GEMS: MOLECULAR STRUCTURE ELUCIDATION AS LEARNED SEARCH OVER ISOMER GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying small-molecule structures from molecular formulas and tandem mass spectra remains a bottleneck in untargeted metabolomics and natural-product discovery. Even when database retrieval or de novo prediction does not identify the exact molecular structure, the resulting candidates can provide starting points for exploring nearby structures through a few bond edits. We introduce GEMS (Guided Editing of Molecular Structures), which refines candidates through learned 2-switch bond edits that preserve the molecular formula and each atom's total valence. Molecular fingerprints encode local structural features and can be computed directly from candidate structures. GEMS uses MIST, a pretrained spectrum-to-fingerprint model, to predict fingerprint probabilities once per spectrum. A neural policy compares the candidate fingerprint with these predictions, links present features to the atoms and bonds that generate them, and scores all valid edits in one forward pass, allowing each evaluation to propose multiple candidate molecules. Shortest-path supervision enables pretraining on molecular structures alone, followed by fine-tuning with spectrum-predicted fingerprints. At inference, GEMS searches from multiple initial candidates matching the supplied molecular formula and ranks candidates by fingerprint agreement and multi-start consensus. With the target structure excluded from PubChem initialization, three edit rounds achieve Hit@1/Hit@10 of 29.64%/50.81% on NPLIB1 and 30.22%/41.33% on MassSpecGym, exceeding published baselines under comparable settings. On NPLIB1, GEMS achieves higher identification accuracy in about 3.3 minutes end to end, compared with 26.2 hours for our FRIGID run. GEMS also improves identification from generated candidates, including those refined by FRIGID. Experiments with exact target fingerprints identify fingerprint prediction error as a bottleneck in target recovery and final ranking. Deeper search recovers additional targets without retraining, with larger gains from weaker initial candidates.

1 INTRODUCTION

Identifying unknown small molecules is a central challenge in untargeted metabolomics and natural-product discovery. Many measured spectra remain unassigned to structures, limiting what experiments reveal about a sample's chemistry (Bushuiev et al., 2024). Tandem mass spectrometry (MS/MS) measures the mass-to-charge ratios and intensities of fragment ions, providing clues about substructures whose arrangement can remain ambiguous. We study the known-formula setting, which fixes elemental composition while leaving connectivity to be determined. Even then, many isomers can be consistent with the observed spectrum.

Database retrieval ranks existing structures, so success requires the target to be in the database (Dührkop et al., 2015; Goldman et al., 2023). Generative methods can propose new structures, but recovering exact connectivity remains difficult (Stravs et al., 2022; Bohde et al., 2025; Zhao et al., 2026). Yet even when these methods miss the target, their candidates are often only a few bond edits away. For each NPLIB1 spectrum, we remove the target from PubChem and retrieve up to five isomers whose fingerprints best match the one predicted from the spectrum. For at least 72.85% of spectra, one of these isomers lies within three bond edits of the target (Appendix I.2). FRIGID, a recent de novo method (Bohde et al., 2026), shows a similar pattern in its initial candidate pool: among spectra with no correct selected starting candidate, at least 47.79% have a starting candidate within three bond edits of the target (Appendix H.3).

