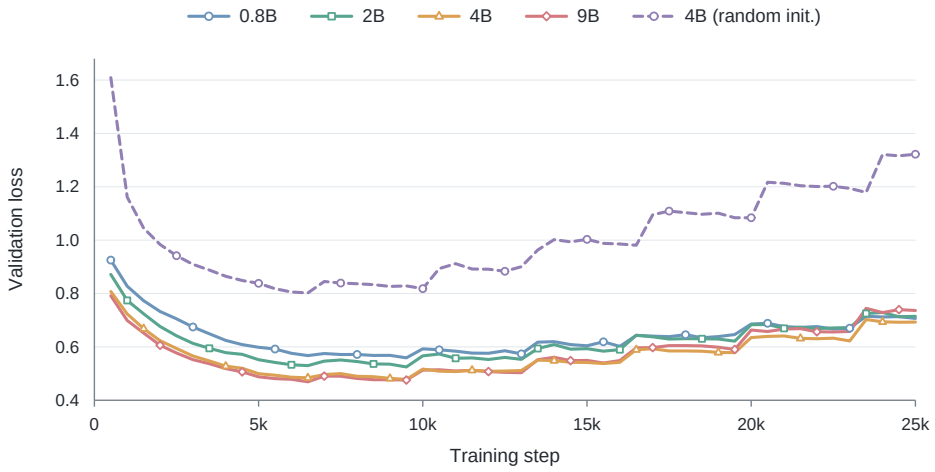

Figure 8: Validation loss during bidirectional SFT on NMRGym. Solid lines denote pretrained Qwen3.5 models; the dashed line denotes the randomly initialized 4B model.

Table 15: Complete results for model size and initialization on formula-conditioned NMRGym. Accuracy is reported in percent; Tanimoto similarity is multiplied by 100.

	Top-1 (% ↑)			**Top-5 (% ↑)**			**Top-10 (% ↑)**		
Model	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto
0.8B	1.31	29.22	35.54	2.86	39.15	43.34	3.49	42.24	45.50
2B	1.85	30.82	37.82	3.96	41.14	46.01	4.82	44.38	48.17
4B	2.52	33.12	41.31	5.80	43.10	48.78	7.00	46.39	50.96
9B	2.63	33.53	41.87	5.84	43.41	49.24	7.06	46.64	51.44
4B (random init.)	0.14	14.71	18.43	0.33	23.52	28.09	0.42	27.92	31.18

Training configuration. Tables 16 and 17 report the complete structure elucidation and spectral prediction results underlying Figure 4(a,b). The three SFT configurations perform comparably in both directions, while DPO improves all reported structure elucidation metrics and maintains comparable spectral prediction performance.

Table 16: Training-configuration ablations for formula-conditioned structure elucidation on NMRGym. Best results are **bold**.

	Top-1 (% ↑)			**Top-5 (% ↑)**			**Top-10 (% ↑)**		
Method	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto
Single-task SFT	2.75	33.36	41.82	5.99	43.51	49.60	7.09	46.68	51.82
Joint SFT (Full)	2.75	32.93	40.89	5.81	43.09	48.83	6.91	46.28	51.00
Joint SFT	2.52	33.12	41.31	5.80	43.10	48.78	7.00	46.39	50.96
Joint SFT + DPO	**3.75**	**35.45**	**44.41**	**7.75**	**46.28**	**52.63**	**9.51**	**50.12**	**55.49**

Policy optimization ablations. Table 18 reports the complete results underlying Figure 4(c) on formula-conditioned NMRGym. *DPO (Ours)* achieves the highest sequence accuracy at all three Top-k levels, supporting our choice of formula-matched, structurally similar negatives.

Effect of DPO on NMRSpec. Table 19 compares NMR-LM before and after DPO on NMRSpec. DPO improves all nine reported metrics, increasing Top-1 sequence accuracy from 45.56% to 47.80%, a gain of 2.24 percentage points, alongside improvements in token accuracy and Tanimoto similarity. These results complement the NMRGym ablation in the main text and show that the benefits of preference post-training extend to NMRSpec.

Table 17: Training-configuration ablations for spectral prediction on NMRGym. Best results are **bold**.

	^{1}H NMR (% ↑)		^{13}C NMR (% ↑)	
Method	Set Sim.	Vec. Sim.	Set Sim.	Vec. Sim.
Single-task SFT	**88.77**	**91.51**	**90.86**	86.42
Joint SFT (Full)	88.37	91.43	90.74	**86.49**
Joint SFT	88.00	91.12	90.65	86.33
Joint SFT + DPO	88.14	91.21	90.67	86.36

Table 18: Policy optimization methods and DPO negative-selection strategies on formula-conditioned NMRGym. Best results are **bold**.

