ANCHORA: RETRIEVAL-GUIDED AEMS COMPLETION FOR MOLECULAR STRUCTURE ELUCIDATION FROM TANDEM MASS SPECTRA

Anonymous authors
Paper under double-blind review

ABSTRACT

Inferring molecular structure from tandem mass spectra is difficult because isomers can share same formula and similar fragments. Structure-database retrieval is applicable when the target or a close analogue is covered, but depends strongly on library coverage and cannot directly recover an out-of-library target. De novo generation can propose unseen molecules, yet constraining an entire molecular graph from sparse spectral evidence creates a large search space and can cause structural drift. To combine retrieval reliability with generative reach, we introduce ANCHORA, a retrieval-augmented generation framework based on an anchor-extended molecular scaffold (AEMS), which explicitly encodes side-chain anchoring points. Given a spectrum and formula, ANCHORA ranks formula-compatible complete molecules, extracts AEMS from the top reference, and represents it as a canonical AEMS–SAFE prefix for side-chain completion by a pretrained discrete diffusion model. Large-scale completion pretraining and spectrum–decoder adaptation support generation, while three ICEBERG-guided refinement rounds edit complete molecules without preserving the prefix, providing a spectrum-guided correction route beyond an erroneous prior. Reference-guided top-1 exact accuracy reaches 38.98% on CANOPUS and 23.50% on MassSpecGym, exceeding the best baseline by 13.95 and 5.21 points, respectively. Zero-shot evaluation further shows that Anchora-Base improves top-1 accuracy over FRIGID-base by 5.10–27.14 percentage points, supporting robustness to distribution shifts. These results prove that ANCHORA bridges chemical-structure retrieval and de novo molecular generation for MS/MS structure elucidation.

1 INTRODUCTION

Tandem mass spectrometry (MS/MS) supports metabolomics, environmental analysis, natural-product discovery, and chemical annotation (Kind et al., 2018). Reference collections such as MassBank and GNPS are widely used (Horai et al., 2010; Wang et al., 2016), yet incomplete, instrument-dependent fragmentation and isomeric ambiguity still hinder exact molecular-graph recovery. Structure elucidation must reconcile sparse spectral evidence with global chemical validity.

Retrieval-based methods rank known molecules from structure databases such as PubChem, ChEBI, and ChemSpider (Kim et al., 2016; Hastings et al., 2016; Pence & Williams, 2010). It is informative when the target or a close analogue is indexed, but cannot return structures outside its candidates. De novo generative models can propose unseen molecules (Shrivastava et al., 2021; Butler et al., 2023; Litsa et al., 2021; Stravs et al., 2022; Wang et al., 2025b; Bohde et al., 2025; 2026), yet reconstructing an entire molecule structure from sparse peaks creates a much larger search space and struggling to generate accurate structure.

In order to reduce the search space of generative models, recent methods make attempts to constrain generation with structural priors. MADGEN fixes a *retrieved* Murcko scaffold (Wang et al., 2025b), while MSAnchor fixes a *predicted* anchor-extended molecular scaffold (AEMS) with explicit side-chain attachment sites (Qin et al., 2026). However, two challenges remain, **(1) Selecting informative priors.** An informative prior should capture both the scaffold and its attachment sites. Matching a spectrum only to reduced scaffolds can discard substituent evidence that helps distinguish candidate

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.

priors. **(2) Keeping the generation flexible.** An inferred scaffold or attachment pattern can be wrong. Keeping it fixed throughout generation excludes the target whenever that prior is incompatible, even if the remaining structure is generated correctly.

Addressing the challenges, we introduce Anchora, which selects references at the full-molecule level and uses their AEMS to guide initial generation. Ranking formula-compatible complete molecules retains both scaffold and substituent information before the selected reference is reduced to a scaffold with explicit attachment sites.

Anchora completes the missing side chains through a reversible AEMS-SAFE representation that preserves attachment correspondence. Subsequent spectrum-guided refinement permits edits to entire candidate molecules, allowing exploration beyond the initial scaffold. Together, these stages use the retrieved prior to focus the initial search while keeping later molecular edits flexible.

Our main contributions are summarized as follows:

- **Retrieval-guided anchored completion.** We introduce Anchora, which selects AEMS priors through complete-molecule retrieval, retaining both scaffold and substituent evidence. It combines anchored completion with full-molecule refinement that permits edits beyond the fixed prefix.
- **Attachment-consistent completion learning.** We develop a reversible AEMS–SAFE representation that preserves attachment correspondence through scaffold canonicalization. This enables completion-only diffusion training with large-scale pretraining and joint spectrum–decoder adaptation.
- **Benchmark and transfer evaluation.** With retrieved priors, Anchora achieves 38.98% and 23.50% top-1 accuracy on CANOPUS and MassSpecGym, respectively. Anchora-Base also outperforms FRIGID-base in zero-shot evaluation across five external spectral libraries.

2 Related Work

2.1 Molecular structure elucidation from MS/MS

Database-search pipelines rank enumerated molecules using in silico fragmentation (Wolf et al., 2010; Ruttkies et al., 2016; Allen et al., 2014; Djoumbou-Feunang et al., 2019), fragmentation trees (Böcker & Dührkop, 2016; Dührkop et al., 2019), or predicted fingerprints (Heinonen et al., 2012; Dührkop et al., 2015). MIST predicts fingerprints with formula-aware transformers (Goldman et al., 2023), whereas CSU-MS2 learns spectrum–structure embeddings (Xie et al., 2025). These methods efficiently exploit known chemistry, but their outputs remain library-bound. Direct spectrum-to-scaffold matching additionally reduces structural evidence before ranking. Anchora instead ranks complete molecules, extracts AEMS only after selection, and generates the unresolved structure.

Learned comparison signals include spectrum embeddings (Huber et al., 2021a;b), structure-to-spectrum predictors (Wei et al., 2019; Zhu et al., 2020; Young et al., 2024; Li et al., 2024), and joint spectrum–molecule embeddings (Kalia et al., 2025). They improve similarity or forward prediction but do not remove library bounds by themselves. ICEBERG models chemically structured fragmentation and supplies the simulated spectra used for Anchora's refinement (Wang et al., 2025a).

Generative methods can leave an indexed collection: MassGenie, MS2Mol, and Spec2Mol map spectra to molecular strings (Shrivastava et al., 2021; Butler et al., 2023; Litsa et al., 2021); MSNovelist decodes predicted fingerprints under formula constraints (Stravs et al., 2022); and graph methods model atoms and bonds directly (Wang et al., 2025b; Bohde et al., 2025). FRIGID combines SAFE generation, molecular pretraining, and iterative correction (Bohde et al., 2026). This openness requires whole-structure inference. Anchora narrows the primary search to AEMS completion, then restores full-molecule editing for prior correction.

