TAG-XMM: MULTIMODAL AND MULTI-MODEL JOINT MULTISPECTRAL REPRESENTATIONS FOR MOLECULAR STRUCTURE ELUCIDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular structure elucidation reconciles complementary spectroscopic measurements that each inform on part of a chemical structure: ^{1}H and ^{13}C NMR on proton/carbon environments, HSQC on which proton sits on which carbon, IR on functional groups and MS/MS on fragments. We introduce Token-Aggregated Generation Across Models and Modalities (TAG-XMM), a formula-conditioned framework that jointly encodes these five analytical spectra into a chemically supervised token bank, which a resampler condenses into memory for an autoregressive (AR) decoder whose beam-search candidates form a pool of candidate structures. To increase robustness to cases where the correct structure is absent from the beam, we introduce the Graph-SSI (Graph Spectrum-Steered Inpainting) mechanism. In Graph-SSI, a learned gate estimates the probability that the true structure is missing from the pool. Small regions where the candidates disagree are located, and their atom states and bond orders are regenerated by bond swaps and a discrete graph diffusion model conditioned on the encoded spectra. Finally, a verifier adds to the pool the few graphs that best fit the observed spectra. To select the answer, an inverse solver ranks the resulting augmented pool by the decoder's score for each candidate, as well as by comparing fit of forward-predicted spectra for each candidate to the initial spectra. Our results show that on a large simulated corpus, the decoder reaches 77.88% Exact@1 on 39,224 test molecules. On all 39,224 test molecules, all components provide additive performance benefits and the final pipeline reaches 90.03% Exact@1 accuracy and Graph-SSI corrections push performance to 91.19%, ranking first 519/2,406 structures the beam never proposed; on each baseline's own spectra, TAG-XMM still leads (86.26% vs 80.02% for NMRViT and, without the formula as in NMIRacle, 72.07% vs 55.63%). Code is provided at this link.

1 INTRODUCTION

Molecular structure elucidation from spectra requires reconciling distinct information and constraints from various sources. Infrared (IR) spectroscopy reports on characteristic functional groups; ^{1}H and ^{13}C nuclear magnetic resonance (NMR) focus on the environments of protons and carbons; HSQC assigns protons to the carbons they are bonded to, while tandem mass spectrometry (MS/MS) reports molecular fragment masses. Finally, high-resolution mass spectrometry (HRMS) helps elucidate the molecular structure by evaluating mass-to-charge ratios, providing the molecular formula. Obtaining the molecular formula narrows the search; however, it leaves many constitutional isomeric options with the same formula but different connectivity. As a consequence, all the information in the spectra must be leveraged jointly for accurate reconstruction. Computer-assisted structure elucidation (CASE) proposes candidate structures, predicts their spectra, and ranks them according to the observed spectra (Elyashberg et al., 2008). Similarly, generative models now propose structures *de novo* from one (Alberts et al., 2023; Bohde et al., 2025) or several spectra (Priessner et al., 2026; Alberts et al., 2025), and their generated candidates are subsequently checked and ranked. However, in either pipeline, two failure modes can weaken the system: a correct structure that is never proposed or generated cannot be identified, and a weak ranker can misrank the suggested structures. Both failures can be mitigated: first, by efficiently encoding the spectra in a joint latent space that maxi-

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.

mizes per-spectrum and cross-spectral information; second, by designing a system where successive components each leverage this information efficiently while compensating for each other's weaknesses: repair for structures never proposed, reranking against the observed spectra for misranked ones. Here, we aim to address four questions about the components of TAG-XMM (Fig. 1): Firstly, can a single joint encoder for all five spectra carry information that late fusion of per-spectrum encodings loses? Secondly, can an AR decoder turn the input into exact structures? Thirdly, when the decoder never proposes the truth in its beam, can we also use spectrum-conditioning methods to steer candidates towards the true molecule? And lastly, can checking each candidate's predicted spectra against the initially measured ones rank the truth first when the decoder does not? TAG-XMM addresses these concerns with a jointly trained encoder that conditions a formula-constrained decoder. The decoder then writes a beam of SMILES strings (Weininger, 1988), deduplicated into a candidate pool which is then augmented by the Graph-SSI module. This module selectively adds new molecular out-of-beam graphs generated by bond swaps and spectrum-conditioned graph diffusion to the pool. Each candidate's spectra are then predicted and compared with the observed ones; this fit, with the decoder score, reranks the augmented pool (the inverse solver).

Our contributions are:

- A joint encoder of five spectra that keeps information late fusion loses (97.85% vs 73.41% same-formula retrieval) and an autoregressive decoder using encoded spectras.
- Graph-SSI, which regenerates uncertain regions of the candidate graphs with bond swaps and spectrum-conditioned diffusion, for pools a learned gate flags, and adds the best-fitting graphs without removing any.
- An inverse solver that reranks each molecule's candidates by a weighted sum of the decoder's log-probability and nine spectral views.

2 Related Work

Our work focuses on developing a spectrum-to-structure pipeline to generate candidate molecular structures and revise them. Many models developed in the past aim to retrieve the analyte structure from a single spectrum. Work has been performed on trying to translate NMR (Alberts et al., 2023; Hu et al., 2024) or IR (Alberts et al., 2024a) spectra into the respective SMILES strings, and retrieving (Dührkop et al., 2015) or generating (Litsa et al., 2023; Stravs et al., 2022) the original analyte molecular structures from MS/MS analysis results. Multimodal models like TAG-XMM on the other hand combine several spectra (Han et al., 2026; Ottomano et al., 2025; Alberts et al., 2025). Others fine-tune on simulated analogues of the target (Priessner et al., 2026) or use evolutionary search (Mirza et al., 2026). MoLFormer (Ross et al., 2022) and ChemBERTa (Chithrananda et al., 2020) learn sequence representations; our encoder uses the MoLFormer embedding as its within-formula contrastive target (App. C). Contrastive learning aligns structures with heterogeneous observations (van den Oord et al., 2018; Radford et al., 2021) such as ^{13}C NMR spectra (Yang et al., 2021).

Our work also ranks the generated candidates, by how well each candidate's forward-predicted spectra match the observed ones, together with the decoder score. Candidates are commonly checked in three ways: by substructure codes for shift lookup (Bremser, 1978); by probabilities from computed shifts, used mostly to distinguish stereoisomers (Smith & Goodman, 2010; Grimblat et al., 2015); and by learned forward models of ^{1}H/^{13}C NMR (Guan et al., 2021; Xu et al., 2025), HSQC (Li et al., 2025) and IR (Stienstra et al., 2024). Lemm et al. (2024) match sorted shift vectors, which cannot merge or skip peaks. Where CASE filters candidates by many hard constraints, TAG-XMM's only hard filter at ranking is the parent mass; it then ranks the candidates, matching predicted to observed peaks with an assignment that tolerates merged or missing peaks (Kuhn, 1955; Crouse, 2016).

One of the key ideas we use in this work is the observation that SMILES strings (Weininger, 1988) convert into molecular graphs without loss, thus acting as a bridge between AR-style and graph-style methods. We exploit this property by using both methods successively: the decoder's strings can be revised or edited as graphs. Graph generators build molecules bond by bond (Jonas, 2019), by diffusion over the graph (Bohde et al., 2025) or over 3D coordinates (Xiong et al., 2025), or by completing a retrieved scaffold (Wang et al., 2025). However, instead of generating whole graphs, Graph-SSI keeps the decoder's proposals and, similar to diffusion inpainting (Lugmayr et al., 2022), regenerates only the regions where the decoder's candidates disagree. This is done through discrete

diffusion (Austin et al., 2021) over molecular graphs (Vignac et al., 2023) conditioned on the encoded spectra. Local search over constitutions is an older concept: SENECA anneals toward spectral agreement (Steinbeck, 2001), Faulon (1996) rewires four atoms at a time, graph-editing generators optimize properties (Jensen, 2019; Xie et al., 2021), and, recently, NMR-Solver optimizes fragments (Jin et al., 2026). Our second generator is inspired by the degree-preserving double-edge swap (Maslov & Sneppen, 2002), akin to Faulon's four-atom moves, and pooled with diffusion: fixing every atom's degree in the graph, it cannot reach every isomer (Fosdick et al., 2018), whereas diffusion can also change degrees and hydrogen counts.

3 THE TAG-XMM METHOD

Given spectra $X = \{X_m\}_{m \in \mathcal{M}_{\text{obs}}}$ (^{1}H, ^{13}C, HSQC, IR, MS/MS) and molecular formula F from HRMS, TAG-XMM seeks the exact-connectivity structure $y^\star$, including atom bonding pattern and order, omitting stereochemistry. The token-aggregated model has five components (Fig. 1): (1) an encoder, (2) a resampler and (3) a decoder, which propose a pool of candidate molecules. (4) Graph-SSI adds structures that are potentially missing in the pool of candidates and decides **which structures** to revise, **where** the candidate graphs are uncertain, **how** to regenerate them and **which proposals** to keep. (5) The inverse solver subsequently selects the answer from the resulting augmented beam. Only the decoder, its score and the diffusion model read the encoded spectra; every other score compares a candidate with the observed spectra, so it does not inherit the encoder's errors.

Proposing candidates (1)–(3). A frozen multimodal encoder that is pretrained with chemical objectives (App. C) has one attention branch per spectrum (the architecture accepts any subset; the deployed models were trained on all five) and maps X to a token bank $\tilde{x}$ of 256-dimensional tokens. A Perceiver-style resampler with 64 learned queries (Jaegle et al., 2021) compresses $\tilde{x}$ into the memory $\mathbf{M}$, one token per heavy (non-hydrogen) element of F is appended, carrying that element's count (twelve element types). An autoregressive decoder pretrained in MolForge (Ucak et al., 2023) and fine-tuned on the training split cross-attends to it: $p_\psi(y \mid X, F) = \prod_t p_\psi(y_t \mid y_{<t}, \mathbf{M}, F)$. A beam of $B = 40$ SMILES strings, constrained to the heavy-atom counts of F, is canonicalized as one string per molecule, no stereochemistry, via RDKit Landrum et al., 2026, and deduplicated into the pool $\mathcal{P}$ of distinct valid (parseable) molecules. A member's decoder score $\log p_k$ is derived from its molecule marginal score by taking the log of the summed probability of its beam strings, and length-normalized as in the beam (Wu et al., 2016). If the truth is not among the B strings, the molecule is out-of-beam: no ranking of $\mathcal{P}$ can recover it, which Graph-SSI aims to address.

Graph-SSI (4). (4a) Whether. Using the label-free pool features (decoder confidence, fit of the pool to each spectrum), a gradient-boosted gate (Friedman, 2001; Pedregosa et al., 2011), fitted on the validation set (App. F), estimates the probability $\hat{p}$ that the truth is missing from $\mathcal{P}$, and sends the molecule to repair iff $\hat{p} \geq 0.01$; otherwise $\mathcal{P}$ passes on unchanged to avoid displacement of a correct answer, which can happen directly or via the standardization in Eq. (2). **(4b) Where.** Pool members share the heavy atoms of F and differ only in bonds and atom states such as hydrogen count and formal charge. We write each member as a Kekulé heavy-atom graph with bond orders $E_{ij} \in \{0, 1, 2, 3\}$, with 0 meaning no bond, map its atoms onto those of the decoder's top-1 member y_1 by an element-preserving atom map. The atom map can be understood as a linear assignment on local bond environments. We read the pool as a distribution over the bond order of each atom pair:

$$q_{ij}(c) = \sum_k w_k \, \mathbb{I}\big[\hat{E}_{ij}^{(k)} = c\big], \qquad w_k = \frac{e^{\ell_k}}{\sum_{k'} e^{\ell_{k'}}}, \qquad \tau_{ij} = 1 - \max_c q_{ij}(c), \tag{1}$$

where $\hat{E}_{ij}^{(k)}$ is member k's aligned bond order and ℓ_k its total decoder log-probability; k, k' range over the members with y_1's heavy atoms. The disagreement τ_{ij} is zero where members agree and marks where the decoder is unsure. A region R is the four atoms of a high-τ atom pair and a disjoint partner pair (the atoms one double-edge swap touches). Their atom states and the bond orders of their six atom pairs form G_R, and the rest, $G_{\setminus R}$, is fixed. Regions are placed in up to four top-ranked members: since $G_{\setminus R}$ is fixed, a repair can only reach a truth that matches the edited member outside its regions. **(4c) How.** Two generators regenerate the regions, and their proposals are pooled. *Bond swaps* are degree-preserving double-edge swaps (Maslov & Sneppen, 2002) of equal-order bonds in R, up to 1,200 proposals per molecule. *Spectrum-conditioned diffusion* is

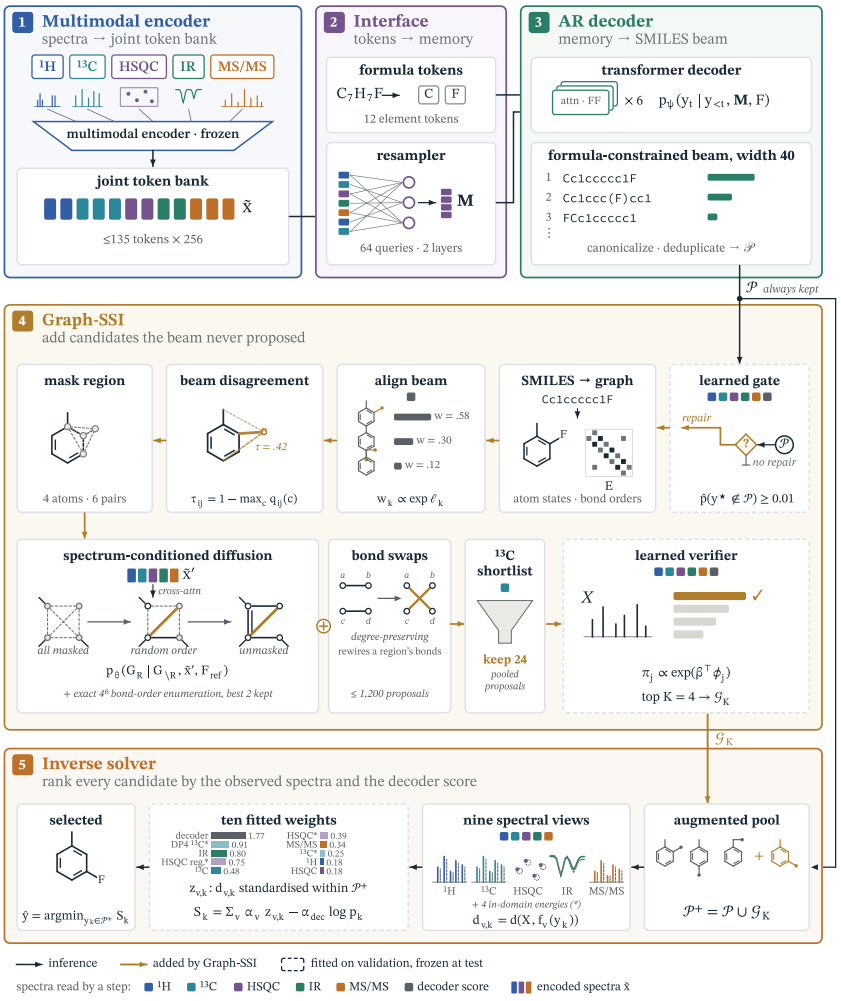


Figure 1: **TAG-XMM.** Illustration of the TAG-XMM architecture (gold: added by Graph-SSI; dashed: fitted on validation). Fig. 2 gives a worked example of how it applies to a molecule.