	Top-1 (% ↑)			**Top-5 (% ↑)**			**Top-10 (% ↑)**		
Method	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto
GRPO	2.65	33.06	41.27	5.86	43.00	48.72	6.98	46.26	50.96
GSPO	2.62	33.07	41.29	5.64	42.89	48.62	6.72	46.06	50.75
DPO (Random)	3.10	34.24	42.44	6.79	45.11	50.76	8.18	48.51	53.16
DPO (Formula)	3.65	35.33	44.29	7.64	45.78	52.00	9.10	49.04	54.17
DPO (Ours)	**3.75**	**35.45**	**44.41**	**7.75**	**46.28**	**52.63**	**9.51**	**50.12**	**55.49**

Candidate-source requirements. We analyze 1,345 cases across the full NMRGym test set where the ground truth is absent from Beam Top-32 but recovered through recombination in the workflow Top-10 (Figure 5(a)). Let M denote Model (Beam $\cup$ Sample) and D denote Database. We define three successful route types: $a = MD$, $b = MM$, and $c = DD$, with donor order ignored. All source memberships of each donor are retained, so a single recorded route may support multiple types. For each case, we aggregate the supported types across all recorded successful routes. This yields seven exact patterns: for example, ab means that types a and b are supported, but c is not. Table 20 groups these patterns according to whether Model alone and Database alone can support a successful route. A source group is required if no recorded successful route can be supported without it. Both the seven exact patterns and the four grouped categories are mutually exclusive and exhaustive.

Contribution of stochastic sampling. For the same recovery cases, we assess whether Sample provides successful routes beyond those supported by other available sources (Figure 5(b)). A case is classified as *Sample not required* if at least one recorded successful route is supported without Sample, *Sample required* if a route becomes supported only after adding Sample, and *Unsupported* if no route is supported even with Sample. Table 21 reports these categories with Beam alone or Beam and Database as the other available sources. Adding Sample increases coverage from 523 to 605 cases in the former setting and from 1,255 to 1,345 in the latter. The corresponding gains of 82 and 90 cases are conditional on different source sets and must not be added. Both source analyses use recorded successful routes within the 1,000-pair execution budget; they exclude unexecuted hypothetical routes and do not rerun the workflow after removing sources.

Workflow ablations. Table 22 reports the complete results underlying Figure 5(c). All variants are evaluated on the same subset of 1,000 formula-conditioned NMRGym test molecules. The database and sampling ablations exclude candidates from the corresponding sources. The support ablation omits the support score when preselecting up to $B_{\text{cand}} = 100$ candidates after recombination using up to $B_{\text{pair}} = 1{,}000$prioritized fragment pairs. The likelihood ablation omits spectral likelihood from final reranking. The complete workflow achieves the highest scores across all reported metrics and values of k.

Table 19: Effect of DPO on structure elucidation on NMRSpec.

	Top-1 (% ↑)			**Top-5 (% ↑)**			**Top-10 (% ↑)**		
Method	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto
NMR-LM (w/o DPO)	45.56	72.14	74.02	62.84	83.28	83.78	66.23	85.63	85.77
NMR-LM	47.80	73.20	75.51	63.61	83.59	84.39	66.48	85.77	86.22

Table 20: Source requirements among 1,345 recombination recoveries. Exact-pattern counts are shown alongside their aggregation into four categories. Percentages use all 1,345 cases as the denominator.

Source requirement	**Exact patterns (counts)**	**Group total**
Model required; Database optional	b (459), ab (134)	593 (44.1%)
Model and Database both required	a (466)	466 (34.6%)
Database required; Model optional	c (148), ac (126)	274 (20.4%)
Either source group sufficient alone	bc (0), abc (12)	12 (0.9%)
Total	Seven exact patterns	1,345 (100.0%)

Table 21: Sampling requirements among the same 1,345 recovery cases. Entries report counts and percentages. Categories are mutually exclusive and exhaustive within each row.