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

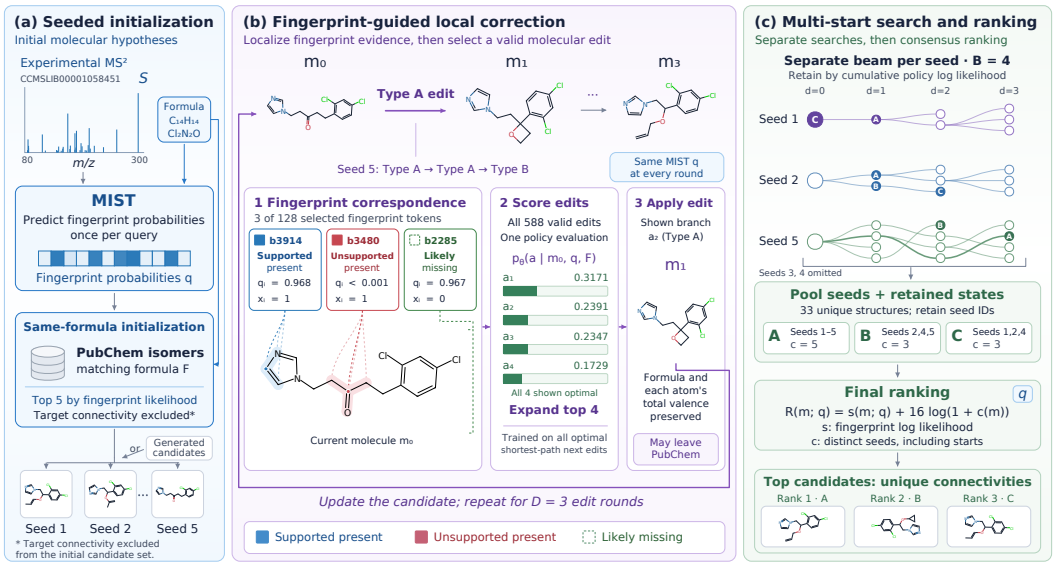


Figure 1: **GEMS: fingerprint-guided search over molecular isomers.** (a) MIST predicts fingerprint probabilities q once from the MS^2 spectrum and molecular formula. Fingerprint likelihood selects five same-formula seeds, excluding the target. (b) Supported present bits have $x_i = 1$ and high q_i, unsupported present bits have $x_i = 1$ and low q_i, and likely missing bits have $x_i = 0$ and high q_i. Present bits are mapped to the atoms and bonds that generate them, while missing bits have no local occurrence. One policy evaluation scores all valid 2-switch edits. For Seed 5, the four highest-ranked edits (green) receive 96.4% of policy probability, and each begins a shortest three-edit path to the reference structure. The illustrated path starts with a_2. (c) Each seed has a separate beam. Initial and retained states are pooled and ranked by $s(m; q) + 16 \log(1 + c(m))$, where $s(m; q)$ measures fingerprint agreement and $c(m)$ counts the distinct seeds whose searches reach m, including through their initial states. The target ranks first among 33 distinct candidate structures. Panel (c) shows Seeds 1, 2, and 5. NPLIB1 query `CCMSLIB00001058451`. The example uses $K = 5$ starts, proposal and beam widths $P = B = 4$, and depth $D = 3$. The main experiments use $P = B = 16$.

These near misses suggest correcting candidates locally rather than discarding them. Two recent methods do this with ICEBERG, a structure-to-spectrum model (Goldman et al., 2024). FRIGID uses discrepancies between a candidate's predicted fragments and the experimental spectrum to set masking probabilities through a fixed rule. Its learned diffusion model then regenerates the masked regions from the predicted fingerprint and formula, without explicitly receiving the discrepancy information (Bohde et al., 2026). FOAM proposes random formula-preserving mutations and crossovers, and uses predicted spectra, together with a structural-complexity penalty, only to select candidates (Manjrekar et al., 2026). We instead supply candidate–fingerprint correspondences directly to a learned policy, which learns to select bond edits from this evidence without simulating spectra.

We introduce GEMS (Guided Editing of Molecular Structures), which formulates structure elucidation as learned search over molecular isomers (Figure 1). Each search step applies a 2-switch edit, which removes one bond unit from each of two bonds and reconnects their endpoints. One bond may be to a hydrogen. These edits preserve the molecular formula and each atom's total valence. MIST, a pretrained spectrum-to-fingerprint model, predicts the target fingerprint probabilities once per spectrum (Goldman et al., 2023). The policy compares these probabilities with the candidate's fingerprint and links the candidate's present features to their generating atoms and bonds. A present feature with low predicted probability, for example, points to where an edit may help. A candidate can admit hundreds or thousands of valid edits, all scored in one forward pass without fingerprinting every possible product. GEMS runs beam searches from several initial candidates, then ranks the retained molecules by fingerprint agreement and by how many independent searches reach them.