an absorbing-state (masked) discrete diffusion model (Austin et al., 2021) on the molecular graph (Vignac et al., 2023), $p_\theta(G_R \mid G_{\setminus R}, \tilde{x}', F_{\text{ref}})$, trained on random regions of true training-split graphs with random subsets of spectra withheld. It regenerates G_R from the per-spectrum encoder tokens $\tilde{x}'$ and the formula F_{ref} of y_1, unmasking in random order; invalid samples are projected to valid graphs where possible, and an exact search over the 4^6 bond-order combinations of R adds the two most probable. Unlike swaps, diffusion can change degrees and hydrogen counts, within F_{ref}. **(4d) Which.** A ^{13}C shortlist keeps the 24 proposal structures of which the predicted ^{13}C shifts best fit the observed spectrum. The verifier then ranks them and the pool members by a linear score $\beta^\top \phi_j$ of nine scores: the decoder's teacher-forced log-probability of j's SMILES and eight spectral mismatches, with the lower score signifying a better match (App. F): ^{13}C spectral lines with their C–F and C–P multiplets, a joint ^{13}C–HSQC assignment, HSQC cross-peaks, ^{1}H peaks matched on shift, on shift and proton count, and on proton count and multiplicity, the IR spectral trace and MS/MS fragments. Each enters as j's value standardized over the shortlist and the pool and as its gap to the best candidate; β is trained on validation swap proposals with a softmax loss that ranks the truth first (a conditional logit, McFadden, 1974). The $K = 4$ best new graphs $\mathcal{G}_K$ are added and nothing is removed, because the pool is right far more often than any single repair. An added graph's decoder score is its teacher-forced log-probability mapped onto the molecule-marginal scale by an isotonic fit on validation.

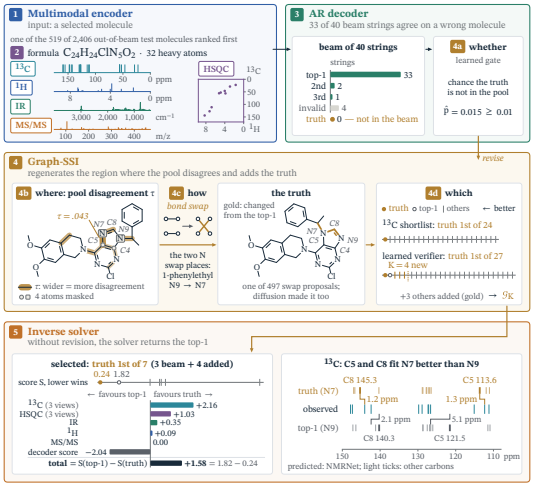


Figure 2: One out-of-beam test molecule ($C_{24}H_{24}ClN_5O_2$) through every stage of Fig. 1. Diffusion also generates the truth from another molecule's spectra, so here the spectra act at selection.

Selecting the answer (5). The verifier only decides what enters the augmented pool $\mathcal{P}^+ = \mathcal{P} \cup \mathcal{G}_K$; it is the inverse solver that makes the final choice among its members using ten views: the decoder score, five external views (NMRNet (Xu et al., 2025) for ^{13}C and ^{1}H, TransPeakNet (Li et al., 2025) for HSQC, Graphormer-IR (Stienstra et al., 2024) for IR fine-tuned on training spectra and, for MS/MS, the potential fragment masses that can be attained by bond breaking (Wolf et al., 2010)) (these first six build what we call the **six-view solver**) ; and four in-domain energies from shift heads trained on the corpus's simulated spectra (^{13}C and HSQC energies, a DP4-style (Smith & Goodman, 2010) ^{13}C likelihood and an HSQC register). Spectral view v scores candidate y_k by a mismatch $d_{v,k} = d\big(X, f_v(y_k)\big)$ between its prediction $f_v(y_k)$ and the observed spectra (lower is better), standardized within $\mathcal{P}_v^+$, the members of $\mathcal{P}^+$ it can score (others contribute zero); the decoder score enters raw (added graphs: isotonic map, 4d):

$$z_{v,k} = \frac{d_{v,k} - \mathrm{mean}_{\mathcal{P}_v^+} d_v}{\mathrm{sd}_{\mathcal{P}_v^+} d_v}, \qquad S_k = \sum_v \alpha_v \, z_{v,k} - \alpha_{\mathrm{dec}} \log p_k, \qquad \hat{y} = \underset{y_k \in \mathcal{P}^+}{\arg\min} \, S_k. \tag{2}$$

The weights are fitted once on validation pools without added graphs (a penalized conditional logit), then frozen, learning that the decoder score has the largest weight ($\alpha_{\mathrm{dec}} = 1.77$; App. E, Table 12).

4 Experiments, Analysis, and Performance

Setup. We use the simulated five-spectrum corpus of Alberts et al. (2024b) (715,840 / 39,284 / 39,238 training / validation / test sets). All results below are on test: 39,224 molecules are evaluated, and the inverse solver and Graph-SSI are scored on all of them. Exact@k denotes the share of molecules whose canonical SMILES structure (disregarding stereochemistry) is contained among the first k distinct candidates. Models are fitted on training or validation data (cf. App. A).

Encoder. We test the encoder by same-formula retrieval on 4,095 test molecules in 1,223 formula families of two or more molecules (mean 3.35 per family). Each query is ranked only against molecules with the same formula, so the task is to tell isomers apart. Given the shared formula, chance of successful isomeric identification is about 30%, whereas the token bank with all five spectra reaches about 98% top-1 (Fig. 3): the representation separates almost all same-formula isomers in these families. Shapley values (Shapley, 1953) over the 32 spectrum subsets attribute the gain to each spectrum (App. Fig. 8). ^{1}H NMR contributes most to the attribution, consistent with its dense record of proton environments, integrals and couplings. HSQC follows in terms of weight of information, linking the local indications obtained from ^{1}H NMR through direct proton–carbon correspondences; IR follows by largely adding non-overlapping functional-group information. To test whether these results require joint encoding, we compare them against the results obtained from a late-fusion baseline. In late-fusion, each spectrum is encoded alone by the jointly trained encoder, for which single-spectrum inputs are out of distribution (App. C), and the retrieval scores are combined by one set of non-negative weights for all sets, with the weights being selected on

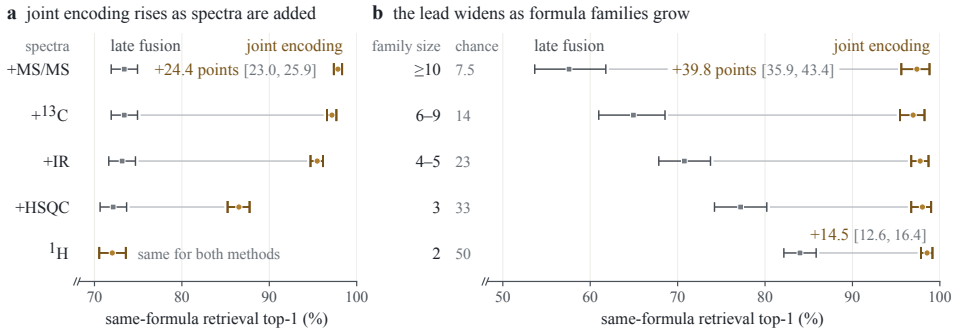


Figure 3: **Joint encoding gains with each added spectrum (a); its lead widens as families grow (b).** Share of 4,095 test molecules whose spectra rank their structure first among same-formula ones (1,223 families). The models is trained by masking spectra in subsets, using encoder trained on all five spectra; late fusion combines single-spectrum scores with validation-fitted weights. (a) Best set per size on validation; each added spectrum helps (paired 95% intervals above zero); chance 29.9%. (b) All five spectra; family size: test molecules per formula. Intervals: 95% formula-family bootstrap, paired for leads.

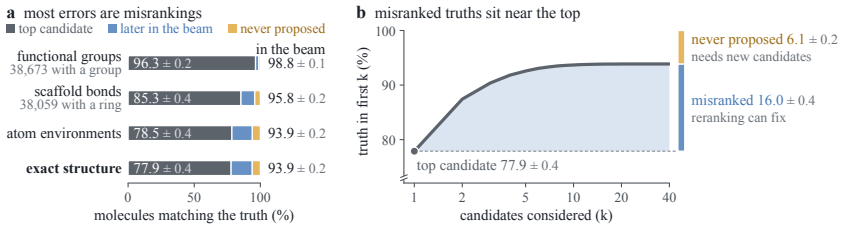


Figure 4: **The decoder's structure contains the truth in its beam for 94% of test molecules but ranks it first for only 78%.**. The correct structure is correctly proposed by the decoder somewhere in the beam for 94% of the tested molecules but at top-1 only for 78% of the tested structures. (a) Agreement with the truth at four levels of detail, for the top candidate and for the whole beam. (b) Exact@k on the 39,224 test molecules; the strip splits the top candidate's misses into misranked and never proposed. CIs are 95% Wilson.

validation and applied unchanged to test. With all five spectra, late fusion reaches 73.41%, 24.44 points below joint encoding (95% CI [23.01, 25.87]; Efron, 1979), and joint encoding performs better on predictions based on every set of two or more spectra (App. Fig. 9). Further, whereas late fusion never beats ^{1}H NMR alone by more than 1.4 points, joint encoding keeps improving as spectra are added: every multimodal set (except ^{13}C + MS/MS) improves its score from its best subset with one spectrum less (95% CI excluding zero; App. C).

Decoder. We score the decoder's suggested molecular candidates at four levels of chemical detail (Fig. 4a; metrics in App. D). Its top candidate correctly predicts all functional groups right for 96.35% of molecules and every scaffold bond for 85.33%, but every atom environment for only 78.49% and the exact structure for 77.88% (Exact@1; 30,548 of 39,224): the decoder identifies the broad chemistry and fails at the wiring. Most of these errors are misrankings: the 40-string beam holds the exact structure for 93.87% (16.0 points more; Exact@10 93.68%), usually second or third (Fig. 4b), which motivates the inverse solver; for the other 6.1% it is never proposed, which motivates Graph-SSI. Widening the beam has diminishing returns: the last 20/40 add only 0.56 points (App. D).

Graph-SSI. Of the 39,224 test molecules, 36,818 are in-beam and 2,406 out-of-beam; no ranking can recover the latter, so this is where Graph-SSI must act. Of those 2,406, Graph-SSI adds the truth to the pool for 717 (29.8%). The inverse solver selects 519 of them at rank one (21.6% of the 2,406; Fig. 5c). On in-beam molecules the cost is small (Exact@1 95.92% $\rightarrow$ 95.74%; 135 gained, 198 lost,

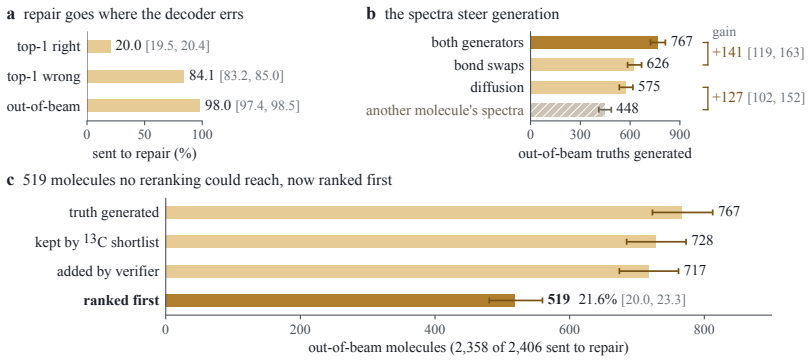


Figure 5: **Graph-SSI correctly ranks first 519 of 2,406 molecules (21.6%) that no reranking can reach.** In-beam: the truth is in the decoder's beam, with its top-1 right (30,593 molecules) or wrong (6,225); out-of-beam: it is not (2,406). (a) Share of each group sent to revise. (b) Out-of-beam truths generated; hatched: control, another molecule's spectra; brackets: paired gains. (c) Of the 2,358 sent to revise, out-of-beam molecules at each step: the ^{13}C shortlist keeps 24 graphs, the verifier adds the best few, the inverse solver ranks them. Intervals: 95% Wilson

exact McNemar $p = 6.6 \times 10^{-4}$; McNemar, 1947). Overall, Exact@1 rises from 90.03% to 91.19% (net $+456$ molecules, $p = 1.2 \times 10^{-57}$; Table 1). The isotonic scoring rule was adopted after an earlier test run (App. A); with that run's pre-registered rule, Graph-SSI still gains 0.72 points (665 gained, 384 lost). The selected share decreases with size (fitted: 30% of out-of-beam molecules at 11 heavy atoms, 17% at 35), but most out-of-beam molecules are large (App. Fig. 10). Revision works best on local errors: Graph-SSI generates the truth for 70.1% of out-of-beam molecules whose error is one bond swap, and for 12.6% of those whose error spans more than two (App. F).

What each component contributes. The gate sends 34.9% of pools to repair: 98.0% of out-of-beam molecules and 84.1% of in-beam ones whose decoder top-1 (best string) is wrong, but only 20.0% of those whose top-1 is right (Fig. 5a). On the 2,358 out-of-beam molecules sent to repair (Fig. 5b), bond swaps (up to 1,200 proposals per molecule) generate 626 truths, diffusion (at most 728 draws) 575 and both pooled 767 ($p = 7 \times 10^{-43}$ vs swaps alone); diffusion reaches more of the 398 errors that move a hydrogen between atoms (43 vs 17). With the same number of proposals per molecule (median 147), diffusion generates 1.6 times as many truths as swaps (571 vs 366, $p = 6 \times 10^{-28}$). The spectra steer diffusion: conditioned on another molecule's spectra (other-molecule control), it generates 22% fewer truths (575 $\rightarrow$ 448). The ^{13}C shortlist keeps 728 of the 767 truths (Fig. 5c; App. Table 13).