2.2 Scaffold-conditioned generation

Bemis–Murcko frameworks retain rings and linkers (Bemis & Murcko, 1996) and organize analogue exploration (Hu et al., 2016). Scaffold- and fragment-conditioned generators preserve supplied cores while constructing substituents (Li et al., 2020; Lim et al., 2020; Podda et al., 2020; Green et al., 2021), but a bare scaffold omits attachment sites. MADGEN fixes a retrieved scaffold; MSAnchor

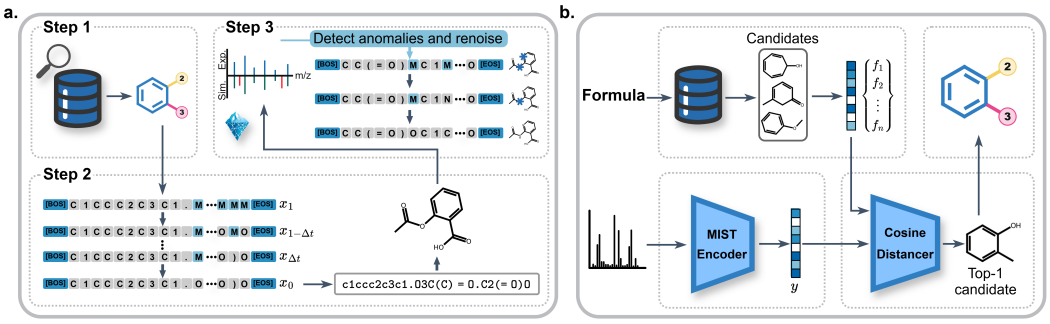

Figure 1: **Anchora overview. a,** Structure-database retrieval supplies an AEMS prior, fixed-prefix diffusion completes its sidechains, and simulated-spectrum feedback refines full molecules to explore alternatives beyond the retrieved prior. **b,** Formula-compatible molecules are ranked against a MIST-predicted spectrum fingerprint; the top reference supplies AEMS rather than a complete output structure.

predicts AEMS and preserves its edges during graph diffusion (Wang et al., 2025b; Qin et al., 2026). These constraints ensure prior consistency but propagate an incompatible prior through conditional generation. Anchora uses AEMS–SAFE completion for efficient constrained search, then sends complete molecules to a generator that need not preserve the prefix.

2.3 Molecular strings and discrete diffusion

SMILES is compact and ubiquitous (Weininger, 1988), while SELFIES guarantees robust decoding (Krenn et al., 2020). SAFE represents connected fragments whose paired closures recover molecular connectivity (Noutahi et al., 2024); its boundaries suit scaffold completion, but standard conversion does not define a canonical AEMS prefix/completion split. Anchora adds canonical anchored-prefix construction, closure reconciliation, and verified reverse conversion, forming a common interface from retrieved AEMS to diffusion training and decoding.

Diffusion reverses progressive corruption (Sohl-Dickstein et al., 2015; Ho et al., 2020); categorical variants act on discrete states (Austin et al., 2021; Hoogeboom et al., 2021) for graph and masked-language generation (Vignac et al., 2023; Sahoo et al., 2024; Lou et al., 2024). Classifier-free guidance adds conditional control (Ho & Salimans, 2022), while DiffMS and FRIGID apply discrete diffusion to molecules (Bohde et al., 2025; 2026). Anchora specializes it to fixed-prefix completion with a completion-only loss and separate length model, then uses ICEBERG as a complementary signal for unrestricted full-molecule refinement (Wang et al., 2025a).

3 Method

3.1 Problem formulation and system overview

Given an MS/MS spectrum s, known molecular formula f, and structure database $\mathcal{D}$, Anchora returns ranked molecular graphs $\{g_i\}$. It retrieves $d^* \in \mathcal{D}$, extracts its AEMS prior a, and samples

$$g \sim p_\theta(g \mid s, f, a). \tag{1}$$

Reference AEMS comes from the top database candidate; target-derived oracle AEMS estimates performance with a correct prior.

The three stages are retrieval, fixed-prefix sidechain completion, and full-molecule refinement (Figure 1a). Refinement can mask and regenerate any molecular region without preserving the prefix, providing a correction route beyond an erroneous prior. Queries without usable AEMS use full-molecule generation.

3.2 Spectrum-guided AEMS retrieval

CSU-MS2-DB contains 672,681 deduplicated structures from 23 databases (Xie et al., 2025). We use its molecular records, not its cross-modal encoder. Retrieval recalls exact-formula candidates

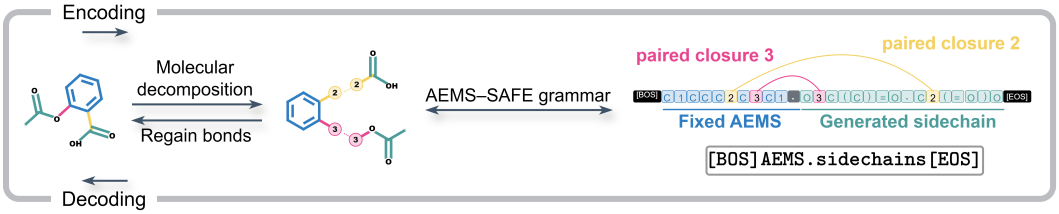

Figure 2: **Canonical AEMS–SAFE grammar.** Aspirin is split into an anchored scaffold and two sidechains. Paired closures restore the cut bonds: encoding gives fixed prefix `c1ccc2c3c1.` and completion `O3C(C)=O.C2(=O)O`, while decoding reverses the process. Diffusion fixes prefix-side closures and generates their sidechain partners.

with nonempty Murcko scaffolds, broadens to mass-compatible candidates when necessary, and canonicalizes and deduplicates the results. MIST predicts $\hat{\mathbf{z}} = E_\phi(s, f)$; candidate d has Morgan fingerprint $\mathbf{m}(d)$ (Morgan, 1965; Rogers & Hahn, 2010). We minimize their cosine distance,

$$\delta(d \mid s, f) = 1 - \frac{\hat{\mathbf{z}}^\top \mathbf{m}(d)}{\|\hat{\mathbf{z}}\|_2 \|\mathbf{m}(d)\|_2} \tag{2}$$

selecting

$$d^* = \arg\min_{d \in \mathcal{C}(f)} \delta(d \mid s, f), \tag{3}$$

where $\mathcal{C}(f)$ is the recalled set (Figure 1b). Full-candidate fingerprints retain core and substituent evidence until selection; only then is the reference compressed into AEMS for generation.

3.3 AEMS–SAFE GRAMMAR AND BIDIRECTIONAL CONVERSION

AEMS extraction. AEMS explicitly encodes side-chain anchoring points on a molecular scaffold (Qin et al., 2026). Given a molecular graph g, we extract its Murcko scaffold $S(g)$ and cut every bond with exactly one endpoint in the scaffold. This yields an anchored scaffold and one or more sidechain components. In SAFE, each cut is represented by a paired closure, with one endpoint in the scaffold prefix and the other in the completion; the bond symbol is retained when reconnecting the two parts.

SMILES to AEMS–SAFE. A direct cut is not yet canonical because serialization order and scaffold symmetry can permute attachment sites. Anchora therefore canonically ranks the atoms of g, selects the lowest-ranked scaffold atom as the serialization root, and applies a scaffold–sidechain SAFE cut. The rooted encoding places the anchored scaffold before the first fragment separator and the detached sidechains after it. We decode the raw prefix into a graph with dummy cut atoms, remove temporary annotations, canonicalize the anchored graph, and re-encode it to obtain the authoritative prefix P_{AEMS}.