We train the policy on shortest edit paths between molecular structures. Several edits can be equally good first steps toward the target, so rather than picking one as the label, we maximize the total probability of all edits that begin a shortest path. Both this supervision and the exact target fingerprint

can be computed from structures alone, so we pretrain the policy on pairs of PubChem isomers without any spectra, then fine-tune it on benchmark training spectra with MIST-predicted fingerprints. Because the best next edit depends only on the current molecule and the target, the policy is Markovian: its inputs, including the fingerprint correspondence, are computed from the current molecule and predicted fingerprint, with no trajectory history or depth index. The same policy can therefore search at greater depths and refine retrieved or generated candidates without retraining or changing the generator architecture.

Our contributions are: **(1)** We formulate known-formula structure elucidation as search over an isomer graph whose edges are formula- and valence-preserving 2-switch edits, allowing retrieved or generated candidates to be corrected locally. Because one close start is enough, we search from several starts, and even a simple Enumerate & Score (E&S) search, which scores every valid one-edit product by fingerprint agreement without a learned policy, matches or exceeds the compared published baselines on the same splits in Hit@1 and Hit@10. **(2)** We introduce GEMS, whose navigation policy links fingerprint evidence to candidate atoms and bonds and scores all valid edits in one forward pass. Even from PubChem starts that exclude the target and every isomer with Tanimoto similarity of at least 0.7 to it, GEMS exceeds all Table 1 baselines in Hit@1 and Hit@10 on both datasets. It ranks more targets first than E&S at matched width four on both benchmarks, and at width sixteen it outperforms E&S at width four with 13–21-fold lower reported runtime. From FRIGID's initial NPLIB1 candidates, without database retrieval, GEMS also outperforms FRIGID's refinement of the same pool. **(3)** We construct supervision from all shortest edit paths. It needs only structures, so it serves both pretraining on PubChem and fine-tuning on predicted fingerprints. A policy with no spectral training already gives 28.89% NPLIB1 Hit@1 with MIST predictions, and in our NPLIB1 comparison this supervision outperforms REINFORCE with exact-fingerprint rewards. **(4)** We study how start quality, search width and depth, and fingerprint accuracy shape recovery and identification. Deeper search helps most from weak starts but raises recovery more than identification. Exact fingerprints improve both: exact navigation raises NPLIB1 reach from 57.78% to 72.85%, and exact reranking raises Hit@1 from 29.64% to 56.66%.

2 Related Work

Candidate ranking. CSI:FingerID and MIST predict molecular fingerprints from MS/MS spectra to rank database candidates (Dührkop et al., 2015; Goldman et al., 2023). Other approaches, including CFM-ID, ICEBERG, MassFormer, FraGNNet, GLACIER, and AIMe, use spectrum prediction to support ranking by spectral similarity (Allen et al., 2014; Goldman et al., 2024; Young et al., 2024; 2025; Wang et al., 2026; Acikalin et al., 2026). These approaches rank a supplied candidate set. GEMS uses MIST retrieval to select starting structures, then edits their connectivity to explore additional isomers.

De novo generation. MSNovelist, Spec2Mol, and MS2Mol generate molecular strings from spectra or predicted fingerprints (Stravs et al., 2022; Litsa et al., 2023; Butler et al., 2023). DiffMS and MBGen generate molecular graphs through diffusion, while FlowMS uses discrete flow matching (Bohde et al., 2025; Sun et al., 2026; Nie & Gao, 2026). All three use MIST encoders and pretrain their decoders on fingerprint–structure pairs. FRIGID-base (without ICEBERG refinement) and MS-GPT generate SAFE strings, a fragment-based molecular representation, from predicted fingerprints through masked diffusion and autoregressive decoding, respectively (Bohde et al., 2026; Zhao et al., 2026). MS-GPT handles fingerprint uncertainty through multiple calibrated binary queries. Generated structures can initialize GEMS, which explores additional isomers through learned bond edits.