Inverse solver. The solver must choose from its pool, which contains the truth for 93.80% of the 39,224 molecules (at most 95.63% with Graph-SSI). Without Graph-SSI, it raises Exact@1 from 78.00% for the decoder's best string to 90.03% (5,322 gained, 601 lost; Table 1), against 85.61% for the six-view solver, the ablation baseline (decoder best-string score plus one external view per spectrum; 2,238 gained, 504 lost Fig. 6a). Each change to the six-view solver helps, in this order (App. E): the molecule-marginal decoder score, which sums a molecule's string probabilities (+1.19 points), the in-domain ^{13}C and HSQC energies (+2.42), and the DP4-style ^{13}C likelihood with the HSQC register (+0.81); the step gains depend on the order, as a pre-registered variant without the in-domain ^{13}C and HSQC energies reaches 90.04%. The gain comes from the spectra: scored against another molecule's spectra, the four in-domain energies add nothing (within 0.03 points; Fig. 6a, diamonds), and on the 37,142 molecules outside the in-domain shift heads' training data (App. A), Exact@1 still rises from 84.88% (six-view solver) to 89.54%. The gain lands where the decoder errs: when its top-1 is wrong but the truth is in the beam, the solver ranks the truth first for 85.5% (six-view solver: 63.5%), and it keeps more correct top-1s (98.04% vs 96.85%; Fig. 6b).

Overall system performance. TAG-XMM reaches 91.19% Exact@1 on all 39,224 test molecules (Table 1). On each baseline's test molecules it beats NMRViT by 11.6 points (7,641 molecules) and NMIRacle by 35.4 points (3,126), reading all five spectra, each baseline its own subset. On each

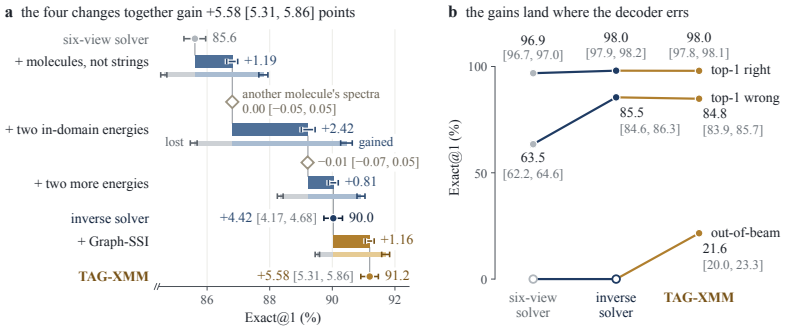


Figure 6: **Every change adds correct answers additively** (a) Exact@1 on 39,224 test molecules as the respective change as indicated on the y axis is added; the improvement is denoted relative to the step before. Fat bars indicated net gain; thin bars indicate the proportion of molecules gained and lost; left of each system indicates the total gain over the six-view solver; diamonds: indicate control, the same energies on another molecule's spectra. (b) Exact@1 when the decoder's top-1 is right (30,593), wrong with the truth in its beam (6,225) or out-of-beam (2,406; hollow: zero). Intervals: 95% Wilson, paired bootstrap for net changes.

Table 1: **TAG-XMM on the test split.** (a) Exact@1 for each ranking stage, overall and split by whether the decoder's beam contains the true structure. (b) Graph-SSI off and on, with the six-view solver and with the inverse solver. (c) Baselines on the same molecules. On all 39,224 test molecules TAG-XMM reaches 91.19% against 77.88% for the decoder's raw top-1.

(a) Ranking stage	Exact@1 (%)		
	All	In-beam	Out-of-beam
molecules	39,224	36,818	2,406
Decoder, best string	78.00	83.09	0.0
Decoder, molecule marginal	81.74	87.09	0.0
Solver, spectral views only	88.56	94.35	0.0
Inverse solver	90.03	**95.92**	0.0
+ Graph-SSI (TAG-XMM)	**91.19**	95.74	**21.6**

(b) Off / on	Graph-SSI	
	off	on
Six-view solver	85.61	86.43†
Inverse solver	90.03	**91.19**

(c) Baseline	Baseline's spectra				
	baseline	our dec.	+ sol.	+ G-SSI	all five
NMRViT	80.02	75.03	85.43	86.26	91.66
NMIRacle	55.63	59.34	72.07	n/r‡	91.07

In-beam: the decoder's beam contains the true structure; *out-of-beam*: it does not, so only Graph-SSI can recover it. *Best string* scores a candidate by its most probable SMILES string, *molecule marginal* by the sum over all its strings; *spectral views only* drops the decoder score. Against the six-view solver, the inverse solver gains 2,238 molecules and loses 504; Graph-SSI then gains 654 (519 out-of-beam) and loses 198. †Sensitivity run from an earlier test (App. A). (c) TAG-XMM's performance compared to baselines on the intersection of our test split and each baseline's (rebuilt from released code): NMRViT, 7,641 molecules (^{1}H, ^{13}C, HSQC, formula, trained on 70% of the corpus vs. 90% for TAG-XMM); NMIRacle, 3,126 (^{1}H, ^{13}C, IR, no formula). *Itself*: the baseline's own Exact@1. *Baseline's spectra*: TAG-XMM sees only those inputs, with its decoder retrained and everything else frozen; these pre-registered runs predate the fix that gave every test molecule a solver pool, so 1,304 NMRViT and 741 NMIRacle molecules keep the retrained decoder's top-1 (3 NMIRacle molecules have no test row and count as misses). ‡Graph-SSI runs only if its gate fires at 0.8–1.25 times its all-five rate; on NMIRacle's inputs it fires 1.55 times as often (NMRViT's: 1.07).

baseline's own spectra, TAG-XMM still leads: 86.26% vs 80.02% for NMRViT and, without the formula as in NMIRacle, 72.07% vs 55.63% (Table 1c). On NMRViT's spectra the decoder alone is below NMRViT (75.03% vs 80.02%), so the lead comes from the solver and Graph-SSI.

On the fitted curves of Fig. 7, TAG-XMM's Exact@1 falls from 97.5% at 8 heavy atoms to 78.6% at 35, but each stage adds most where the decoder is weakest: the solver's gain over the decoder's best string grows from 4.1 to 19.8 points, and Graph-SSI's from 0.1 to 2.5. The molecules Graph-SSI gains have a median of 27 heavy atoms.

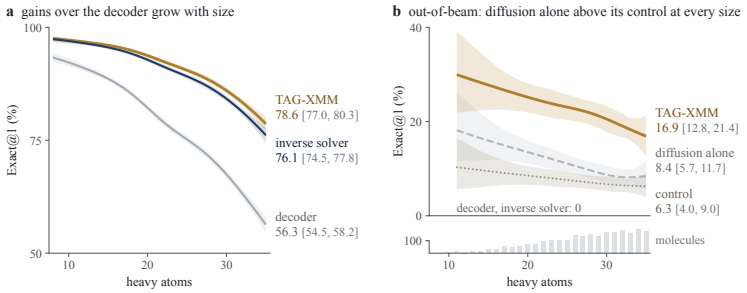


Figure 7: **Gains from the decoder grow with molecule size.** Fits with 95% bootstrap bands (2,000 resamples; Efron, 1979). Labels: value at 35 heavy atoms [95% band]. (a) All 39,224 test molecules; TAG-XMM's paired gain over the inverse solver is positive from 10 heavy atoms and reaches +2.5 [1.7, 3.4] at 35. (b) 2,406 out-of-beam molecules, never proposed by the decoder; dashed: diffusion alone and, dotted, given another molecule's spectra (control), ranked by the six-view solver, unlike TAG-XMM in generator and solver; dashed is above dotted at every size (paired 95% band).

5 Limitations

While our results are promising, we must acknowledge three main limitations to their scope and interpretation: **(1)**, all spectra are simulated; therefore, accuracy will likely drop when measured spectra as used, as impurity peaks, noise and referencing will reduce the signal-to-noise ratio. Further, the in-domain energies fit the corpus's own simulator (Alberts et al., 2024b); the other spectral views use external predictors, and the other-molecule controls (Figs. 5b and 6a) tie the gains to each molecule's own spectra. **(2)**, The encoder and decoder were trained on all five spectra, so subsets are out of distribution. This means that without retraining, the decoder's output collapses; App. D; however, with decoder retrained, TAG-XMM still beats both baselines on their own spectra (Table 1c). **(3)**, The true formula is given, which is equivalent to the HRMS fixing it unambiguously, asan accurate mass can fit several formulas. Without HRMS and on NMIRacle's inputs, TAG-XMM still reaches 72.07% (Table 1c), but its gate fires 1.55 times as often as with all five spectra, outside the pre-registered 0.8–1.25 range, so Graph-SSI was not run in this case, (outlind by n/r in the table).

6 Conclusion and Future Work

TAG-XMM is a model that elucidates molecular structure from five spectra and the molecular formula using five computational components in its pipeline: a jointly trained multimodal encoder, a resampler, a formula-constrained autoregressive decoder, the molecular pool revision component Graph-SSI, and an inverse solver that subsequently ranks every suggested molecular candidate with the decoder score and the observed spectra. On the 39,224 held-out test split of the simulated corpus, TAG-XMM reaches 91.19% Exact@1. Each component within this pipeline has a distinct role: The decoder recovers broad chemistry of almost every molecule but the exact structure of only 77.88%. Graph-SSI addresses the decoder's blind spots: by regenerating uncertain regions of the candidate molecular graphs with bond swaps and spectrum-conditioned diffusion, it adds the true structure for 717 of the 2,406 out-of-beam molecules. The inverse solver succeeds to rank 519 of them first (21.6% of the 2,406). Where the decoder's top-1 structure does not match the truth, but the latter is within the beam, the inverse solver ranks the truth first for 85.5% of molecules, against 63.5% for the six-view solver. Three directions follow from where the system still fails. Firstly, Graph-SSI generates the truth for only a third of the out-of-beam molecules sent to review (767 of 2,358), and rarely when the error spans more than two bond swaps (12.6%), because both generators edit only small local regions; larger, non-local regions and edits that may change the formula F_{ref} would be required for verifiying the entire structure. Secondly, the solver selects 519 of the 717 added truths and loses 198 in-beam answers; fitting its weights on pools with added graphs is the direct next step. Thirdly, the limitations mentionned earlier define the future path and potential applicability of the model, and in chemical research: training and validation on experimental spectra, different subsets of spectra according to availability in the wet chemical laboratory , the model's performance without known formula, as well as for molecules with stereochemistry.

AI USE STATEMENT

In this work, we used generative AI tools to help develop theoretical models or conceptual frameworks (specifically, refinements and correctness verifications on details of the GraphSSI design), to propose or refine hypotheses, to design or provide feedback on research methodology or experiments (initiatives and ideas on the experiment methodology all originate from the authors directly, after which refinements and precise experiment selection feedback were discussed with AI tools), to implement methods (main method design ideas originate from authors and Generative AI helped write, iterate on, and refine code used to implement these ideas) and support qualitative and thematic data analysis (most of the plots present in the work, specifically, were designed at least in part with the help of Generative AI tools) and were lightly used to interpret results (a majority of results interpretation comes from authors directly, with assistance of Generative AI in reformulation and clarity of interpretations provided for example in annotations of the figures). Generative AI was also lightly used to clean and reformat the dataset (for occasional uses such as sharding the dataset for easier and faster processing, Generative AI tools never modified spectra or labels directly). AI coding agents also ran parts of the experimental pipeline under the authors' direction (evaluations, controls, statistical analyses and figure generation) and every reported number traces to logged outputs checked by the authors.

We have not used generative AI tools for the rest of the required disclosure tasks (Generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, assist with translation), as most of these are not directly relevant to this work.

Additionally, we used generative AI tools to create or modify scientific figures or images (design of figures was assisted by Generative AI tools for refinement), suggest experimental parameters (e.g., hyperparameter search), summarize or analyze existing literature (e.g., in Related Works section), discover research topics or identify gaps, brainstorming, sourcing/searching for information, identify relevant literature, format references, and also draft parts of a research paper (though most of the initial drafts come from the authors, which were then refined using Generative AI tools). Additionally, AI tools were heavily used to reformat and summarize main paper and appendix text (edit a research paper to improve readability) after first writing by authors to improve clarity, remove unnecessary text and verify consistency of results and interpretations and to create or edit software code (e.g., to write experiments, training, inference, and data analysis code).

Generative AI was not used to propose the title, keywords or paper structure, and those choices were made exclusively by the authors. The remainder of recommended disclosure tasks (Formulate questions for surveys or interviews, creation of artifacts, transcribe recordings of research material) are not applicable to this work.

We have reviewed all AI-assisted work. LLM-written code was consistently manually reviewed by at least one author for mathematical correctness, consistency with theoretical models. Training flow and testing was always manually monitored by authors for consistency with expectations, quality and interpretability of results contained. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

This paper aims to be a fully reproducible report of the TAG-XMM method: the main text explains each component (Sec. 3, Fig. 1), the appendix gives the fine-grained details needed to reproduce each one, and we took the following steps to make every result checkable:

- **Data and splits.** All experiments use the public simulated corpus of Alberts et al. (2024b). App. A gives the split sizes, every evaluation cohort and how it was formed (Table 2), and the evaluation units.
- **Test-set use.** Every model was fitted on training or validation data. App. A ("Test-set use and overlaps") lists each exception, every read of the test labels and the overlaps between test and training data. The test evaluations were pre-registered before scoring, those of Graph-SSI with their files locked by md5; extending the controls to all 39,224 test molecules was decided post hoc (App. A).

- **Statistics.** App. A ("Statistical comparisons") gives the paired tests and the bootstrap intervals with their resampling settings.
- **Spectral inputs.** App. B specifies how each spectrum becomes tokens (Table 4), their normalization (Table 5) and how the formula is used (Table 6).
- **Encoder.** App. C gives the architecture, the chemical training objectives (Table 7), the configuration and optimization settings (Table 8), and the retrieval, attribution and late-fusion protocols.
- **Resampler and decoder.** App. D gives their training, the beam search, the candidate pool and decoder score (Eq. 5), and the reconstruction metrics.
- **Inverse solver.** App. E gives the score (Eq. 6), its forward models and in-domain energies, the ten fitted weights (Table 12) and how they were fitted.
- **Graph-SSI.** App. F gives the gate, the alignment and region construction, both generators with their budgets, the shortlist and verifier, the scoring of added graphs, and the other-molecule control (Table 13).
- **Baselines.** We rebuilt NMRViT's split with its released builder and used NMIRacle's released test manifest (Table 2); App. B ("Runs on the baselines' spectra") describes the input masking and decoder retraining behind Table 1c.
- **Worked example.** Fig. 2 follows one test molecule through every stage with its intermediate values.

A complete version of the code, train/test/validation split builders on the dataset, as well as per-molecule outcome files is available at this link.

REFERENCES

Marvin Alberts, Federico Zipoli, and Alain C. Vaucher. Learning the language of NMR: Structure elucidation from NMR spectra using transformer models. ChemRxiv preprint, 2023. URL https://doi.org/10.26434/chemrxiv-2023-8wxcz.