Non-sampling sources	**Sample not required**	**Sample required**	**Unsupported**
Beam	523 (38.9%)	82 (6.1%)	740 (55.0%)
Beam + Database	1,255 (93.3%)	90 (6.7%)	0 (0.0%)

Search-budget and reranking sensitivity. We vary five search budgets and the likelihood weight one at a time on the same 1,000 formula-conditioned NMRGym test molecules used in Table 5, under Exclude-Scaffold retrieval. Figure 9(a)–(e) varies the beam, sampling, and retrieval limits, the fragment-pair assembly budget B_{pair}, and the candidate-preselection budget B_{cand}, with defaults 32, 32, 50, 1,000, and 100, respectively. Larger budgets do not uniformly improve accuracy; $B_{\mathrm{cand}} = 200$ yields the highest observed Top-5/Top-10 accuracy among the tested configurations. Panel (f) varies λ_{rank} from 0 (preliminary ranking only) to 1 (spectral likelihood only). Tested intermediate weights outperform either signal alone: the default $\lambda_{\mathrm{rank}} = 0.32$ achieves 8.1% Top-1 accuracy, compared with 4.1% and 4.3% at the two endpoints.

Table 22: Workflow ablations on 1,000 formula-conditioned NMRGym test molecules. bold marks the best result in each column.

	Top-1 (% ↑)			**Top-5 (% ↑)**			**Top-10 (% ↑)**		
Method	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto	Seq. Acc.	Tok. Acc.	Tanimoto
NMR-CRAFT	**8.10**	**45.10**	**56.12**	**14.80**	**54.97**	**60.72**	**17.50**	**58.33**	**62.50**
w/o Database	7.20	43.52	52.85	13.30	51.53	56.98	15.40	54.67	58.41
w/o Sampling	7.40	44.39	54.66	14.20	53.45	59.32	16.30	56.49	60.88
w/o Support	7.10	44.37	55.17	14.40	54.08	59.99	16.70	57.49	61.56
w/o Likelihood	4.10	38.93	47.78	10.80	51.26	55.93	14.00	55.41	58.92

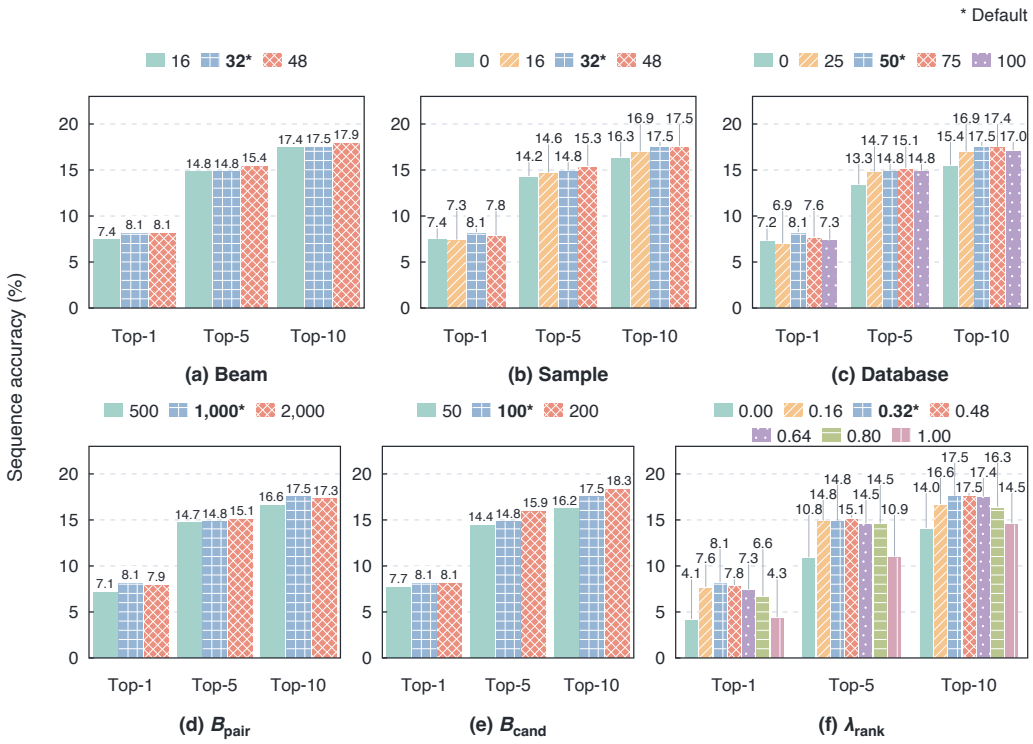

Figure 9: Sensitivity to search budgets (a–e) and likelihood weight (f) on NMRGym. Bars show Top-1/5/10 sequence accuracy; legend values denote the varied setting. Blue bars and starred legend values mark the defaults. Thin lines connect offset value labels to their bars.