Because canonicalization may permute symmetry-equivalent anchors, prefix and completion must be constructed jointly. We enumerate graph isomorphisms from the raw anchored scaffold to its canonical form, apply each induced closure mapping to the sidechain string, and keep only mappings whose full SAFE string decodes to the same canonical molecular graph as the input. The resulting grammar is

$$[\text{BOS}]; P_{\text{AEMS}}; .; C_{\text{sidechain}}; [\text{EOS}], \tag{4}$$

where the period marks the prefix/completion boundary and $C_{\text{sidechain}}$ contains the missing fragments. Closure numbers pair serialized endpoints rather than denote atom attributes or chemical labels. For aspirin, the procedure yields

$$\texttt{c1ccc2c3c1.O3C(C)=O.C2(=O)O}, \tag{5}$$

where `c1ccc2c3c1.` is the fixed prefix and the remaining tokens form the completion. For this molecule, the raw and canonical anchored serializations coincide, so no closure relabeling is needed.

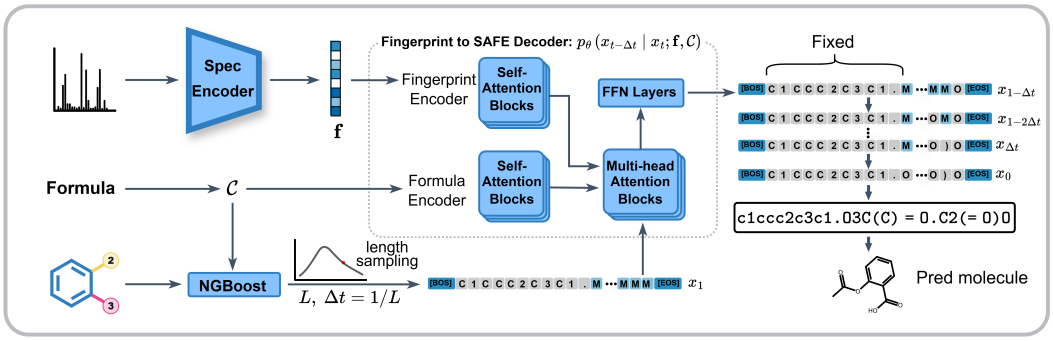

Figure 3: **AEMS-conditioned completion diffusion.** Spectrum, formula, and AEMS determine the fingerprint, completion length, and decoder conditions. The decoder restores the canonical AEMS prefix at every step, generates only the masked completion, and decodes the resulting SAFE sequence to a molecular graph.

AEMS–SAFE to SMILES. The reverse converter joins prefix and completion, resolves paired closures, restores cut bonds, and emits canonical SMILES. Training examples must reconstruct the input graph. Generated candidates instead undergo decoding and formula/AEMS checks, without access to the target structure.

3.4 AEMS COMPLETION DIFFUSION

Figure 3 shows how Anchora turns AEMS completion into conditional masked diffusion. The spectrum encoder maps the experimental spectrum and molecular formula to a probabilistic fingerprint $\hat{\mathbf{z}} = E_\phi(s, f)$. Separate self-attention encoders transform the active fingerprint bits and elemental composition into condition tokens, which the SAFE decoder reads through cross-attention. In parallel, a probabilistic length model estimates

$$L \sim q_\psi(L \mid f, a), \tag{6}$$

where L counts tokens after the AEMS boundary, excluding the observed prefix. Generation starts from

$$x_1 = [\mathrm{BOS}]\ P_{\mathrm{AEMS}}\ .\ [\mathrm{MASK}]^L\ [\mathrm{EOS}]. \tag{7}$$

Let J_p and J_c denote prefix and completion positions. A log-linear schedule masks completion tokens; prefix tokens are restored after forward corruption and every reverse transition,

$$x_{t,j} = x_{0,j} \qquad \forall j \in J_p. \tag{8}$$

The decoder thus preserves AEMS and prefix-side closures while generating all completion atoms, bonds, branches, separators, and closure partners. Denoising loss applies only to J_c:

$$\mathcal{L}_{\mathrm{comp}} = \mathbb{E}_{t,x_t}\left[\frac{1}{|J_c|}\sum_{j\in J_c} \ell(x_{0,j}, p_\theta(x_{0,j} \mid x_t, t, \hat{\mathbf{z}}, f, P_{\mathrm{AEMS}}))\right]. \tag{9}$$

This construction makes training and inference identical at the prefix/completion boundary. In particular, the completion-side digits in Equation 5 are generated by diffusion; they are not inserted by a deterministic post-processing rule.

During joint adaptation, spectrum-predicted fingerprints replace structural fingerprints. The encoder and decoder optimize completion loss together with binary cross-entropy $\mathcal{L}_{\mathrm{FP}}$ against target molecular fingerprints:

$$\mathcal{L}_{\mathrm{joint}} = \mathcal{L}_{\mathrm{comp}} + \lambda_{\mathrm{fp}}\mathcal{L}_{\mathrm{FP}}. \tag{10}$$

At inference, the same spectrum-predicted fingerprint is also compared with candidate Morgan fingerprints to rank generated structures, aligning generation and selection around a shared spectrum-derived signal.

3.5 TRAINING CURRICULUM

Training has four stages. **General SAFE initialization** learns molecular syntax, token co-occurrence, and full-molecule editing without a fixed/generated boundary. **Large-scale AEMS completion pretraining** learns sidechain and closure statistics from codec-verified structures beyond the paired benchmarks (Section 4). It conditions on formula, structural fingerprint, and canonical prefix, restricting masking and loss to sidechains.**Dataset adaptation** continues completion training on CANOPUS or MSG training structures, using structure-derived fingerprints to adapt the decoder to benchmark-specific sidechain chemistry without updating the spectrum encoder. **Spectrum–decoder joint adaptation** replaces structural with predicted fingerprints and trains both modules under $\mathcal{L}_{\text{joint}}$, reducing the conditioning mismatch. Section 4.6 evaluates these transitions.

3.6 THREE-ROUND FULL-MOLECULE REFINEMENT

Fixed-prefix completion cannot revise retrieval errors, so refinement decodes candidates into ordinary SMILES and permits editing anywhere. Each round combines fresh completions with edits from a general full-molecule diffusion model. Discrepancies between ICEBERG-predicted and observed spectra guide masking. New and retained candidates are merged, checked for formula/AEMS consistency, and ranked by Morgan Tanimoto similarity to the jointly adapted encoder's predicted fingerprint. Thus AEMS guides completion without constraining all refinement proposals, providing a correction route for imperfect retrieval. Queries lacking usable AEMS use full-molecule generation followed by the same refinement.

4 EXPERIMENTS

Our evaluation try to answer three research questions: *RQ1 Main Comparison.* Compared with baselines, how does ANCHORA perform on commonly used benchmarks? *RQ2 Ablation Study.* How do the components in the method affect performance? *RQ3 Zero-Shot Transferring.* Can the model be transferred to external spectral libraries? We compare against published baselines on two benchmarks, analyze retrieval, training, and refinement, and evaluate frozen models on five external libraries.