Marvin Alberts, Teodoro Laino, and Alain C. Vaucher. Leveraging infrared spectroscopy for automated structure elucidation. *Communications Chemistry*, 7(1):268, 2024a. doi: 10.1038/s42004-024-01341-w. URL https://doi.org/10.1038/s42004-024-01341-w.

Marvin Alberts, Oliver Schilter, Federico Zipoli, Nina Hartrampf, and Teodoro Laino. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 125780–125808. Curran Associates, Inc., 2024b. doi: 10.52202/079017-3996. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e38e60b33bb2c6993e0865160cdb5cf1-Paper-Datasets_and_Benchmarks_Track.pdf. Datasets and Benchmarks Track.

Marvin Alberts, Nina Hartrampf, and Teodoro Laino. Automated structure elucidation at human-level accuracy via a multimodal multitask language model. ChemRxiv preprint, 2025. URL https://doi.org/10.26434/chemrxiv-2025-q80r9.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928. URL https://doi.org/10.1021/jm9602928.

Montgomery Bohde, Mrunali Manjrekar, Runzhong Wang, Shuiwang Ji, and Connor W. Coley. DiffMS: Diffusion generation of molecules conditioned on mass spectra. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4737–4756. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/bohde25a.html.

Wolfgang Bremser. HOSE — a novel substructure code. *Analytica Chimica Acta*, 103(4):355–365, 1978. doi: 10.1016/S0003-2670(01)83100-7. URL https://doi.org/10.1016/S0003-2670(01)83100-7.

Chein-I Chang. An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis. *IEEE Transactions on Information Theory*, 46(5):1927–1932, 2000. doi: 10.1109/18.857802. URL https://doi.org/10.1109/18.857802.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885, 2020. URL https://arxiv.org/abs/2010.09885.

David F. Crouse. On implementing 2D rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, August 2016. doi: 10.1109/TAES.2016.140952. URL https://doi.org/10.1109/TAES.2016.140952.

Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3(10):1503–1507, 2008. doi: 10.1002/cmdc.200800178. URL https://doi.org/10.1002/cmdc.200800178.

Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu, and Sebastian Böcker. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences*, 112(41):12580–12585, 2015. doi: 10.1073/pnas.1509788112. URL https://doi.org/10.1073/pnas.1509788112.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

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. doi: 10.1016/j.pnmrs.2007.04.003. URL https://doi.org/10.1016/j.pnmrs.2007.04.003.

Jean-Loup Faulon. Stochastic generator of chemical structure. 2. Using simulated annealing to search the space of constitutional isomers. *Journal of Chemical Information and Computer Sciences*, 36(4):731–740, 1996. doi: 10.1021/ci950179a. URL https://doi.org/10.1021/ci950179a.

Bailey K. Fosdick, Daniel B. Larremore, Joel Nishimura, and Johan Ugander. Configuring random graph models with fixed degree sequences. *SIAM Review*, 60(2):315–355, 2018. doi: 10.1137/16M1087175. URL https://doi.org/10.1137/16M1087175.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451. URL https://doi.org/10.1214/aos/1013203451.

Nicolás Grimblat, María M. Zanardi, and Ariel M. Sarotti. Beyond DP4: an improved probability for the stereochemical assignment of isomeric compounds using quantum chemical calculations of NMR shifts. *The Journal of Organic Chemistry*, 80(24):12526–12534, 2015. doi: 10.1021/acs.joc.5b02396. URL https://doi.org/10.1021/acs.joc.5b02396.

Yanfei Guan, S. V. Shree Sowndarya, Liliana C. Gallegos, Peter C. St. John, and Robert S. Paton. Real-time prediction of ^{1}H and ^{13}C chemical shifts with DFT accuracy using a 3D graph neural network. *Chemical Science*, 12(36):12012–12026, 2021. 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. doi: 10.1039/D6SC02352E. URL https://doi.org/10.1039/D6SC02352E.

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. doi: 10.1186/s13321-015-0068-4. URL https://doi.org/10.1186/s13321-015-0068-4.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022. URL https://arxiv.org/abs/2207.12598.

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. doi: 10.1021/acscentsci.4c01132. URL https://doi.org/10.1021/acscentsci.4c01132.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Jan H. Jensen. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chemical Science*, 10(12):3567–3572, 2019. doi: 10.1039/C8SC05372C. URL https://doi.org/10.1039/C8SC05372C.

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*, 17(1): 4740, 2026. doi: 10.1038/s41467-026-71315-0. URL https://doi.org/10.1038/s41467-026-71315-0.

Eric Jonas. Deep imitation learning for molecular inverse problems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/b0bef4c9a6e50d43880191492d4fc827-Abstract.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109. URL https://doi.org/10.1002/nav.3800020109.

Greg Landrum et al. RDKit: Open-source cheminformatics. https://www.rdkit.org, 2026. Zenodo, doi: 10.5281/zenodo.591637.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Dominik Lemm, Guido Falk von Rudorff, and O. Anatole von Lilienfeld. Impact of noise on inverse design: the case of NMR spectra matching. *Digital Discovery*, 3(1):136–144, 2024. doi: 10.1039/d3dd00132f. URL https://doi.org/10.1039/d3dd00132f.

Yunrui Li, Hao Xu, Ambrish Kumar, Duo-Sheng Wang, Christian Heiss, Parastoo Azadi, and Pengyu Hong. TransPeakNet for solvent-aware 2D NMR prediction via multi-task pretraining and unsupervised learning. *Communications Chemistry*, 8(1):51, 2025. doi: 10.1038/s42004-025-01455-9. URL https://doi.org/10.1038/s42004-025-01455-9.

Eleni E. Litsa, Vijil Chenthamarakshan, Payel Das, and Lydia E. Kavraki. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Communications Chemistry*, 6(1):132, 2023. doi: 10.1038/s42004-023-00932-3. URL https://doi.org/10.1038/s42004-023-00932-3.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1711.05101.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11451–11461, 2022. doi: 10.1109/CVPR52688.2022.01117. URL https://openaccess.thecvf.com/content/CVPR2022/html/Lugmayr_RePaint_Inpainting_Using_Denoising_Diffusion_Probabilistic_Models_CVPR_2022_paper.html.

Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002. doi: 10.1126/science.1065103. URL https://doi.org/10.1126/science.1065103.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka (ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press, New York, 1974. URL https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf.

Charles McGill, Michael Forsuelo, Yanfei Guan, and William H. Green. Predicting infrared spectra with message passing neural networks. *Journal of Chemical Information and Modeling*, 61(6): 2594–2609, 2021. doi: 10.1021/acs.jcim.1c00055. URL https://doi.org/10.1021/acs.jcim.1c00055.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996. URL https://doi.org/10.1007/BF02295996.

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. doi: 10.1038/s41467-026-73846-y. URL https://doi.org/10.1038/s41467-026-73846-y.

Federico Ottomano, Yingzhen Li, and Alex M. Ganose. NMIRacle: Multi-modal generative molecular elucidation from IR and NMR spectra. arXiv preprint arXiv:2512.19733, 2025. URL https://arxiv.org/abs/2512.19733. AI for Accelerated Materials Design (AI4Mat) workshop at ICLR 2026.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85): 2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Martin Priessner, Richard J. Lewis, Isak Lemurell, Magnus J. Johansson, Jonathan M. Goodman, Jon Paul Janet, and Anna Tomberg. Advancing structure elucidation with a flexible multi-spectral AI model. *Angewandte Chemie International Edition*, 65(2):e17611, 2026. doi: 10.1002/anie.202517611. URL https://doi.org/10.1002/anie.202517611.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022. doi: 10.1038/s42256-022-00580-7. URL https://doi.org/10.1038/s42256-022-00580-7.

L. S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, number 28 in Annals of Mathematics Studies, pp. 307–318. Princeton University Press, 1953. doi: 10.1515/9781400881970-018. URL https://doi.org/10.1515/9781400881970-018.

Steven G. Smith and Jonathan M. Goodman. Assigning stereochemistry to single diastereoisomers by GIAO NMR calculation: The DP4 probability. *Journal of the American Chemical Society*, 132(37):12946–12959, 2010. doi: 10.1021/ja105035r. URL https://doi.org/10.1021/ja105035r.

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. doi: 10.1021/ci000407n. URL https://doi.org/10.1021/ci000407n.

Cailum M. K. Stienstra, Liam Hebert, Patrick Thomas, Alexander Haack, Jason Guo, and W. Scott Hopkins. Graphormer-IR: Graph transformers predict experimental IR spectra using highly specialized attention. *Journal of Chemical Information and Modeling*, 64(12):4613–4629, 2024. doi: 10.1021/acs.jcim.4c00378. URL https://doi.org/10.1021/acs.jcim.4c00378.

Michael A. Stravs, Kai Dührkop, Sebastian Böcker, and Nicola Zamboni. MSNovelist: de novo structure generation from mass spectra. *Nature Methods*, 19(7):865–870, 2022. doi: 10.1038/s41592-022-01486-3. URL https://doi.org/10.1038/s41592-022-01486-3.

Umit V. Ucak, Islambek Ashyrmamatov, and Juyong Lee. Reconstruction of lossless molecular representations from fingerprints. *Journal of Cheminformatics*, 15(1):26, 2023. doi: 10.1186/s13321-023-00693-0. Code and trained models released as MolForge: https://github.com/knu-lcbc/MolForge.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018. URL https://arxiv.org/abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. DiGress: Discrete denoising diffusion for graph generation. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=UaAD-Nu86WX.

Fei Wang, Jaanus Liigand, Siyang Tian, David Arndt, Russell Greiner, and David S. Wishart. CFM-ID 4.0: More accurate ESI-MS/MS spectral prediction and compound identification. *Analytical Chemistry*, 93(34):11692–11700, 2021. doi: 10.1021/acs.analchem.1c01465. URL https://doi.org/10.1021/acs.analchem.1c01465.

Yinkai Wang, Xiaohui Chen, Liping Liu, and Soha Hassoun. MADGEN: Mass-spec attends to de novo molecular generation. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=78tc3EiUrN.

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. doi: 10.1021/ci00057a005. URL https://doi.org/10.1021/ci00057a005.

Sebastian Wolf, Stephan Schmidt, Matthias Müller-Hannemann, and Steffen Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11(1):148, 2010. doi: 10.1186/1471-2105-11-148. URL https://doi.org/10.1186/1471-2105-11-148.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144, 2016. URL https://arxiv.org/abs/1609.08144.

Yutong Xie, Chence Shi, Hao Zhou, Yuwei Yang, Weinan Zhang, Yong Yu, and Lei Li. MARS: Markov molecular sampling for multi-objective drug discovery. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=kHSu4ebxFXY.

Ziyu Xiong, Yichi Zhang, Foyez Alauddin, Chu Xin Cheng, Joon Soo An, Mohammad R. Seyedsayamdost, and Ellen D. Zhong. Atomic diffusion models for small molecule structure elucidation from NMR spectra. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-3868. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a845fdc3f87751710218718adb634fe7-Abstract-Conference.html.

Fanjie Xu, Wentao Guo, Feng Wang, Lin Yao, Hongshuai Wang, Fujie Tang, Zhifeng Gao, Linfeng Zhang, Weinan E, Zhong-Qun Tian, and Jun Cheng. 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. doi: 10.1038/s43588-025-00783-z. URL https://doi.org/10.1038/s43588-025-00783-z.

Zhuo Yang, Jianfei Song, Minjian Yang, Lin Yao, Jiahua Zhang, Hui Shi, Xiangyang Ji, Yafeng Deng, and Xiaojian Wang. Cross-modal retrieval between ^{13}C NMR spectra and structures for compound identification using deep contrastive learning. *Analytical Chemistry*, 93(50):16947–16955, 2021. doi: 10.1021/acs.analchem.1c04307. URL https://doi.org/10.1021/acs.analchem.1c04307.

A Reproducibility and evaluation protocol

Data and splits. We use the simulated multimodal corpus of Alberts et al. (2024b), whose molecules are sampled from the USPTO patent dataset: 715,840/39,284/39,238 training/validation/test records, 39,224 test molecules pass record-level decoder eligibility (Table 2). Each record contains ^{1}H NMR, ^{13}C NMR, HSQC, IR, and positive- and negative-mode MS/MS spectra with the molecular formula. Its ^{1}H NMR, ^{13}C NMR and HSQC spectra are simulated with MestReNova ($CDCl_3$), its IR spectra (400–4,000 cm^{-1}, 1,800 points) are computed from the dipole moment along GAFF molecular-dynamics trajectories (LAMMPS), and the MS/MS spectra we use are simulated with CFM-ID 4.0 (Wang et al., 2021) in positive and negative mode at 10, 20 and 40 eV (the corpus's additional SCARF and ICEBERG positive-mode spectra are not used). Model selection, loss weighting, the fitting of the solver, gate and verifier, and checkpoint selection use training or validation data only.

Evaluation units. We canonicalize every valid prediction with RDKit (Landrum et al., 2026) to a canonical non-stereo SMILES before deduplicating a beam; formula-budget masking during beam search keeps every candidate within the heavy-atom counts of the formula. Exact@k is the fraction of molecules whose true canonical non-stereo SMILES is among the first k distinct candidates. No test truth and no top-1 candidate carries stereo marks, so Exact@k scores the structure (constitution); matching on the first block of the InChIKey (Heller et al., 2015) instead would give the decoder 78.18% rather than 77.88% Exact@1 (that key also merges tautomers). A tie at the minimum of the solver score counts as an error, and so does a molecule whose truth is not among the solver's candidates (24 in-beam molecules, all with a wrong decoder top-1, lose the truth when the solver builds its candidate set, so the solver's ceiling is 36,794), so every solver number has the denominator 39,224. Units are test records; a few molecules occur in more than one record (65 repeat records), and each record is scored. Every reported result is on the test split; quantities fitted on validation (solver weights, gate, verifier, isotonic map, and the noise model and line costs of the DP4-style ^{13}C likelihood and the event costs of the HSQC register) are marked as such.

Statistical comparisons. Comparisons between two systems are always paired over the same molecules; rankers share candidate pools except the Graph-SSI arm, which ranks the augmented pool. For the inverse solver and Graph-SSI we report the two discordant counts (molecules gained and lost) and an exact two-sided McNemar test (McNemar, 1947), a binomial test of the discordant counts at probability $1/2$; p-values below 10^{-200} are reported as a bound. Confidence intervals and bands are bootstrap intervals (Efron, 1979): over formula families for the encoder, so that the isomers of one family are resampled together (Sec. 4; Figs. 8 and 9), and over molecules, with 2,000 resamples, for Figs. 7 and 10. The Graph-SSI pre-registrations fixed their primary comparisons and verdict rule in advance (an improvement needs a net gain and $p < 0.05$): Test A declared six

Table 2: Evaluation cohorts (all test; the last row is the validation population used for fitting and selection). The inverse solver and Graph-SSI are evaluated on all 39,224 test molecules.