Scaffold conditioning and candidate refinement. MADGEN completes a retrieved scaffold, while MSAnchor predicts a scaffold and its attachment sites to guide generation (Wang et al., 2025; Qin et al., 2026). FRIGID refines complete candidates using ICEBERG fragment predictions to select regions to remask and regenerate (Bohde et al., 2026; Goldman et al., 2024). FOAM searches from retrieved or generated structures through formula-constrained mutation and crossover, scoring candidates with ICEBERG-predicted spectra (Manjrekar et al., 2026). GEMS learns to select valid bond edits using the same MIST fingerprint prediction throughout search, without simulating candidate spectra.

Molecular graph editing. Chemical graph enumeration has a long history, from DENDRAL to Surge (Lindsay et al., 1993; McKay et al., 2022), and learned graph edits have also been used for chemical reaction prediction (Do et al., 2019). GEMS builds on classical degree-preserving 2-switches (Senior, 1951; Hakimi, 1962; Taylor, 1981; Fernandes, 2022), adapting them to molecular bonds and attached hydrogens and learning which valid edits to apply from spectral evidence.

3 Structure Elucidation as Search over Isomers

3.1 Molecular Candidates and Bond Edits

We represent each candidate by its heavy atoms, integer bond orders, and each heavy atom's hydrogen count. We consider connected, neutral molecules without isotopic labels or radicals, and evaluate connectivity with stereochemistry removed. Each search step applies a formula-preserving bond edit.

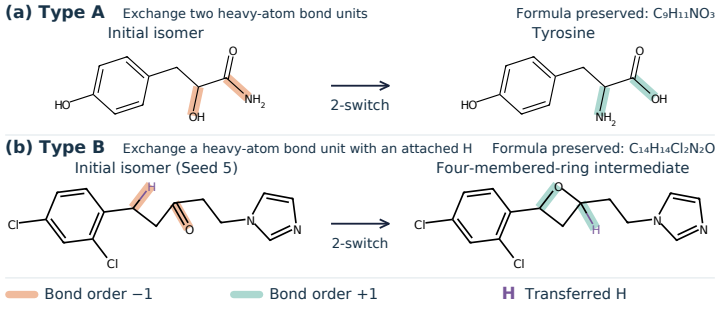


Figure 2: **Formula- and valence-preserving 2-switch edits.** (a) Type-A exchanges hydroxyl and amino attachments to reach tyrosine. (b) Type-B exchanges one carbonyl bond unit and one C–H bond for a new C–O bond and a new C–H bond, forming a four-membered ring. This is action a_1 from Seed 5 in Figure 1. Each affected atom loses and gains one bond unit. Orange and teal mark bond-order decreases and increases. Purple marks the transferred hydrogen.

We use two types of 2-switch edits (Figure 2), where a bond unit is one bond-order unit. Type-A removes one bond unit from each of two heavy-atom bonds, (u, v) and (w, z), with four distinct atoms, and adds one unit each to (u, w) and (v, z). The other pairing adds (u, z) and (v, w) instead.

Type-B uses the same switch with a hydrogen endpoint. Let u, v, and w be distinct heavy atoms, with a hydrogen on u and a bond between v and w. The edit removes the u–H bond, reduces the v–w bond order by one, adds a u–v bond unit, and attaches the hydrogen to w. Exchanging v and w gives the other pairing. For example, keto–enol interconversion can be represented by one Type-B edit.

In both edit types, every affected atom loses and gains one bond unit. The edits therefore preserve the formula and each atom's total bond order, including attached hydrogens. We retain products with nonnegative integer bond orders up to three, nonnegative attached-hydrogen counts, and a connected heavy-atom graph. Appendix A.1 gives the matrix representation and invariants.