4.1 DATASETS AND BASELINES

We evaluate on CANOPUS and MassSpecGym (MSG) using established splits (Dührkop et al., 2021; Bushuiev et al., 2024). CANOPUS contains curated GNPS spectrum–structure pairs, while MSG uses a standardized split with stronger structural separation between train and test molecules. All methods receive the molecular formula.

For zero-shot transfer, we additionally use five external GNPS libraries (Wang et al., 2016): ECRFS, Pesticides, MIADB, IQAMDB, and WineDB, from the public exports dated 20 September 2026. Molecules overlapping CANOPUS supervised training data are removed using the connectivity block of non-stereochemical InChIKeys. These spectra are used only for evaluation.

We compare with MADGEN, DiffMS, MSAnchor, and FRIGID using published top-1 and top-10 results (Wang et al., 2025b; Bohde et al., 2025; Qin et al., 2026; Bohde et al., 2026). Anchora uses either a retrieved *reference* AEMS or a target-derived *oracle* AEMS. Anchora-Base performs initial completion with full-molecule routing, while Anchora adds three refinement rounds. Zero-shot transfer compares frozen CANOPUS-trained Anchora-Base and FRIGID-base models.

4.2 EVALUATION METRICS

At $k \in \{1, 10\}$, Acc@k measures exact target recovery, Sim@k the maximum target–candidate Morgan Tanimoto similarity, and MCES@k the minimum maximum-common-edge-subgraph distance (Kretschmer et al., 2023). Acc and Sim are maximized; MCES is minimized. Predictions are canonicalized and deduplicated. Scaffold Acc. measures exact Murcko-scaffold recovery, and AEMS Acc. exact constraint-signature recovery.

Table 1: Results on CANOPUS and MassSpecGym (MSG). Prior quality is exact Murcko-scaffold and strict AEMS-signature accuracy on queries with defined target AEMS; oracle is 100% by construction, and MADGEN reports published SPA. Anchora-Base includes initial AEMS completion and full-molecule routing; Anchora adds three refinement rounds. Baselines are from the cited studies. Bold and underline mark the best and second-best non-oracle structure results; prior columns are not ranked.

	Prior quality		Top-1			Top-10		
Model	Scaf. (%) ↑	AEMS (%) ↑	Acc. (%) ↑	MCES ↓	Sim. ↑	Acc. (%) ↑	MCES ↓	Sim. ↑
CANOPUS								
MADGEN	17.70	–	2.10	20.56	0.22	2.39	12.69	0.27
DiffMS	–	–	8.34	11.95	0.35	15.44	9.23	0.47
MSAnchor	–	–	8.51	11.12	0.38	16.90	8.95	0.49
FRIGID	–	–	25.03	7.10	0.58	33.37	5.56	0.64
Anchora-Base (Reference)	66.53	59.78	<u>36.86</u>	<u>5.54</u>	<u>0.64</u>	<u>39.85</u>	<u>4.69</u>	**0.68**
Anchora (Reference)	66.53	59.78	**38.98**	**5.42**	**0.65**	**41.72**	**4.63**	**0.68**
Anchora-Base (Oracle)	100.00	100.00	60.27	2.39	0.83	68.74	1.47	0.89
Anchora (Oracle)	100.00	100.00	64.38	2.22	0.85	70.86	1.48	0.89
MassSpecGym								
MADGEN	13.20	–	1.31	27.47	0.20	1.54	16.84	0.26
DiffMS	–	–	2.30	18.45	0.28	4.25	14.73	0.39
MSAnchor	–	–	4.73	13.92	0.37	6.16	12.53	0.46
FRIGID	–	–	18.29	13.49	0.43	22.00	11.65	0.47
Anchora-Base (Reference)	45.14	43.27	<u>21.63</u>	<u>10.70</u>	<u>0.49</u>	<u>22.95</u>	<u>9.45</u>	<u>0.53</u>
Anchora (Reference)	45.14	43.27	**23.50**	**10.57**	**0.50**	**25.04**	**9.42**	**0.54**
Anchora-Base (Oracle)	100.00	100.00	49.33	3.52	0.81	54.75	2.36	0.86
Anchora (Oracle)	100.00	100.00	52.94	3.30	0.82	58.37	2.31	0.87

4.3 Experimental settings

Following DiffMS (Bohde et al., 2025), we pretrain on 2.71 million AEMS completions from HMDB, DSSTox, COCONUT, and MOSES (Wishart et al., 2022; Williams et al., 2017; Sorokina et al., 2021; Polykovskiy et al., 2020). Molecules are deduplicated by non-stereochemical InChI, excluding CANOPUS/MSG validation and test overlaps. NGBoost predicts completion length (Duan et al., 2020). Anchora uses 128 initial attempts; each of three refinement rounds adds 64 completions and up to 64 full-molecule edits, for up to 512 generation attempts. Valid unique candidates are ranked, with at most 100 returned; $K = 16$ transition candidates each receive $M = 4$ masked variants.

For external GNPS evaluation, we retain spectra with supported positive-ion adducts, valid single-component structures with computable formulas, and nonempty MS/MS peaks, yielding 1,066 spectra. CANOPUS-trained models are evaluated directly without external-library training or fine-tuning.

4.4 Main results

Table 1 shows that Anchora improves both exact recovery and structural similarity. Reference-guided Acc@1 reaches 38.98% on CANOPUS and 23.50% on MSG, exceeding FRIGID by 13.95 and 5.21 percentage points, respectively. Higher Acc@10 and Sim, together with lower MCES, indicate that the benefit extends beyond top-ranked exact matches. This pattern is consistent with the role of AEMS in narrowing the search around a plausible molecular core while retaining flexibility for spectrum-conditioned side-chain reconstruction.

Prior-quality results support this interpretation. Reference Scaffold/AEMS Acc. reaches 66.53%/59.78% on CANOPUS and 45.14%/43.27% on MSG. Scaffold accuracy exceeds MADGEN's published results, while both metrics outperform CSU-MS2 (Xie et al., 2025) under matched test subsets and prior-quality definitions (Figure 4c). The latter comparison shows that the advantage extends beyond core identity to attachment-site specification, which determines where molecular completion can proceed. These results support the design rationale of retaining core and substituent evidence during complete-molecule ranking before compressing the selected reference into an AEMS prior.

Table 2: Zero-shot transfer to five GNPS libraries using frozen CANOPUS generation models without refinement; n counts eligible spectra. Scaf./AEMS evaluate generated Top-1 structures. Other metrics follow Table 1; bold marks the best result per library.