Cohort (use)	Starting	Eligible	Excluded	Requirement
Full test split (decoder, full system)	39,238	39,224	14	record-level decoder eligibility (13 of the 14 excluded records contain a [PH] atom)
Solver cohort (inverse solver, Graph-SSI)	39,224	39,224	0	solver pool built
In-beam subset (inverse solver)	39,224	36,818	–	truth in the beam; decoder top-1 (best string) right for 30,593, wrong for 6,225
Out-of-beam subset (Graph-SSI)	39,224	2,406	–	truth not in the beam; 2,358 sent to repair
Sent to repair (Graph-SSI)	39,224	13,705	–	gate fired (label-free)
Leak-free subset (inverse solver)	39,224	37,142	2,082	outside the in-domain shift heads' training data
Same-formula retrieval (encoder)	–	4,095	–	1,223 formula families with two or more members
Matched NMRViT	–	7,641	–	in our test split and NMRViT's, rebuilt with its released builder (seed 3245; NMRViT reports 79.12% on its own split (Han et al., 2026))
Matched NMIRacle	–	3,126	–	in our test split and NMIRacle's released test manifest
Fitting and selection (validation)	39,284	32,696	6,588	pools of the 39,276 evaluable records; 6,580 have none, by the same stereo-aware matching

primary comparisons with a Holm–Bonferroni adjustment, and Test A2 one primary comparison, scored together with its pre-registered sensitivity row. Each pre-registration was written before its score stage. Test A's and its amendment were made read-only at their lock, with their md5 sums entered in an append-only log of every operation on test inputs; the pre-registrations of the swap-only precursor and of Test A2 are also read-only, and Test A2's carries a lock section, appended before its score stage, with its md5 and those of its frozen inputs. In Test A, each label-free computation on test inputs was logged with its code md5, on loaders that delete the truth and evaluation fields (the solver's feature builder reads the true SMILES only as the key under which a molecule's observed spectra are stored). Arm results were read only after the configuration had matched the frozen one by md5 and a fixed anchor had been reproduced exactly: for Test A, the six-view solver's 28,053 of 32,571 correct and the decoder's 25,607; for Test A2, the inverse solver's 29,485 of 32,571, with no per-molecule mismatch. Secondary analyses report nominal p.

Notation. Table 3 lists the symbols and terms of Secs. 3 and 4.

Table 3: Notation and terms.

Symbol or term	Meaning
X; F; $y^\star$	observed spectra; molecular formula (from HRMS); true structure (which atoms are bonded, and by which bond orders; stereochemistry is not scored)
$\tilde{x}$; $\mathbf{M}$	encoder token bank (at most 135 tokens of width 256; Table 4); decoder memory: the resampler's 64 outputs and one token per heavy element of F
beam; pool $\mathcal{P}$	the 40 SMILES strings of formula-constrained beam search; their distinct valid molecules (canonical non-stereo SMILES)
in-beam; out-of-beam	truth among the 40 beam strings; truth not among them
$\log p_k$ (decoder score)	log molecule marginal of pool member y_k (summed probability of its beam strings, length-normalized); for an added graph, its teacher-forced log-probability mapped by the isotonic fit
y_1; ℓ_k; w_k	top pool member (the decoder's top-1); total decoder log-probability of member k; its weight in Eq. (1)
$\hat{E}_{ij}^{(k)}$; $q_{ij}(c)$; τ_{ij}	bond order of aligned atoms i and j in member k (0: no bond); weighted share of members with bond order c; disagreement $1 - \max_c q_{ij}(c)$
$\hat{p}$; sent to repair	the gate's estimate that the truth is missing from $\mathcal{P}$; $\hat{p} \geq 0.01$
R; G_R; $G_{\setminus R}$	region (four atoms, six atom pairs); its atom states (hydrogen count, formal charge) and bond orders; the rest of the graph, held fixed
$\tilde{x}'$; F_{ref}	per-spectrum encoder tokens read by the diffusion model p_θ; formula of y_1
ϕ_j; β	verifier features of candidate j (nine scores, each standardized and as a gap to the best); their weights, fitted on validation
K; $\mathcal{G}_K$; $\mathcal{P}^+$	number of added graphs ($K = 4$); the added graphs; the augmented pool $\mathcal{P} \cup \mathcal{G}_K$
generated; shortlisted; added; selected	truth among the proposals; kept by the ^{13}C shortlist (24); in $\mathcal{G}_K$; ranked first (Table 13)
f_v; $d_{v,k}$; $z_{v,k}$	forward model of view v; mismatch of its prediction for y_k with the observed spectrum (lower is better); $d_{v,k}$ standardized within $\mathcal{P}_v^+$
α_v; α_{dec}; S_k	weight of spectral view v; weight of the decoder score (1.77); the solver score, minimized over $\mathcal{P}^+$
inverse solver; six-view solver	the ten-view ranker of Eq. (2); the ablation that scores the decoder's best string plus one external view per spectrum
solver cohort; full test split	the test molecules with a solver pool (all 39,224); all 39,224 evaluated test molecules
Exact@k	share of molecules whose truth is among the first k distinct candidates

B SPECTRAL REPRESENTATIONS

Table 4 defines the fixed-capacity observations presented to the multimodal encoder. Padding is masked in the encoder and the resampler and excluded from all reported token counts. Each spectrum is processed by a separate branch; after the HSQC correspondence stage (App. C), the tokens of all branches are concatenated with the summary token and fused.

Table 4: Spectral inputs and their contribution to the encoder token bank. The first dimension is a fixed event capacity. IR is read as 120 patches and pooled to 16 tokens; MS/MS reserves 20 tokens per ionization mode. The complete bank contains at most $1+25+24+29+16+40 = 135$ tokens of width 256. †C–H join field, read from HSQC (see text). Neutral-loss masses and parities, fragment and loss compositions and 11 summary descriptors derive from the simulator's annotations of the true structure (App. A).

Input	Stored shape	Bank capacity	Physical information retained	Notes
^{1}H NMR	25×8	25	Chemical shift, log integration, four largest couplings, carbon-membership flag†, and matched carbon shift†	multiplicity embedded separately; descending integral order
^{13}C NMR	24×4	24	Chemical shift, protonation flag†, partner proton count†, and matched proton shift†	descending shift order
HSQC	29×4	29	Proton and carbon shifts, proton count, and merged-cross-peak flag	explicit H–C correspondence; descending carbon-shift order
IR	120×15	16	The 1,800-point intensity trace reshaped into contiguous patches	learned attention pooling
MS/MS	40×26	40	Mass-to-charge ratio, three collision-energy intensities, intensity summary, precursor fraction, neutral loss, parity, and fragment/loss compositions	20 positive- and 20 negative-mode slots; the three energies merged by m/z; descending precursor-fraction order
Summary token	1×256	1	Spectrum-presence vector (six bits, two for MS/MS), molecular-formula residual, and 418 engineered global descriptors (209 per-spectrum aggregates and element counts, 209 cross-modal statistics)	prepended before fusion

Capacities. On the 39,238 test records the bank holds 30 to 130 tokens per molecule (median 88, mean 85.8), summary token included. The HSQC capacity is never reached (at most 28 cross-peaks) and the ^{1}H capacity is reached by 2 records. The ^{13}C capacity binds: 15.0% of test records have more than 24 ^{13}C peaks, and peaks beyond the 24 most downfield are dropped. MS/MS keeps the 20 fragments per ionization mode with the largest mass fraction, which truncates most molecules (78.3% of test records have more than 20 positive-mode fragments, 55.0% more than 20 negative-mode ones). These limits apply to the encoder only: the inverse solver compares candidates with the observed spectra.

C–H join fields (†). Five columns of the one-dimensional NMR tokens are not read from the one-dimensional spectra. The featurizer matches each ^{1}H peak to the HSQC cross-peak nearest on the proton axis (within 0.05 ppm) and each ^{13}C peak to the cross-peak nearest on the carbon axis (within 1.0 ppm). A matched ^{1}H peak is flagged as bonded to carbon and receives that carbon's shift; a matched ^{13}C peak is flagged as protonated and receives the cross-peak's proton count and proton shift. These fields are therefore HSQC information: the protonation flag and proton count carry the carbon multiplicity that a multiplicity-edited experiment such as DEPT provides (the count is the cross-peak's total, so it can exceed three for equivalent or merged carbons), and the matched shifts are HSQC's one-bond correlations. HSQC also reaches the summary token, through per-spectrum

Table 5: Deterministic input normalization. These transformations are applied before padding and encoding. δ_H and δ_C denote chemical shifts (ppm) and couplings are in Hz; n_H denotes proton count; M_F is the monoisotopic mass of the formula (Da).

Spectrum	Quantity	Transformation
^{1}H NMR	shift / integration / couplings	$(\delta_H{+}1)/11$; $\log(1{+}n_H)/\log 41$; the four largest couplings in descending order, each divided by 20, zero-padded
^{1}H NMR	join fields†	carbon-membership flag in $\{0,1\}$; matched $\delta_C/230$
^{13}C NMR	shift	$\delta_C/230$
^{13}C NMR	join fields†	protonation flag in $\{0,1\}$; partner $n_H/3$; matched δ_H uses $(\delta_H+1)/11$
HSQC	proton shift / carbon shift / proton count	$(\delta_H+1)/11$; $\delta_C/230$; $\min(n_H,9)/3$
HSQC	merged-cross-peak flag	1 if $n_H > 3$ and $n_H \notin \{4,6,8,9\}$ (a merge of inequivalent carbons)
IR	intensity	negative values (a baseline artifact, at most about 2% of the maximum in magnitude) clipped at zero; 1,800 samples reshaped to 120×15 without spectral resampling
MS/MS	mass / collision-energy intensities	m/z divided by 1,000; intensities at 10, 20, and 40 eV (base peak 100 at each energy) divided by 100
MS/MS	intensity summary / precursor fraction	$0.2\log(1 + I_{\max})$ over the three energies; $\min(m/z/M_F,\, 2)$
MS/MS	neutral loss / parity / compositions	loss mass (capped at 999 Da) divided by 100; 1 for an odd nominal loss mass; counts of C, N, O, F, Cl, S, Br, I and P in fragment and loss transformed as $\log(1+n)/\log 41$

cross-peak statistics and cross-modal statistics among the 418 descriptors. Every TAG-XMM system result uses all five spectra, except the runs on the baselines' spectra (below), so the fields are backed by an observed HSQC. Whenever HSQC is withheld they are set to -1, the value the encoder was trained to read as unknown: in the encoder's spectrum-subset analysis (App. C), whose HSQC contribution therefore includes them, and in the run on NMIRacle's spectra. The inputs $\tilde{x}'$ of Graph-SSI's diffusion model have them set to -1 for every spectrum set.

Runs on the baselines' spectra. For the column of Table 1c in which TAG-XMM sees only the matched baseline's spectra, the frozen encoder encodes each molecule with the other spectra removed, and the decoder is retrained on these banks. A removed spectrum has all its slots marked as padding and its presence bits cleared, and a dimension of the 418 summary descriptors is kept only if every input it reads is kept. NMRViT's set (^{1}H, ^{13}C, HSQC and the formula) zeroes 92 of them. NMIRacle's set (^{1}H, ^{13}C, IR) also removes HSQC, so the join fields are set to -1; like NMIRacle, this run receives no formula: the encoder's formula input is emptied, the decoder's formula tokens are masked, formula-constrained decoding is off, and 161 summary dimensions are zeroed. The solver's and the verifier's views of removed spectra abstain. Without MS/MS, the parent-mass guard of App. E finds no parent mass and removes no candidate, and on NMIRacle's spectra no formula check replaces it: all 24,013 candidates of the 2,382 solver pools reach the solver. On 2,048 training molecules, a label-free check found no masking violation, and replacing the removed inputs by noise left the encoder's output unchanged. Graph-SSI's gate reads each feature of a removed input at its median over validation pools, and Graph-SSI was to run only if the gate's firing rate stayed within 0.8–1.25 times its all-five rate on the same molecules: it fired on 34.20% of NMRViT's 6,337 pooled molecules against 32.10% with all five (ratio 1.07; run) and on 50.50% of NMIRacle's 2,382 against 32.66% (1.55; not run). On NMRViT's spectra, Graph-SSI adds 0.82 points to the solver (95 gained, 32 lost, exact McNemar $p = 1.9 \times 10^{-8}$); in 84 of the 95 gains the truth is a graph it added. These runs were pre-registered; NMIRacle's formula-free arm was fixed by an addendum written before its decoder was trained.

Formula information. The molecular formula is assumed to be available from high-resolution mass spectrometry. It enters the encoder and decoder in four ways (Table 6); in addition, Graph-SSI's gate reads the presence of fluorine and phosphorus and the heavy-atom count, its diffusion model is conditioned on the formula F_{ref} of the top pool member y_1 (App. F), and encoder training groups molecules by formula group (Table 7). This assumption makes the task constitutional-isomer resolution given the composition. Hydrogen counts are not enforced, so a candidate's hydrogens may differ from F. The same formula is supplied to every TAG-XMM condition except the run on NMIRacle's spectra (above), which receives none; among the matched baselines, NMRViT also receives it, while NMIRacle, run from its released manifest, does not.

Table 6: Uses of the HRMS-derived molecular formula.

Location	Role
Encoder summary	A compact formula residual contributes to the global summary token alongside spectrum presence and engineered descriptors, some of which also read the formula's element counts.
MS/MS tokens	The precursor fraction divides each m/z by the monoisotopic mass of the formula; neutral-loss compositions subtract the annotated fragment from the formula plus one hydrogen.
Decoder memory	Twelve heavy-element tokens (C, N, O, S, P, F, Cl, Br, I, Si, B, Se; hydrogen has none) expose element budgets directly to cross-attention; absent elements are masked.
Beam search	Formula-budget masking prevents heavy-atom counts from being exceeded, and the end token is permitted only after the budget is exhausted (at the maximum length, ending early instead costs 2 nats per missing heavy atom).

C MULTIMODAL ENCODER

Architecture. Proton and carbon shifts are embedded by shared Fourier-feature encoders, whether they come from the ^{1}H or ^{13}C peak lists or from the two axes of an HSQC cross-peak. The ^{1}H, ^{13}C, HSQC, and MS/MS branches each use two induced set-attention blocks with 32 inducing points (Lee et al., 2019). The IR branch uses two self-attention blocks followed by attention pooling to 16 tokens (Lee et al., 2019). A dedicated correspondence stage exchanges information between HSQC and its ^{1}H and ^{13}C partners: the HSQC tokens cross-attend to the ^{1}H and to the ^{13}C tokens, which then attend back to the HSQC tokens. After concatenation with the summary token, two global self-attention layers produce a bank of width 256. An attention pool over the bank feeds the fingerprint head, and a pooled 768-dimensional molecular embedding is used only by the contrastive terms and diagnostics.

Chemical supervision. The encoder is trained on the training split with

$$\begin{aligned}\mathcal{L}_{\text{enc}} = {} & 1.00\,\mathcal{L}_{\text{C-env}} + 0.50\,\mathcal{L}_{\text{join}} + 0.50\,\mathcal{L}_{\text{corr}} + 0.50\,\mathcal{L}_{\text{formula-NCE}} \\ & + 0.10\,\mathcal{L}_{\text{global-NCE}} + 0.40\,\mathcal{L}_{\text{fp}} + 0.05\,\mathcal{L}_{\text{VICReg}}. \end{aligned} \tag{3}$$

The local terms preserve chemically interpretable environments and observed cross-modal joins; the molecular terms organize the fused representation around target identity. Table 7 gives the target and matching rule for every active term, and Table 8 the architecture and optimization settings.

Table 7: Active encoder objectives. The within-formula term and the batch sampler group molecules by the corpus's precomputed formula identifier, which is not one-to-one with the molecular formula (hence "formula group"): in the 4,095-molecule retrieval cohort, 1,223 formulas map to 3,094 identifiers, 1,153 formulas are split across several identifiers, and 186 identifiers mix formulas. Retrieval evaluation groups by the exact formula.

Supervision signal	Weight	Objective	Constraint
Carbon-environment inventory	1.00	Permutation-invariant cross-entropy over 12 environments with Sinkhorn assignment ($\varepsilon = 0.1$)	Retain local carbon chemistry without peak-to-atom labels
Masked C–H join recovery	0.50	Mask the join fields of 30% of ^{13}C tokens and of the ^{1}H tokens joined to them, and recover protonation, partner proton count, and matched proton shift	Reconstruct cross-modal context from masked fields; input columns cannot be copied
HSQC–NMR correspondence	0.50	Negative log-likelihood on HSQC$\rightarrow$ ^{1}H and HSQC$\rightarrow$ ^{13}C attention from one dedicated head	Preserve peak-level correspondence while leaving other heads unrestricted
Within-formula alignment	0.50	Symmetric InfoNCE (van den Oord et al., 2018; Radford et al., 2021) against a pretrained MoLFormer embedding (Ross et al., 2022), with negatives restricted to batch molecules of the same formula group (temperature 0.07)	Separate isomers within a formula group
Global molecular alignment	0.10	Symmetric InfoNCE against the same embedding over each device's batch	Maintain broad chemical organization of the pooled representation
Fingerprint prediction	0.40	Binary cross-entropy to a 2,048-bit radius-3 circular fingerprint (Rogers & Hahn, 2010)	Expose substructure information across the fused bank
Within-molecule token regularization	0.05	Variance–covariance penalty (Bardes et al., 2022) over each molecule's ^{13}C tokens	Prevent collapse and discourage redundant dimensions

Same-formula retrieval. The final encoder checkpoint is evaluated frozen on the 4,095 test molecules of Table 2: 1,223 formula families with at least two members, drawn whole and at random from the test split. For each spectrum set, the fingerprint head predicts 2,048 bit probabilities from the token bank, and the true fingerprints of all members of the molecule's family are ranked

Table 8: Encoder architecture and optimization.

Stage	Setting	Value
Architecture	Branch attention	two blocks; 32 inducing points for NMR, HSQC, and MS/MS; two self-attention blocks and 16 pooling queries for IR (Lee et al., 2019)
	Cross-modal / global mixing	one HSQC–NMR correspondence stage; two global self-attention layers
	Width / heads / dropout	256 / 8 / 0.1
	Bank / pooled embedding	at most 135 positions / 768 dimensions
Optimization	Optimizer	AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) $(\beta_1, \beta_2) = (0.9, 0.98)$; learning rate 3×10^{-4}; weight decay 10^{-4}; gradient clipping at 1.0
	Schedule / precision	2,000-step warmup, cosine decay, stepped per epoch, to 5% of peak; bfloat16 mixed precision
	Batch / duration	four devices, 1,024 molecules per device (effective batch 4,096), 25 epochs (4,350 steps), seed 0; each device batch draws 8 molecules from each of 128 formula groups (with replacement from smaller groups; groups of one are never drawn)

by cosine similarity to this prediction; top-1 counts the molecule's own fingerprint ranked first, with ties broken in a fixed order. The formula is present in every condition. A spectrum is removed by zeroing its features, masking its tokens and clearing its presence bits; removing HSQC also sets the join fields to -1 (App. B). The 418 summary descriptors mix spectra, so they are zeroed in every condition, including the one with all five spectra. Spectra are removed only at evaluation: the encoder was trained without spectrum dropout, so every condition with a spectrum removed is out of distribution for it. With the formula alone, all members of a family receive the same input and the same predicted fingerprint, so exactly one member per family ranks first and top-1 equals chance, $1{,}223/4{,}095 = 29.87\%$.

Shapley allocation. For molecule i in a family of n_i members, let $h_i(S) \in \{0, 1\}$ be its top-1 hit with spectrum set $S \subseteq \mathcal{M}$, where $\mathcal{M}$ is the set of five spectra, and let $g_i(S) = \big(h_i(S) - 1/n_i\big)/\big(1 - 1/n_i\big)$ be its chance-normalized gain. The Shapley value (Shapley, 1953) of spectrum m is

$$\Delta_m = \frac{1}{N} \sum_{i=1}^{N} \sum_{S \subseteq \mathcal{M} \setminus \{m\}} \frac{|S|!\,(4 - |S|)!}{5!} \big[g_i(S \cup \{m\}) - g_i(S)\big], \tag{4}$$

computed exactly from all 32 spectrum sets for each of the $N = 4{,}095$ molecules. Formula-only retrieval scores exactly one hit per family, so its mean gain is zero and the five values sum to the mean gain with all five spectra, 97.0 points: ^{1}H NMR contributes 46.0, HSQC 23.9, IR 13.7, MS/MS 8.7 and ^{13}C NMR 4.7 (Fig. 8). They measure what each spectrum adds to isomer retrieval through this encoder under the removal protocol above. The neutral-loss and element-count columns of the MS/MS tokens are computed from the fragments that the simulator annotates on each peak, that is, from the true structure (App. A, item (v)); the summary descriptors derived the same way are zeroed here, and we have not measured how much of the MS/MS value comes from these columns.

Late fusion. The late-fusion baseline encodes each spectrum alone with the same encoder, scores the members of a family by cosine similarity for each single-spectrum prediction, standardizes each query's scores, and averages them over the spectra of the set with one non-negative weight vector, fitted on validation with a pairwise logistic ranking loss; this rule had the best multimodal validation top-1 of four candidate rules and is applied unchanged to test. Because the encoder saw no withheld spectra in training (above), the single-spectrum encodings that late fusion combines are out of distribution for it, whereas the all-five joint input differs from training only in its zeroed summary descriptors: the baseline is late fusion of this encoder's single-spectrum outputs, not of encoders trained on one spectrum each. Against its best subset with one spectrum fewer, every multimodal set except ^{13}C + MS/MS gains, with a 95% interval that excludes zero.

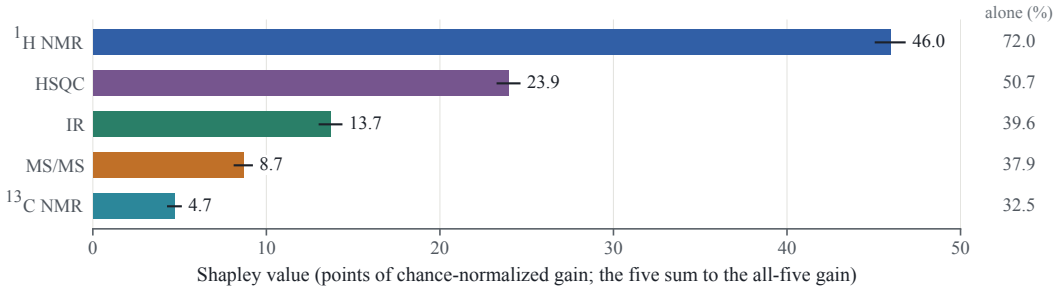


Figure 8: **How much each spectrum contributes to same-formula retrieval** (4,095 test molecules, 1,223 formula families). Bars: each spectrum's Shapley value, its average contribution to the chance-normalized gain over all orders in which the five spectra can be added; for a molecule in a family of n the gain is $(h - 1/n)/(1 - 1/n)$, with $h \in \{0, 1\}$ its top-1 hit, and the five values sum to the gain of all five spectra. Right: retrieval top-1 with that spectrum alone; chance, the formula alone, is 29.87%. A short bar means a spectrum adds little, on average, to the spectra added before it, not that it is uninformative. Whiskers: 95% formula-family bootstrap intervals (2,000 resamples of the 1,223 families); the right-hand values' intervals, at most ± 1.8 points, are not drawn.

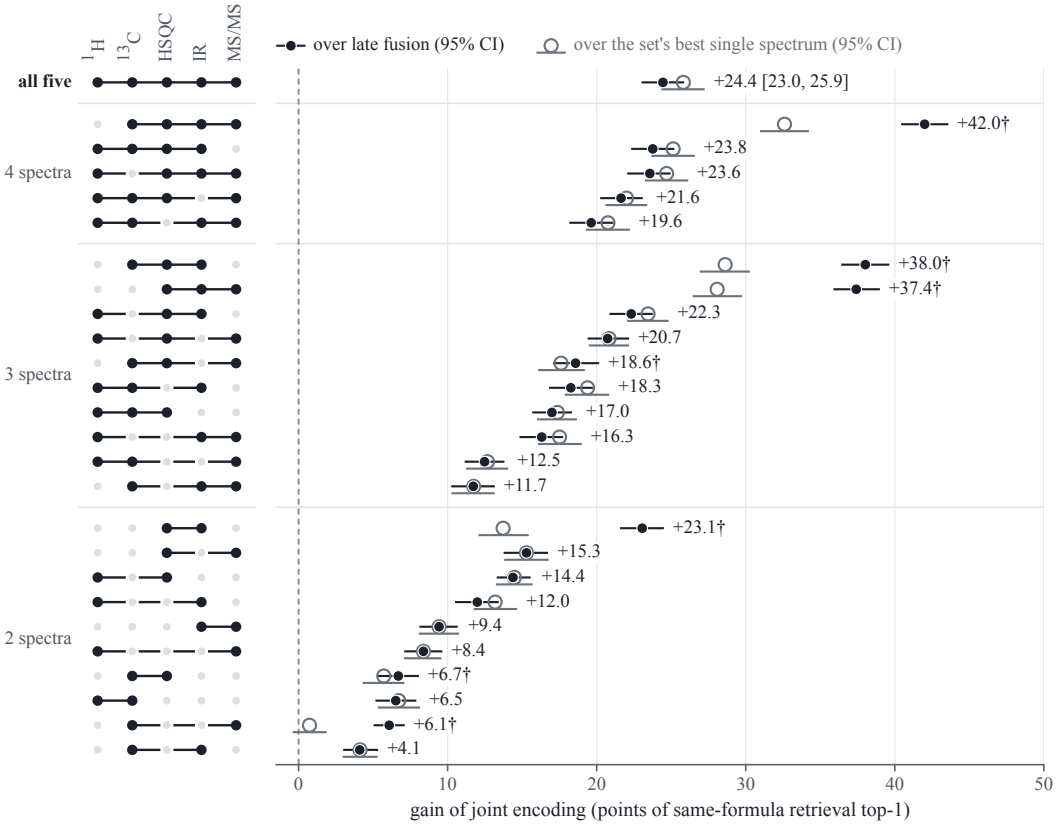


Figure 9: **Joint encoding beats late fusion on every set of two or more spectra** (4,095 test molecules). The top row gives the gain with all five spectra, 24.44 points (95% CI [23.01, 25.87]). Each row is one of the 26 multimodal spectrum sets (dots on the left mark its spectra). Filled: joint encoding minus late fusion of the same spectra; open: joint encoding minus the set's best single spectrum, its interval drawn just below the circle. The shared late-fusion weights are dominated by ^{1}H, so on sets without ^{1}H late fusion can fall below the set's best single spectrum (†; the open circle is the fairer comparison there). Intervals: 95% formula-family bootstrap (2,000 resamples of the 1,223 formula families, paired on the same molecules); only the open-circle interval of ^{13}C + MS/MS includes zero.

D DECODER AND RECONSTRUCTION ANALYSIS

Decoder. We retain the SMILES vocabulary, token and position embeddings, causal Transformer stack (Vaswani et al., 2017), cross-attention, and output head of a pretrained ECFP4-to-SMILES target-side decoder (ECFP4: Rogers & Hahn, 2010) from Ucak et al. (2023). Its six layers have width 512 (eight heads, feed-forward width 2,048), its vocabulary contains 109 tokens, and sequences are limited to 130 positions. A two-layer resampler (Jaegle et al., 2021) with 64 learned queries compresses the encoder's token bank into the cross-attention memory $\mathbf{M}$, to which up to 12 heavy-element formula tokens are appended. The resampler projects the bank tokens to width 512 and passes the queries through two post-norm cross-attention layers (eight heads, GELU feed-forward of width 2,048, dropout 0.1, padded bank positions masked); every query attends over the whole bank, so the 64 memory slots are not aligned with individual bank tokens and do not preserve their order. The formula enters as one token per heavy element present in F, over 12 element slots (C, N, O, S, P, F, Cl, Br, I, Si, B, Se); each token combines a learned element embedding with Fourier features of $\log(1+n_e)$, n_e being the element's count, so the decoder cross-attends to at most $64 + 12 = 76$ keys. With the encoder frozen at its final checkpoint (Table 8; the same checkpoint produces Graph-SSI's inputs $\tilde{x}'$), the resampler and the decoder are trained on the training split with teacher forcing, label smoothing 0.1 and an exponential moving average of the weights (decay 0.999), on five Kekulé SMILES spellings per molecule without stereochemistry (the canonical one and four from random atom orders) and an effective batch of 4,096, for 40,000 steps (seed 0, mixed precision); the checkpoint with the best in-run validation Exact@1 on 2,048 validation molecules, at step 24,000, is deployed. Neither the encoder nor the decoder was trained with spectra withheld: the decoder saw only banks of all five spectra. Without retraining, its output collapses when spectra are withheld: in a label-free check on 200 test molecules, withholding IR and MS/MS left its top-1 unchanged for 11 of 200 (17 had no valid candidate), and withholding HSQC and MS/MS for 1 of 200 (16).