	Structural match		Top-1			Top-10		
Model	Scaf. (%) ↑	AEMS (%) ↑	Acc. (%) ↑	MCES ↓	Sim. ↑	Acc. (%) ↑	MCES ↓	Sim. ↑
ECRFS ($n = 56$)								
FRIGID-base	32.14	19.64	14.29	7.09	0.39	21.43	5.50	0.48
Anchora-Base (Reference)	**60.71**	**51.79**	**25.00**	**5.43**	**0.48**	**30.36**	**4.57**	**0.55**
Pesticides ($n = 90$)								
FRIGID-base	37.78	35.56	25.56	9.00	0.47	28.89	7.26	0.53
Anchora-Base (Reference)	**84.44**	**83.33**	**34.44**	**6.02**	**0.63**	**36.67**	**4.61**	**0.69**
MIADB ($n = 239$)								
FRIGID-base	0.84	0.00	0.00	15.03	0.23	0.00	13.01	0.27
Anchora-Base (Reference)	**27.20**	**25.52**	**13.81**	**10.06**	**0.43**	**14.23**	**8.93**	**0.45**
IQAMDB ($n = 269$)								
FRIGID-base	7.43	1.86	1.12	11.08	0.33	1.12	9.86	0.37
Anchora-Base (Reference)	**67.29**	**43.49**	**28.25**	**3.28**	**0.70**	**37.17**	**2.50**	**0.74**
WineDB ($n = 412$)								
FRIGID-base	30.10	13.11	8.50	13.11	0.34	10.44	9.92	0.40
Anchora-Base (Reference)	**51.46**	**26.21**	**13.59**	**7.99**	**0.44**	**16.99**	**6.89**	**0.49**

In targeted or multimodal applications, external structural evidence—such as a known synthetic scaffold, a characterized congener, or NMR-resolved attachment sites—may specify an AEMS while leaving the complete molecule unknown. We therefore evaluate target-derived oracle AEMS to model this setting and separate prior selection from molecular completion. With oracle AEMS, Acc@1 rises to 64.38% on CANOPUS and 52.94% on MSG. The strong performance under known AEMS shows that Anchora has a high performance ceiling when supplied with an accurate structural prior, while the reference–oracle gap highlights the importance of prior quality.

4.5 Zero-shot transfer to external spectral libraries

To test whether Anchora's gains extend beyond benchmark-specific distributions, we evaluate frozen CANOPUS-trained models on five external GNPS libraries after removing overlaps with CANOPUS training data. No external fine-tuning or refinement is used, isolating the transferability of retrieval-guided AEMS completion.

Anchora-Base improves Acc@1 over FRIGID-base by 5.10–27.14 percentage points across all five libraries (Table 2), with consistent gains in Acc@10, MCES, and Sim. Notably, it recovers exact structures on MIADB and IQAMDB, where FRIGID-base achieves little exact recovery. These results show that reference-guided completion remains effective when full-molecule generation transfers poorly, supporting generalization beyond the CANOPUS distribution while indicating library-dependent transfer difficulty.

4.6 Ablation studies

Figure 4a shows complementary gains from large-scale completion pretraining and joint adaptation. Relative to the standard curriculum, their combination raises reference Acc@1/Acc@10 from 33.25%/36.49% to 36.86%/39.85% and oracle accuracy from 52.55%/62.76% to 60.27%/68.74%. Intermediate variants indicate that joint adaptation contributes the larger exact-recovery gain, while large-scale pretraining supplies complementary improvements. Appendix Table 3 reports all variants.

The first refinement round contributes the largest step. Across three rounds, Acc@1 gains are 2.12/4.11 points for CANOPUS reference/oracle and 1.87/3.61 for MSG (Figure 4b). Target-AEMS Top-1 rises from 57.49% to 59.51% and from 39.72% to 41.98%, respectively (Figure 4d). Because

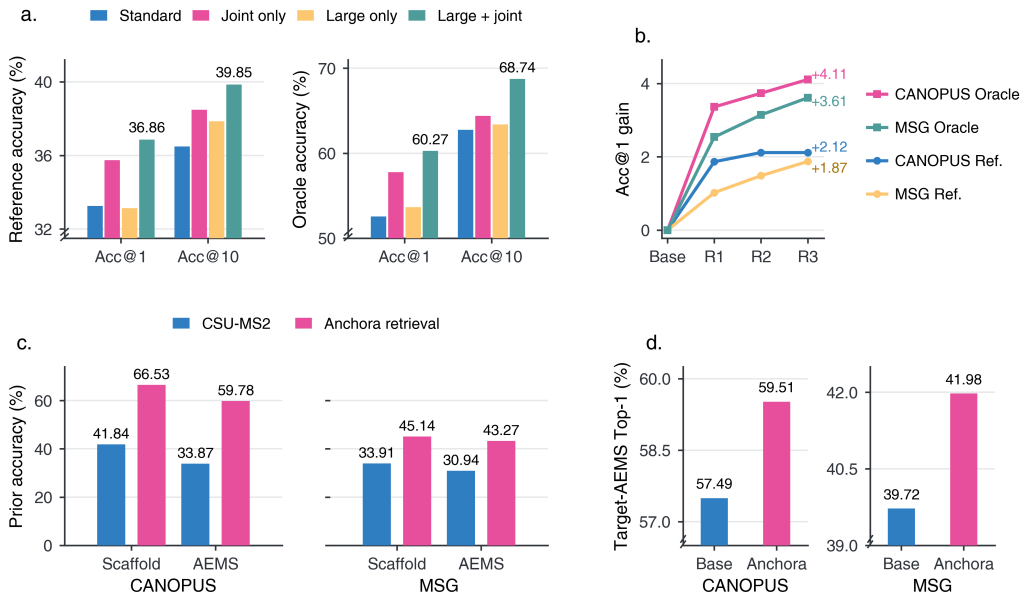

Figure 4: **Retrieval, training, and refinement analysis. a,** CANOPUS training-curriculum ablations under reference (left) and oracle (right) priors. **b,** Acc@1 gain across three refinement rounds. **c,** Scaffold and AEMS prior accuracy of the CSU-MS2 contrastive retriever and Anchora retrieval on CANOPUS and MSG. **d,** Top-1 predictions satisfying the target AEMS before and after three refinement rounds. Complete training and refinement values are provided in Appendix B.

full-molecule editing is not confined to completion, these gains support a route beyond the fixed prefix without implying universal correction. Appendix Table 4 gives absolute values.

Predicting completion rather than full target length reduces token RMSE from 5.33 to 3.18 and raises AEMS pass from 62.92% to 74.99%, while preserving Acc@1 (Appendix Table 6).

5 Discussion

Anchora combines complementary retrieval and generation. Formula-constrained full-molecule retrieval retains core and substituent evidence before compressing the reference into AEMS; AEMS–SAFE completion then narrows the generation space. Prefix-free full-molecule refinement can correct an incompatible prior, but correction still depends on ranking. The reference–oracle gap therefore shows that prior selection remains the main bottleneck.

6 Conclusion

Anchora unifies retrieval and de novo generation for known-formula MS/MS elucidation through AEMS retrieval, reversible AEMS–SAFE completion, curriculum training, and full-molecule refinement. On CANOPUS and MassSpecGym, it improves exact recovery and structural similarity while retaining a route beyond the retrieved molecule. Zero-shot evaluation on five external GNPS libraries further shows consistent gains over FRIGID-base, supporting transfer beyond the benchmark distributions. Future work will explore uncertainty-aware multi-prior retrieval.

AI USE STATEMENT

Generative AI tools assisted with code inspection and manuscript drafting, LaTeX preparation, and build-quality checks. They were not used to create experimental spectra, molecular labels, or numerical measurements. The authors verified technical descriptions and reported values, reviewed the manuscript and citations, and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The experimental settings describe the data sources, training, inference budget, and metrics. The anonymous supplement provides the AEMS–SAFE codec, data-construction, training, and inference code, configuration examples, environment information, and data-access instructions. Public datasets remain governed by their providers' terms.

REFERENCES

Felicity Allen, Allison Pon, Michael Wilson, Russ Greiner, and David Wishart. CFM-ID: A web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Research*, 42(W1):W94–W99, 2014. doi: 10.1093/nar/gku436.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html.

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.

Sebastian Böcker and Kai Dührkop. Fragmentation trees reloaded. *Journal of Cheminformatics*, 8(1):5, 2016. doi: 10.1186/s13321-016-0116-8.

Montgomery Bohde, Mrunali Manjrekar, Runzhong Wang, Shuiwang Ji, and Connor W. Coley. DiffMS: Diffusion generation of molecules conditioned on mass spectra. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4737–4756. PMLR, 2025. URL https://proceedings.mlr.press/v267/bohde25a.html.

Montgomery Bohde, Hongxuan Liu, Mrunali Manjrekar, Magdalena Lederbauer, Shuiwang Ji, Runzhong Wang, and Connor W. Coley. FRIGID: Scaling diffusion-based molecular generation from mass spectra at training and inference time. *arXiv preprint arXiv:2604.16648*, 2026. doi: 10.48550/arXiv.2604.16648. URL https://arxiv.org/abs/2604.16648.

Roman Bushuiev, Anton Bushuiev, Niek F. de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, Marcus Ludwig, Nils A. Haupt, Apurva Kalia, Corinna Brungs, Robin Schmid, Russell Greiner, Bo Wang, David S. Wishart, Li-Ping Liu, Juho Rousu, Wout Bittremieux, Hannes Rost, Tytus D. Mak, Soha Hassoun, Florian Huber, Justin J. J. van der Hooft, Michael A. Stravs, Sebastian Böcker, Josef Sivic, and Tomáš Pluskal. MassSpecGym: A benchmark for the discovery and identification of molecules. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c6c31413d5c53b7d1c343c1498734b0f-Abstract-Datasets_and_Benchmarks_Track.html.

Thomas Butler, Abraham Frandsen, Rose Lightheart, Brian Bargh, James Taylor, TJ Bollerman, Thomas Kerby, Kiana West, Gennady Voronov, Kevin Moon, et al. MS2Mol: A transformer model for illuminating dark chemical space from mass spectra. *ChemRxiv*, 2023. doi: 10.26434/chemrxiv-2023-vsmpx-v2.

Yannick Djoumbou-Feunang, Allison Pon, Naama Karu, Jiamin Zheng, Carin Li, David Arndt, Maheswor Gautam, Felicity Allen, and David S. Wishart. CFM-ID 3.0: Significantly improved ESI-MS/MS prediction and compound identification. *Metabolites*, 9(4):72, 2019. doi: 10.3390/metabo9040072.

Tony Duan, Anand Avati, Daisy Yi Ding, Khanh K. Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. NGBoost: Natural gradient boosting for probabilistic prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2690–2700. PMLR, 2020. URL https://proceedings.mlr.press/v119/duan20a.html.

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.

Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Alexander A. Aksenov, Alexey V. Melnik, Marvin Meusel, Pieter C. Dorrestein, Juho Rousu, and Sebastian Böcker. SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods*, 16(4):299–302, 2019. doi: 10.1038/s41592-019-0344-8.

Kai Dührkop, Louis-Félix Nothias, Markus Fleischauer, Raphael Reher, Marcus Ludwig, Markus A. Hoffmann, Daniel Petras, William H. Gerwick, Juho Rousu, Pieter C. Dorrestein, and Sebastian Böcker. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology*, 39:462–471, 2021. doi: 10.1038/s41587-020-0740-8.

Samuel Goldman, Jeremy Wohlwend, Martin Stražar, Guy Haroush, Ramnik J. Xavier, and Connor W. Coley. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence*, 5:965–979, 2023. doi: 10.1038/s42256-023-00708-3.

Harrison Green, David R. Koes, and Jacob D. Durrant. DeepFrag: A deep convolutional neural network for fragment-based lead optimization. *Chemical Science*, 12(23):8036–8047, 2021. doi: 10.1039/D1SC00163A.

Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, 44(D1):D1214–D1219, 2016. doi: 10.1093/nar/gkv1031.

Markus Heinonen, Huibin Shen, Nicola Zamboni, and Juho Rousu. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*, 28(18):2333–2341, 2012. doi: 10.1093/bioinformatics/bts437.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. doi: 10.48550/arXiv.2207.12598.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12454–12465, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/67d96d458abdef21792e6d8e590244e7-Abstract.html.

Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, et al. MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714, 2010. doi: 10.1002/jms.1777.

Ye Hu, Dagmar Stumpfe, and Jürgen Bajorath. Computational exploration of molecular scaffolds in medicinal chemistry. *Journal of Medicinal Chemistry*, 59(9):4062–4076, 2016. doi: 10.1021/acs.jmedchem.5b01746.

Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H. Spaaks, Faruk Diblen, Simon Rogers, and Justin J. J. van der Hooft. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology*, 17(2):e1008724, 2021a. doi: 10.1371/journal.pcbi.1008724.

Florian Huber, Sven van der Burg, Justin J. J. van der Hooft, and Lars Ridder. MS2DeepScore: A novel deep learning similarity measure to compare tandem mass spectra. *Journal of Cheminformatics*, 13(1):84, 2021b. doi: 10.1186/s13321-021-00558-4.

Apurva Kalia, Yan Zhou Chen, Dilip Krishnan, and Soha Hassoun. JESTR: Joint embedding space technique for ranking candidate molecules for the annotation of untargeted metabolomics data. *Bioinformatics*, 41(7):btaf354, 2025. doi: 10.1093/bioinformatics/btaf354.

Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, et al. PubChem substance and compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, 2016. doi: 10.1093/nar/gkv951.

Tobias Kind, Hiroshi Tsugawa, Tomas Cajka, Yan Ma, Zijuan Lai, Sajjan S. Mehta, Gert Wohlgemuth, Dinesh K. Barupal, Megan R. Showalter, Masanori Arita, and Oliver Fiehn. Identification of small molecules using accurate mass MS/MS search. *Mass Spectrometry Reviews*, 37(4):513–532, 2018. doi: 10.1002/mas.21535.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020. doi: 10.1088/2632-2153/aba947.

Fleming Kretschmer, Jan Seipp, Marcus Ludwig, Gunnar W. Klau, and Sebastian Böcker. Small molecule machine learning: All models are wrong, some may not even be useful. *bioRxiv*, 2023. doi: 10.1101/2023.03.27.534311.

Yibo Li, Jianxing Hu, Yanxing Wang, Jielong Zhou, Liangren Zhang, and Zhenming Liu. DeepScaffold: A comprehensive tool for scaffold-based de novo drug discovery using deep learning. *Journal of Chemical Information and Modeling*, 60(1):77–91, 2020. doi: 10.1021/acs.jcim.9b00727.

Xinmeng Li, Yan Zhou Chen, Apurva Kalia, Hao Zhu, Li-Ping Liu, and Soha Hassoun. An ensemble spectral prediction (ESP) model for metabolite annotation. *Bioinformatics*, 40(8):btae490, 2024. doi: 10.1093/bioinformatics/btae490.