At inference, formula-constrained beam search uses width 40 and ranks finished strings by the length-normalized score $\log p_\psi(y \mid X, F) \big/ \big(\frac{5+|y|}{6}\big)^{0.6}$ (Wu et al., 2016), where $|y|$ counts the start and end tokens. The formula budget masks tokens that would exceed the permitted heavy-atom counts and admits the end token only once every count is reached, which is well defined because every token is one SMILES symbol; at the maximum length, ending early instead costs 2 nats per missing heavy atom. Hydrogens are implicit in SMILES and are not constrained: on the 39,224 evaluated test molecules the top-1 string is a valid molecule for all of them and matches the full molecular formula, hydrogens included, for 98.64%; for 99.75%, at least one of the first ten distinct candidates does. This constraint is part of every reported decoding condition except the run on NMIRacle's spectra (App. B).

Candidate pool and decoder score. Unparseable strings are dropped and the rest are deduplicated in beam order into the pool $\mathcal{P}$ (App. A); a test pool holds 4.69 distinct molecules on average, and the pool order follows each molecule's best string. Because a molecule is spelled by many strings, the decoder score used by the inverse solver is instead the molecule marginal of Sec. 3,

$$\log p_k = \Big(\log \sum_{s \in \mathcal{S}_k} p_\psi(s \mid X, F)\Big) \Big/ \Big(\tfrac{5+\overline{|s|}_k}{6}\Big)^{0.6}, \tag{5}$$

where $\mathcal{S}_k$ holds the beam strings that canonicalize to y_k and $\overline{|s|}_k$ is their mean length. It has no fitted parameters and adds no spectral term beyond the decoder's own conditioning. On all 39,224 test molecules, the inverse solver run with the decoder score as its only view reaches 78.00% Exact@1 with the best-string score and 81.74% with Eq. (5) (Table 1a).

Hierarchical error analysis. Exact match hides whether a failure has the wrong composition, functional groups, scaffold, or only the final bonding. We therefore score each pool at increasingly discriminative chemical resolutions. The metric families are defined in Table 9. At rank k, continuous criteria use the best value among the first k distinct candidates, while exact criteria ask whether any of those candidates satisfies the criterion. Each level is scored independently, so the candidate that satisfies one level need not satisfy another.

Results. Table 10 gives the levels of Fig. 4 with their denominators and graded scores. Radius-2 atom environments are almost as strict as the structure itself (78.49% against 77.88% at top-1): once

Table 9: Hierarchy used to localize reconstruction errors. Figure 4 reports the exact-match rate at four levels: functional groups (on the 38,673 test molecules with at least one group), scaffold bonds (on the 38,059 with a ring), atom environments and exact structure; the remaining criteria are defined for completeness and not reported. Functional groups are counts of 24 fixed SMARTS patterns; atom environments are the multiset of Morgan identifiers emitted at exactly radius r; scaffolds are Bemis–Murcko frameworks, scored on molecules with a ring.

Level	Question	Exact criteria	Graded criteria
L0	Is the output chemically parseable?	molecular validity	–
L1	Is the composition correct?	molecular formula (hydrogens included) and degree-of-unsaturation (rings + π bonds) match	element-count similarity
L2	Are local environments correct?	non-empty functional-group inventory match; radius-2 atom-environment F1 $= 1$	count-aware functional-group F1; radius-1/2 atom-environment F1
L3	Is the scaffold correct?	typed and generic cyclic Murcko (Bemis & Murcko, 1996) scaffold match; scaffold-bond F1 $= 1$ (molecules with a ring)	scaffold-bond F1; weighted and attachment-aware BRICS (Degen et al., 2008) F1
L4	Is the global topology correct?	ring-system match	ring-system bond F1; ECFP4 (Rogers & Hahn, 2010) Tanimoto; MCES distance and similarity
L5	Is the molecule recovered?	exact structure: canonical non-stereo SMILES identity	–

every two-bond neighborhood is right, the molecule is nearly always right. Because pools average 4.69 distinct molecules, top-10 is close to the recall of the whole beam, 93.87% of the 39,224 molecules; counting raw strings, the truth is among the first 10 of the 40 for 92.19% and among the first 20 for 93.31%, so the last 20 strings add 0.56 points. Beyond the label-free check above, we did not measure the effect of removing individual spectra on the deployed decoder on the test split, and therefore report no per-spectrum decoder ablation; per-spectrum contributions are reported for the encoder (App. C, Fig. 8).

Table 10: Reconstruction by level on the full test split (deployed decoder, formula-constrained beam 40). *Recovered*: share of molecules for which the top candidate (top-1) or the best of the first ten distinct candidates (top-10) scores 1 on the level. *Mean F1*: the level's graded score (best over the same candidates), averaged over molecules. L0 is 100% by construction at top-10, since unparseable strings never enter the pool; at top-1 it records that the first beam string always parsed. n excludes molecules without any listed functional group (L2, groups) or without a ring (L3). For the rows from functional groups to exact structure, 95% confidence intervals lie within ± 0.41 points of every entry.

			Recovered (%)		Mean F1 (%)	
Level	Criterion	n	top-1	top-10	top-1	top-10
L0	valid molecule	39,224	100.00	100.00	–	–
L1	molecular formula (with H)	39,224	98.64	99.75	–	–
L2	functional-group inventory	38,673	96.35	98.77	99.02	99.76
L2	atom environments ($r = 2$)	39,224	78.49	93.74	90.63	97.40
L3	scaffold bonds	38,059	85.33	95.61	93.42	98.27
L5	exact structure	39,224	77.88	93.68	–	–

E INVERSE SOLVER

The inverse solver reranks the augmented pool $\mathcal{P}^+ = \mathcal{P} \cup \mathcal{G}_K$: the decoder's deduplicated pool plus the (at most $K = 4$) graphs that Graph-SSI adds (App. F). For candidate y_k and spectral view v, a forward model f_v predicts $f_v(y_k)$: for MS/MS, an enumeration of bond-breaking fragments; for the four in-domain energies, the in-domain shift heads, predictors of ^{13}C and ^{1}H shifts trained on molecules of this corpus (App. A). Both heads are message-passing networks over the heavy-atom graph, trained on the corpus's simulated spectra without atom-level labels, by a minimum-cost assignment of their predictions to the observed peaks. The in-domain ^{13}C energy reads the observed line list as it is, so the extra lines of C–F and C–P multiplets count as unmatched; the DP4-style likelihood expands each carbon orbit into its multiplet. Graphormer-IR is fine-tuned on the first 200,000 training molecules, with the epoch-28 checkpoint chosen on 83 validation pools. A view-specific comparison with the observed spectrum gives the discrepancy $d_{v,k}$ (lower is better). Spectral discrepancies are standardized within $\mathcal{P}_v^+$, the members of $\mathcal{P}^+$ that view v can score (the decoder score is not standardized), and combined as in Eq. (2):

$$z_{v,k} = \frac{d_{v,k} - \operatorname{mean}_{\mathcal{P}_v^+} d_v}{\operatorname{sd}_{\mathcal{P}_v^+} d_v}, \qquad S_k = \sum_{v \in \mathcal{V}_k} \alpha_v \, z_{v,k} - \alpha_{\mathrm{dec}} \log p_k, \qquad \hat{y} = \operatorname*{arg\,min}_{y_k \in \mathcal{P}^+} S_k, \tag{6}$$

where $\mathcal{V}_k$ is the set of spectral views that can score y_k and p_k is the molecule marginal of Eq. (5), the summed probability of y_k's beam strings, length-normalized as in the beam (for an added graph, its teacher-forced log-probability mapped onto this scale by an isotonic fit on validation, App. F). The standard deviation is the population value. A view that cannot score a candidate contributes zero, as does, for a whole pool, a view with fewer than two scored candidates or with identical values; the other weights are not renormalized. A parent-mass guard first removes from $\mathcal{P}^+$ the candidates whose mass matches no parent mass read from the observed MS/MS (the top peak of the lowest-energy spectrum, as $[M+H]^+$ or $M^{+\bullet}$) and reverts to the full pool if it would remove every candidate; no candidate is removed for a poor spectral fit. For added graphs, NMRNet's conformers are embedded with at most 30 ETKDG iterations instead of 200, and its ^{13}C and ^{1}H views abstain for a graph that needs more: they score 47,823 of the 51,612 added graphs (92.7%), and every other view scores all of them. The weights (Table 12) are fitted once on validation pools without added graphs, by maximizing a ridge-penalized conditional-logit likelihood of the truth (McFadden, 1974), with the penalty strength λ chosen from $\{0, 0.003, 0.01, 0.03, 0.1, 0.3, 1\}$ by the Exact@1 of an inner five-fold cross-validation over pools ($\lambda = 0.003$ for the inverse solver), and then frozen. The fit uses the 32,696 validation molecules with a solver pool (of 39,276 evaluated; the other 6,580 have none, by the same stereo-aware matching as on test). The six-view solver weights the decoder's best-string log-probability by 2.32 and its five external views by 0.99 (^{13}C), 0.90 (IR), 0.47 (MS/MS), 0.34 (HSQC) and 0.23 (^{1}H).

Table 11: **Inverse-solver build-up on test.** Exact@1 (%) on all 39,224 test molecules (All) and on the 37,142 of them outside the in-domain shift heads' training data (Leak-free). Each fitted solver is fitted once on validation and applied unchanged; the two decoder rows are not fitted, and the last row sets $\alpha_{\mathrm{dec}} = 0$ in the inverse solver's weights without refitting. Δ: points against the row above, computed before rounding; for a control, against the solver without the controlled views; for the variant, against the inverse solver. Controls feed the named views another molecule's spectra.

Solver	All	Δ	Leak-free
Decoder, best string	78.00	–	76.85
Decoder, molecule marginal	81.74	–	–
Six-view solver	85.61	–	84.88
+ molecule marginal	86.80	+1.19	86.13
+ in-domain ^{13}C and HSQC energies	89.22	+2.42	88.68
+ DP4-style ^{13}C likelihood and HSQC register (inverse solver)	90.03	+0.81	89.54
Control: in-domain ^{13}C and HSQC energies	86.80	0.00	86.13
Control: DP4-style likelihood and HSQC register	89.21	−0.01	88.67
Variant without the in-domain ^{13}C and HSQC energies	90.04	+0.01	89.55
Inverse solver, spectral views only ($\alpha_{\mathrm{dec}} = 0$)	88.56	–	–

Table 11 gives the build-up on test. The six-view solver had been evaluated on test before this work; the new solver arms, their controls and the primary comparisons were pre-registered before any of

them was run on test, and the DP4-style likelihood, the register, the variant below and its control were added in an addendum written before their test run. The last step alone gains 663 molecules and loses 345 ($p = 8 \times 10^{-24}$). The controls feed the added views the spectra of another test molecule with the same number of ^{13}C lines (for the in-domain HSQC energy, the same number of cross-peaks); the views then add nothing (0.00 and -0.01 points, the latter 72 gained and 76 lost, $p = 0.81$), so the gain comes from pairing each candidate with its own spectra. The step gains of Sec. 4 depend on the order in which views are added, because the in-domain energies share their shift heads and overlap: adding the DP4-style likelihood and the register lowers the weight of the in-domain ^{13}C energy from 1.08 and that of the in-domain HSQC energy from 0.86 to their values in Table 12, and a pre-registered variant without the in-domain ^{13}C and HSQC energies reaches 90.04%. We keep the inverse solver, the pre-registered primary arm. The solver cannot recover a structure absent from its pool: the decoder's pool contains the truth for 93.80% (36,794) of the 39,224 test molecules, and the pool augmented by Graph-SSI for at most 95.63% (37,511).

Table 12: Inverse-solver views (ten), candidate-side predictions, comparison functions and weights α (fitted once on validation pools without added graphs). Spectral views are standardized within $\mathcal{P}_v^+$, the members of $\mathcal{P}^+$ each view can score; the decoder score, with weight α_{dec}, is not.

View	Candidate-side prediction	Comparison with the observed spectrum	α
External forward models			
^{13}C NMR (NMRNet, Xu et al., 2025)	Predicted carbon shifts averaged over symmetry-equivalent carbons (graph-automorphism orbits), which give a single peak (topological symmetry; diastereotopic carbons treated as equivalent)	One-dimensional minimum-cost assignment (Kuhn, 1955) with unmatched-peak penalty	0.48
^{1}H NMR (NMRNet, Xu et al., 2025)	Per-hydrogen shift predictions	Observed peaks expanded by integral, then one-dimensional minimum-cost assignment (Kuhn, 1955) with unmatched-peak penalty	0.18
HSQC (TransPeakNet, Li et al., 2025; 2D checkpoint, solvent input $CDCl_3$)	Predicted two-dimensional cross-peaks	Minimum-cost two-axis assignment (Kuhn, 1955) after scaling proton and carbon differences by 0.10 and 1.50 ppm	0.18
IR (Graphormer-IR, Stienstra et al., 2024)	Predicted intensity trace (checkpoint fine-tuned on training spectra)	Spectral information divergence (Chang, 2000; McGill et al., 2021) on 400–2,000 cm^{-1} after non-negativity and baseline treatment	0.80
MS/MS (fragment enumeration)	Masses reachable by breaking up to two acyclic single bonds	Coverage of the observed peaks of the 10, 20 and 40 eV spectra within a mass error of $\max(20\,\mathrm{ppm}, 0.005\,\mathrm{Da})$	0.34
In-domain energies (in-domain ^{13}C and ^{1}H shift heads)			
DP4-style ^{13}C likelihood	Predicted carbon shifts, with the splitting by fluorine and phosphorus	DP4-style likelihood (Smith & Goodman, 2010) of the best one-to-one assignment of predicted to observed ^{13}C lines (Student-t errors with scales by carbon class), with costs fitted on validation for missing and extra lines	0.91
HSQC register	Predicted carbon and proton shifts of each protonated carbon; carbon shifts of the others	One assignment (Crouse, 2016) of protonated carbons to HSQC cross-peaks and of the other carbons to the ^{13}C peaks no cross-peak claims (Student-t shift errors; fitted costs for split, merged, missing and spurious peaks; proton counts not used)	0.75
In-domain HSQC energy (HSQC-n_{H})	Predicted carbon and proton shifts and the proton count of each protonated carbon	Minimum-cost assignment to the HSQC cross-peaks with cost $\lvert\Delta\delta_{\mathrm{C}}\rvert + \lvert\Delta\delta_{\mathrm{H}}\rvert/0.35 + 2\lvert\Delta n_{\mathrm{H}}\rvert$ (shifts in ppm) and 15 per unmatched orbit or peak, divided by the larger count	0.39
In-domain ^{13}C energy (E_{res})	Predicted carbon shifts averaged over symmetry-equivalent carbons	Minimum-cost one-to-one assignment (Crouse, 2016) of orbits to the observed ^{13}C lines: summed $\lvert\Delta\delta_{\mathrm{C}}\rvert$ plus 15 ppm per unmatched orbit or line, divided by the larger count	0.25
Decoder			
Decoder score	Molecule-marginal log-probability (summed over the molecule's beam strings, length-normalized; added graphs: teacher-forced log-probability mapped by an isotonic fit on validation)	Negative log-probability, entered raw (not standardized)	1.77

F GRAPH-SSI DETAILS

Graph-SSI reads no label, and its components were frozen before Test A, except the decoder score of added graphs, which was changed for Test A2 (*Scoring added graphs*). All counts are on all 39,224 test molecules. Counts selected by the inverse solver come from Test A2; those selected by the six-view solver come from Test A rescored with the isotonic map, the pre-registered sensitivity row whose rule Test A2 adopted (under Test A's pre-registered rule: 343 with both generators, 260 with diffusion alone and 198 for the control; App. A).

Gate. A histogram-based gradient-boosted classifier (Friedman, 2001; the scikit-learn implementation, Pedregosa et al., 2011), fitted on validation from 43 label-free features of the pool (decoder confidence and how well the pool fits each spectrum), estimates the probability that the truth is missing from the pool; a molecule is sent to repair when this estimate is at least 0.01. It is fitted on 11,126 validation pools (all 1,981 out-of-beam, all 5,145 in-beam with a wrong top-1, and 4,000 sampled from those with a right top-1, weighted to their population share). The threshold was chosen on validation together with the parameters of an earlier rule that replaced the decoder's top-1 by one repaired graph; the deployed stage adds graphs instead and reuses the threshold unchanged. A pool that is not sent passes on unchanged, because every added graph can displace a correct answer; the threshold is low, so the gate also sends many in-beam pools. On test the gate sends 34.9% of pools to repair: 98.0% of out-of-beam pools (2,358 of 2,406), 84.1% of in-beam pools whose decoder top-1 (best string) is wrong (5,235 of 6,225) and 20.0% of those whose decoder top-1 is right (6,112 of 30,593).

Alignment and regions. Each pool member is written as a Kekulé heavy-atom graph and mapped onto the top member y_1 by an element-preserving atom map (a linear assignment between same-element atoms on local bond environments, refined by pairwise exchanges; not a maximum common substructure); a pool member yields a graph only if it is connected and decodes back to its own canonical SMILES. A member votes, i.e. enters τ, if its heavy atoms match those of y_1; its hydrogens may differ. A molecule is repaired only if at least two members, y_1 included, vote; diffusion also needs an observed ^{13}C spectrum and spectral tokens. The disagreement τ of Eq. (1) is computed over all atom pairs, order 0 meaning no bond. Members are kekulized independently, so an aromatic ring shared by all members can still show disagreement where their Kekulé forms differ. In each of up to four top members, the anchors are those of the 12 atom pairs with the highest τ that are bonds of the member; diffusion pairs each anchor (u, s) with up to eight disjoint bonds (v, w) ranked by $\tau_{vw} + \max(\tau_{uv} + \tau_{sw}, \tau_{uw} + \tau_{sv})$, and bond swaps pair it with every disjoint bond of equal order; an anchor and a partner define a region R of four atoms (six atom pairs); its atom states and bond orders form G_R, and the rest of the graph, $G_{\setminus R}$, is fixed. A double region joins two disjoint regions of the same member. A repair can therefore only reach a truth that matches the edited member outside its regions.

Generators. Two generators edit the same regions, and their proposals are pooled. Bond swaps are degree-preserving double-edge swaps (Maslov & Sneppen, 2002) of two equal-order bonds of a region, up to 1,200 proposals per molecule; they keep every atom's degree and hydrogen count, and undecodable graphs, radicals, duplicates and pool members are dropped. Spectrum-conditioned diffusion is an absorbing-state discrete diffusion model (Austin et al., 2021) on the molecular graph (Vignac et al., 2023), $p_\theta(G_R \mid G_{\setminus R}, \tilde{x}', F_{\mathrm{ref}})$, a graph transformer (eight blocks of width 384, eight heads, 20.9M parameters; atom self-attention biased by pair features, and cross-attention to $\tilde{x}'$) with $T = 32$ absorbing steps, trained on 75,785 training-split molecules with regions drawn at random on their true graphs, never on decoder outputs, and the spectra admitted all together (probability 0.5), each independently (0.3) or not at all (0.2). The deployed checkpoint is the last step (58,251) of this run, added to the pre-registered checkpoint rule by a dated amendment written before any of its validation-pool metrics was computed; placing training regions where the decoder errs gave far fewer valid samples on validation. It reads $\tilde{x}'$, the outputs of the encoder's five spectrum branches before the spectra exchange information, exported from a wider build of the same corpus (25, 48, 29, 16 and 80 slots for ^{1}H, ^{13}C, HSQC, IR and MS/MS; the encoder was trained with at most 24 ^{13}C and 40 MS/MS events), and the formula F_{ref} of y_1, whose hydrogen count can differ from F. It unmasks the region in random order and can change degrees and hydrogen counts; an invalid sample is projected onto the nearest valid graph by a mixed-integer program that keeps $G_{\setminus R}$ fixed and stops

after 5 s; a sample it cannot repair is dropped, and valid samples and exact decodes are not projected. In addition, an exact decode scores all 4^6 bond-order combinations of a region under p_θ and keeps the best two; it is part of the deployed diffusion arm. An exact decode keeps the edited member's hydrogen count, so one from a member whose hydrogen count differs from that of F_{ref} is invalid and dropped. Each single region gets four samples and two exact decodes and each double region eight samples (at most 48, 24, 16 and 12 single regions and 8, 4, 2 and 2 double regions in the four members), so a molecule gets at most 728 draws; valid draws not in $\mathcal{P}$ are kept up to 1,200.

Samplers not deployed. Three spectral schedules inside the sampler (a learned per-spectrum logit schedule with coarse-to-fine, atoms-before-bonds unmasking; tuned spectral guidance (Ho & Salimans, 2022); and learned resampling of particles) failed their pre-registered validation tests; the first improved held-out likelihood slightly, but not recovery. The deployed sampler unmasks in random order without resampling.

Shortlist and verifier. The ^{13}C shortlist ranks the pooled proposals by the fit of their predicted ^{13}C lines (in-domain ^{13}C head, including the splitting by fluorine and phosphorus that proton decoupling does not remove) to the observed ^{13}C spectrum and keeps 24. The verifier then gives each shortlisted proposal and each pool member j a linear score $\beta^\top \phi_j$ over nine scores: the decoder's teacher-forced log-probability of the candidate's canonical Kekulé SMILES and eight mismatch scores (lower is better) between the observed spectra and predictions from the in-domain ^{13}C and ^{1}H heads and an IR forward model: (1) a minimum-cost assignment of the observed ^{13}C lines to the predicted carbons, each expanded into its C–F and C–P multiplet (the shortlist's score); (2) one assignment of the protonated carbons to the HSQC cross-peaks and of the other carbons to the ^{13}C lines that no cross-peak claims; (3) an assignment of the protonated carbons to the HSQC cross-peaks on carbon and proton shifts; (4)–(6) assignments of the observed ^{1}H peaks on shift, on shift and proton count, and on proton count and the number of coupling partners implied by each peak's multiplicity (the $n + 1$ rule); (7) a spectral-similarity distance between the predicted and observed IR traces; and (8) the share of observed MS/MS intensity that no fragment from breaking up to two acyclic single bonds explains. Each enters as its value standardized over these candidates and as its gap to the best candidate. β is trained on validation swap proposals (decoys on molecules whose truth is in the pool come from a smaller swap budget than deployed) with a population-weighted, L2-penalized softmax loss that ranks the truth first (a conditional logit, McFadden, 1974); no input identifies the generator, so β applies unchanged to diffusion proposals. The $K = 4$ highest-scoring proposals, $\mathcal{G}_K$, are added to the pool; no member is removed. $K = 4$ was fixed on validation before the test runs.

Scoring added graphs. An added graph has no beam strings, so its decoder score is its teacher-forced log-probability (the best length-normalized value over its canonical Kekulé spelling and two spellings from random atom orders) mapped onto the molecule-marginal scale by an increasing isotonic regression fitted on the validation pools that the gate sends to repair in-sample (of its 11,126 fitting molecules) and clipped to the fitted range; the length-normalized form was preferred to the total by its validation Spearman correlation with the molecule marginal. Pool members keep their own scores. Test A's pre-registered primary rule instead shifted each teacher-forced score by the pool's median offset between beam and teacher-forced scores; Test A2 adopted the isotonic map, one of Test A's pre-registered sensitivity rows, after that analysis (App. A). Through the inverse solver, the offset rule also improves Exact@1 (665 gained, 384 lost; $p = 3 \times 10^{-18}$; 90.75%) but displaces more in-beam answers than the isotonic map (384 vs 198 lost). For the six-view solver the same construction maps onto its best-string score. Graphs are added to 13,173 of the 13,705 pools sent to repair; the solver's feature builder kept 51,612 of the 52,689 added graphs, in Test A as in Test A2. All 717 added out-of-beam truths are among them. In Test A, one added graph without a C–H bond had an empty predicted HSQC and, by an amendment written before any Test A metric was computed, counts as a missing HSQC view; the solver's candidate filter removes this graph, so the amendment changed no solver pool, and it is not among the 51,612 graphs scored in Test A2.

Control and error classes. The other-molecule control reruns diffusion conditioned on another test molecule's encoded spectra; the regions, F_{ref}, the shortlist, the verifier and the solver keep the molecule's own. The donor is chosen without labels: a different test molecule whose top pool member has the same heavy-atom formula, else the same heavy-atom count, else any. Error classes

compare each out-of-beam truth with the six-view solver's answer without Graph-SSI: *one swap* or *more than two swaps* apart (double-edge swaps through valid graphs), or *hydrogen moves* (same heavy atoms and formula, hydrogens on different atoms). Graph-SSI generates 70.1% (483 of 689) of the one-swap truths and 12.6% (71 of 562) of those more than two swaps away. Bond swaps keep each member's hydrogen counts, so they reach a hydrogen move only from a member whose hydrogens already sit as in the truth. Neither generator reaches a truth whose formula differs from the edited member's in its hydrogens: a diffusion proposal has the formula F_{ref}, and a swap keeps its member's.

Results. Table 13 follows the out-of-beam truths through each stage. The share of out-of-beam molecules ranked first falls with size, from 30% at 11 heavy atoms (95% band 21.5–39.5%; 12 out-of-beam molecules at that size) to 17% at 35 (13.1–21.2%; 172 molecules), on the fitted curve of Fig. 10; without Graph-SSI it is zero. Added graphs also change the within-pool standardization of every spectral view, so an in-beam answer can change even when no added graph is ranked first.

Table 13: **Graph-SSI on the 2,406 out-of-beam test molecules.** Generated counts are on the 2,358 molecules sent to repair; a molecule counts when its truth is among the proposals (generated), kept by the ^{13}C shortlist (shortlisted), added to the pool by the verifier (added) or ranked first (selected). Equal budget: each generator's first n valid proposals, with n the smaller of the two generators' counts for that molecule. p: exact McNemar test on the paired molecules.

Stage or arm	Molecules	p
Out-of-beam molecules	2,406	
Sent to repair by the gate	2,358	
Generated, deployed budgets		
Bond swaps (up to 1,200 proposals)	626	
Spectrum-conditioned diffusion	575	
Both, pooled (deployed)	767	7×10^{-43} [a]
Diffusion, other-molecule control	448	1×10^{-26} [b]
Diffusion / bond swaps, on the 398 hydrogen moves	43 / 17	
Generated, equal budget per molecule (median 147 proposals)		
Diffusion	571	6×10^{-28} [a]
Bond swaps	366	
Deployed funnel (both generators)		
Shortlisted (^{13}C shortlist keeps 24)	728	
Added (verifier, $K = 4$)	717	
Selected by the inverse solver	519	
Selected by the six-view solver (Test A, isotonic map)	320	
Diffusion alone, selected by the six-view solver (Test A, isotonic map)		
Own spectra	249	
Other-molecule spectra (control)	175	
Graph-SSI vs. inverse solver alone: gained / lost		
In-beam molecules	135 / 198	6.6×10^{-4}
All test molecules	654 / 198	1.2×10^{-57}

[a] vs. bond swaps in the same block; [b] vs. diffusion with the molecule's own spectra. Swap and pooled counts, and the comparisons that involve swaps, use the swap list regenerated for this analysis, whose top 48 differ from the deployed list on 62 of the 13,705 pools sent to repair; on the other 2,348 out-of-beam molecules, swaps generate 622 truths, both generators 762, and, at equal budget, diffusion 567 against 363 for swaps.

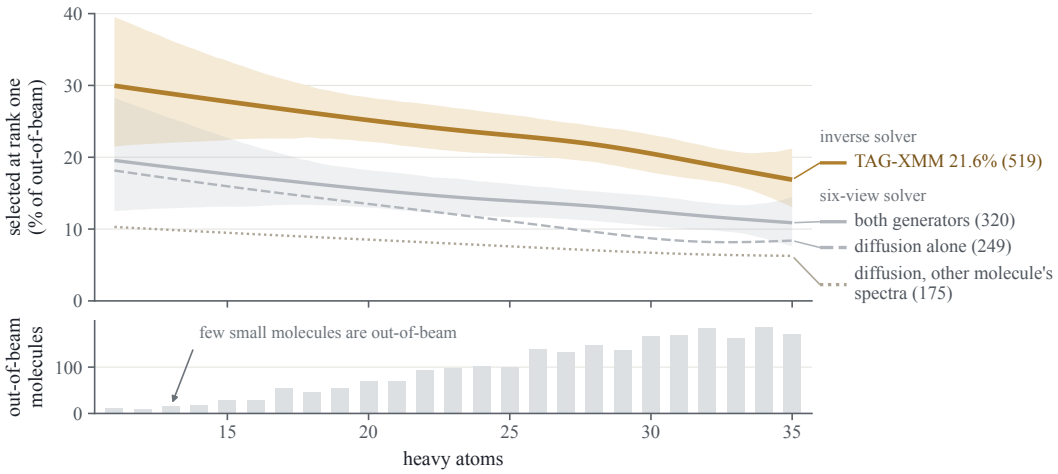


Figure 10: **Out-of-beam molecules selected at rank one, against molecule size** (the 2,406 test molecules whose truth is not in the beam; the decoder and the inverse solver alone select none). Gold: TAG-XMM, 519 of 2,406 (21.6%), from 30% at 11 heavy atoms to 17% at 35, with a 95% bootstrap band. Grey, all with the six-view solver (Test A, isotonic-map rescoring): Graph-SSI with both generators (320), diffusion alone (249) and diffusion conditioned on another molecule's spectra (175, control). Logistic-spline fits (bands for TAG-XMM and both generators: 95% bootstrap, 2,000 resamples); the strip below counts out-of-beam molecules per size, from 11 heavy atoms (the few smaller ones count in the totals). Few are small, hence the wide bands; the rate falls with size, but the share of molecules that are out-of-beam rises faster, so Graph-SSI's gain grows with size (Fig. 7).