Jaechang Lim, Sang-Yeon Hwang, Seokhyun Moon, Seungsu Kim, and Woo Youn Kim. Scaffold-based molecular design with a graph generative model. *Chemical Science*, 11(4):1153–1164, 2020. doi: 10.1039/C9SC04503A.

Eleni Litsa, Vijil Chenthamarakshan, Payel Das, and Lydia E. Kavraki. Spec2Mol: An end-to-end deep learning framework for translating MS/MS spectra to de novo molecules. *ChemRxiv*, 2021. doi: 10.26434/chemrxiv-2021-6rdh6.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024. URL https://arxiv.org/abs/2310.16834.

Harry L. Morgan. The generation of a unique machine description for chemical structures—a technique developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5(2):107–113, 1965. doi: 10.1021/c160017a018.

Emmanuel Noutahi, Cristian Gabellini, Michael Craig, Jonathan S. C. Lim, and Prudencio Tossou. Gotta be SAFE: A new framework for molecular design. *Digital Discovery*, 3:796–804, 2024. doi: 10.1039/D4DD00019F.

Harry E. Pence and Antony Williams. ChemSpider: An online chemical information resource. *Journal of Chemical Education*, 87(11):1123–1124, 2010. doi: 10.1021/ed100697w.

Marco Podda, Davide Bacciu, and Alessio Micheli. A deep generative model for fragment-based molecule generation. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2240–2250. PMLR, 2020. URL https://proceedings.mlr.press/v108/podda20a.html.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (MOSES): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11:565644, 2020. doi: 10.3389/fphar.2020.565644.

Xiaohan Qin, Chao Wang, Zhengyang Zhou, Linjiang Chen, Wenjie Du, and Yang Wang. MSAnchor: De novo molecular generation from mass spectrometry data with anchor-extended molecular scaffolds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 953–961, 2026. doi: 10.1609/aaai.v40i2.37064.

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

Christoph Ruttkies, Emma L. Schymanski, Stefan Wolf, Juliane Hollender, and Steffen Neumann. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics*, 8(1):3, 2016. doi: 10.1186/s13321-016-0115-9.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130136–130184, 2024. doi: 10.52202/079017-4135.

Aditya Divyakant Shrivastava, Neil Swainston, Soumitra Samanta, Ivayla Roberts, Marina Wright Muelas, and Douglas B. Kell. MassGenie: A transformer-based deep learning method for identifying small molecules from their mass spectra. *Biomolecules*, 11(12):1793, 2021. doi: 10.3390/biom11121793.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265. PMLR, 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Maria Sorokina, Peter Merseburger, Kohulan Rajan, Mehmet Aziz Yirik, and Christoph Steinbeck. COCONUT online: Collection of open natural products database. *Journal of Cheminformatics*, 13(1):2, 2021. doi: 10.1186/s13321-020-00478-9.

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

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

Mingxun Wang, Jeremy J. Carver, Vanessa V. Phelan, Laura M. Sanchez, Neha Garg, Yao Peng, Don D. Nguyen, Jeramie Watrous, Clifford A. Kapono, Tal Luzzatto-Knaan, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature Biotechnology*, 34(8): 828–837, 2016. doi: 10.1038/nbt.3597.

Runzhong Wang, Mrunali Manjrekar, Babak Mahjour, Julian Avila-Pacheco, Joules Provenzano, Erin Reynolds, Magdalena Lederbauer, Eivgeni Mashin, Samuel Goldman, Mingxun Wang, Jing-Ke Weng, Desiree L. Plata, Clary B. Clish, and Connor W. Coley. Neural spectral prediction for structure elucidation with tandem mass spectrometry. *bioRxiv*, 2025a. doi: 10.1101/2025.05.28.656653. URL https://doi.org/10.1101/2025.05.28.656653.

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

Jennifer N. Wei, David Belanger, Ryan P. Adams, and D. Sculley. Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS Central Science*, 5(4):700–708, 2019. doi: 10.1021/acscentsci.9b00085.

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.

Antony J. Williams, Christopher M. Grulke, Jeff Edwards, Andrew D. McEachran, Kamel Mansouri, Nancy C. Baker, Grace Patlewicz, Imran Shah, John F. Wambaugh, Richard S. Judson, and Ann M. Richard. The CompTox chemistry dashboard: A community data resource for environmental chemistry. *Journal of Cheminformatics*, 9(1):61, 2017. doi: 10.1186/s13321-017-0247-6.

David S. Wishart, An Chi Guo, Eponine Oler, Fei Wang, Afia Anjum, Harrison Peters, Raynard Dizon, Zinat Sayeeda, Siyang Tian, Brian L. Lee, et al. HMDB 5.0: The human metabolome database for 2022. *Nucleic Acids Research*, 50(D1):D622–D631, 2022. doi: 10.1093/nar/gkab1062.

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

Ting Xie, Hailiang Zhang, Qiong Yang, Jinyu Sun, Yue Wang, Jia Long, Zhimin Zhang, and Hongmei Lu. CSU-MS2: A contrastive learning framework for cross-modal compound identification from MS/MS spectra to molecular structures. *Analytical Chemistry*, 97(25):13350–13360, 2025. doi: 10.1021/acs.analchem.5c01594.

Adamo Young, Hannes Röst, and Bo Wang. Tandem mass spectrum prediction for small molecules using graph transformers. *Nature Machine Intelligence*, 6(4):404–416, 2024. doi: 10.1038/s42256-024-00816-8.

Hao Zhu, Liping Liu, and Soha Hassoun. Using graph neural networks for mass spectrometry prediction. *arXiv preprint arXiv:2010.04661*, 2020. doi: 10.48550/arXiv.2010.04661.

Table 3: CANOPUS training-curriculum ablation at the base-generation stage before iterative refinement. "Large AEMS" denotes large-scale completion pretraining; "Full joint" denotes spectrum–decoder adaptation. All variants use the same B=128 evaluation and full-molecule routing. Values are percentages.

Training stage		Reference prior		Oracle prior		AEMS pass (%)	
Large AEMS	Full joint	Acc@1	Acc@10	Acc@1	Acc@10	Reference	Oracle
No	No	33.25	36.49	52.55	62.76	64.30	62.37
No	Yes	35.74	38.48	57.78	64.38	67.45	65.84
Yes	No	33.13	37.86	53.67	63.39	74.19	72.43
Yes	Yes	**36.86**	**39.85**	**60.27**	**68.74**	**75.69**	**73.77**

Table 4: End-to-end accuracy across refinement rounds under a common output protocol. Target-AEMS Top-1 compares top-ranked predictions before and after three refinement rounds.

Dataset	Prior	Anchora-Base	+Round 1	+Round 2	Anchora (+Round 3)
Acc@1 (%)					
CANOPUS	Reference	36.86	38.73	38.98	38.98
CANOPUS	Oracle	60.27	63.64	64.01	64.38
MSG	Reference	21.63	22.65	23.11	23.50
MSG	Oracle	49.33	51.87	52.48	52.94
Acc@10 (%)					
CANOPUS	Reference	39.85	41.10	41.72	41.72
CANOPUS	Oracle	68.74	69.49	70.61	70.86
MSG	Reference	22.95	23.96	24.58	25.04
MSG	Oracle	54.75	56.46	57.51	58.37
Target-AEMS Top-1 (%)					
CANOPUS	Reference	57.49	–	–	59.51
MSG	Reference	39.72	–	–	41.98

A Detailed experimental settings

The completion decoder is pretrained for 200,000 steps with global batch size 256 and learning rate 10^{-5}. Dataset adaptation uses 10,000 steps, batch size 32, and learning rate 5×10^{-6}. During joint adaptation, the spectrum encoder and decoder use learning rates of 10^{-5} and 2×10^{-6}, respectively, with fingerprint-loss weight 0.1; CANOPUS uses 5,000 steps and MSG uses 10,000. Training uses bfloat16 precision. Experiments were run on one NVIDIA GeForce RTX 5090 GPU (32 GB) with an Intel Core Ultra 9 285K processor and 32 GB system memory.

B Additional ablation results

Table 3 gives the numerical training-curriculum results summarized in Figure 4a, including AEMS consistency. Table 4 gives the absolute Acc@1 and Acc@10 values behind Figure 4b and the Target-AEMS comparison in panel d.

Table 5 expands the round-wise analysis to all six structure metrics. Every displayed stage is a complete ranked output: candidates satisfying both molecular-formula and AEMS constraints are placed first, followed by formula-compatible and remaining candidates in joint-fingerprint order. Queries without a usable AEMS follow the corresponding full-molecule route. This common protocol makes Accuracy, Sim, and MCES directly comparable across rounds.

Target-AEMS Top-1 compares the AEMS of the highest-ranked generated structure with the target AEMS on eligible reference queries. It rises from 57.49% for Anchora-Base to 59.51% for Anchora on CANOPUS, and from 39.72% to 41.98% on MSG. This diagnostic complements exact whole-molecule recovery by asking whether refinement improves the retrieved structural core represented in the top prediction.

Table 6 separates validation error of the probabilistic length model from the realized generation-length error. Training directly on completion length improves both quantities and substantially increases the fraction of generated structures that retain the AEMS condition.

Table 5: End-to-end round-wise results under a unified evaluation protocol. Round 1 and Round 2 outputs are ranked from their accumulated candidate pools using the same constraint-prioritized completion and full-molecule routing rules as the complete model.

Dataset	Prior	Stage	Acc@1 (%)	MCES@1	Sim@1	Acc@10 (%)	MCES@10	Sim@10
CANOPUS	Reference	Anchora-Base	36.86	5.54	0.64	39.85	4.69	0.68
		Round 1	38.73	5.45	0.64	41.10	4.67	0.68
		Round 2	38.98	5.43	0.65	41.72	4.62	0.68
		Anchora (Round 3)	38.98	5.42	0.65	41.72	4.63	0.68
CANOPUS	Oracle	Anchora-Base	60.27	2.39	0.83	68.74	1.47	0.89
		Round 1	63.64	2.27	0.85	69.49	1.51	0.89
		Round 2	64.01	2.28	0.85	70.61	1.50	0.89
		Anchora (Round 3)	64.38	2.22	0.85	70.86	1.48	0.89
MSG	Reference	Anchora-Base	21.63	10.70	0.49	22.95	9.45	0.53
		Round 1	22.65	10.63	0.50	23.96	9.45	0.53
		Round 2	23.11	10.60	0.50	24.58	9.44	0.54
		Anchora (Round 3)	23.50	10.57	0.50	25.04	9.42	0.54
MSG	Oracle	Anchora-Base	49.33	3.52	0.81	54.75	2.36	0.86
		Round 1	51.87	3.41	0.82	56.46	2.34	0.87
		Round 2	52.48	3.35	0.82	57.51	2.33	0.87
		Anchora (Round 3)	52.94	3.30	0.82	58.37	2.31	0.87

Table 6: Completion-length prediction on the CANOPUS oracle setting. Errors are measured in decoder tokens. Predicting only the missing completion improves calibration and AEMS consistency while preserving exact recovery.

Length target	Acc@1 (%)	AEMS pass (%)	Formula success (%)	Gen. MAE	Gen. RMSE	Val. MAE	Val. RMSE
Full target sequence	42.71	62.92	95.27	3.82	5.33	2.92	4.03
Completion only	**42.84**	**74.99**	**95.52**	**2.32**	**3.18**	**1.59**	**2.22**

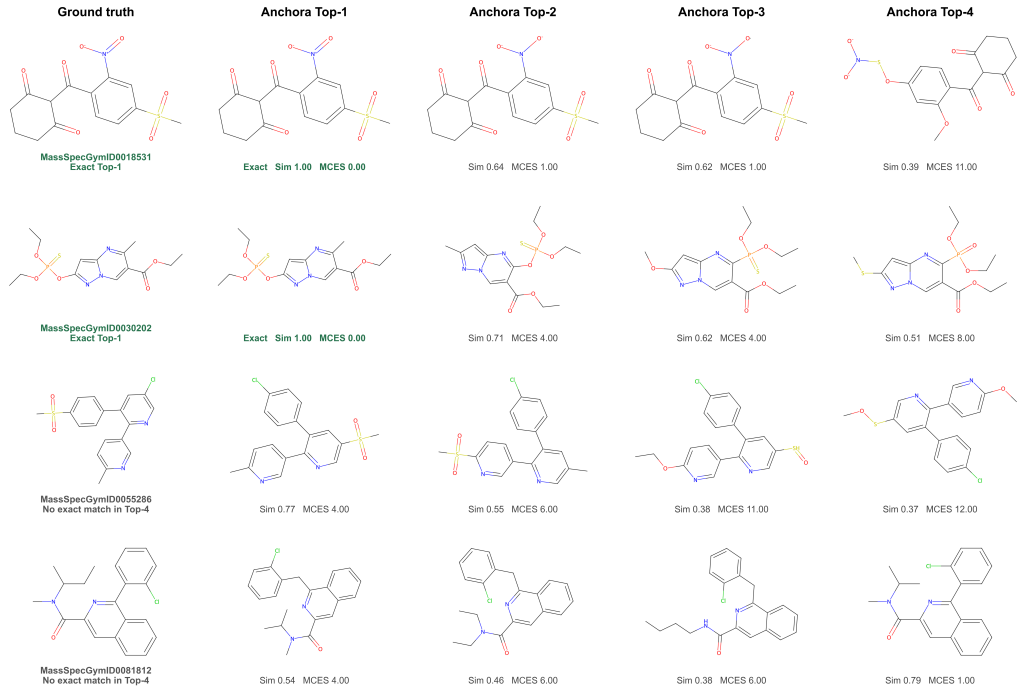

Figure 5: **Anchora predictions on MassSpecGym test samples.** Ground-truth molecules are shown in the left column and the Top-1 to Top-4 predictions in the remaining columns. Tanimoto similarity (Sim) and MCES are listed below every prediction using the definitions in Table 1. The first two rows are exact Top-1 recoveries; the bottom two rows contain no exact match among the first four predictions. These examples visualize representative successes and errors rather than aggregate performance.