3.2 Search over Isomers

Applying a valid edit produces another candidate with the same molecular formula. These transitions define an implicit isomer graph $\mathcal{G}_{\mathcal{F}} = (\mathcal{V}_{\mathcal{F}}, \mathcal{E}_{\mathcal{F}})$. Each node is a molecular state as defined in Section 3.1 with formula $\mathcal{F}$, and an edge corresponds to a valid edit between two states. A sequence of edits traces a path through this graph. The edits also preserve each atom's total valence, so sharing a molecular formula does not guarantee that two states can be connected by a path of valid edits.

We enumerate outgoing edits only when expanding a state. For E heavy-atom bond units and n heavy atoms, there are $O(E^2)$ potential Type-A and $O(nE)$ potential Type-B edits, each checked for validity. Rapid neighborhood growth (Appendix K.3) motivates learned beam search that selects edits before fingerprinting their products.

3.3 Candidate Recovery and Ranking

Search starts from initial isomers and returns a pool of candidate structures. It *reaches* the target when the pool contains a structure with the target's InChIKey14, the first 14 characters of the InChIKey. GEMS can fail at three stages. Initialization may supply no candidate from which the target is reachable within the search depth. Search may discard a path to the target. Ranking may place other candidates above a recovered target. We examine initialization and depth in Section 5.5, and fingerprint errors in navigation and ranking in Section 5.6.

4 GEMS

At each step, the candidate molecule specifies present local structural features and their locations. Fingerprint probabilities predicted from the experimental spectrum indicate the unknown target's likely features. The navigation policy uses their agreement and discrepancies to choose bond edits.

Given a spectrum S and molecular formula $\mathcal{F}$, GEMS uses MIST, a frozen spectrum-to-fingerprint model, to predict $q \in [0,1]^{4096}$. These probabilities are computed once and remain fixed throughout search. GEMS selects initial isomers, searches from each independently, and ranks the retained candidates by fingerprint agreement and support from distinct starts:

$$\begin{aligned} q &= g_{\mathrm{MIST}}(S, \mathcal{F}), \\ \mathcal{M}_0 &= \mathcal{I}(q, \mathcal{F}), \\ \mathcal{C}_\theta &= \mathrm{Search}_\theta(q, \mathcal{M}_0, \mathcal{G}_\mathcal{F}), \\ \widehat{m} &= \arg\max_{m \in \mathcal{C}_\theta} R(m; q). \end{aligned} \tag{1}$$

4.1 Initialization

GEMS starts from chemically valid molecules with formula $\mathcal{F}$. These can come from database retrieval or a de novo generator such as FRIGID (Bohde et al., 2026) or MS-GPT (Zhao et al., 2026). Our main configuration selects initial PubChem isomers by agreement with the predicted fingerprint.

For a candidate fingerprint $x(m) \in \{0,1\}^{4096}$, we measure agreement with q using the independent-bit Bernoulli log likelihood

$$s(m; q) = \sum_i \big[x_i(m) \log \widetilde{q}_i + (1 - x_i(m)) \log(1 - \widetilde{q}_i)\big], \tag{2}$$

where $\widetilde{q}_i = \mathrm{clip}(q_i, 10^{-7}, 1 - 10^{-7})$. This score is used for both retrieval and final ranking.

We rank PubChem molecules with formula $\mathcal{F}$ by $s(m; q)$, retain the highest-scoring representative of each InChIKey14 group, and select up to $K = 5$ initial structures. In target-excluded evaluation, we remove the target InChIKey14 group before selecting the starts. A correct prediction therefore requires at least one edit to a retrieved structure.

4.2 Candidate–Fingerprint Correspondence

GEMS compares the candidate's radius-two, 4,096-bit Morgan fingerprint $x(m)$ (Rogers & Hahn, 2010) with the spectrum-predicted probabilities q (Figure 1b). *Supported present bits* occur in the candidate and have high q_i, while *unsupported present bits* occur in the candidate and have low q_i. *Likely missing bits* are absent from the candidate but have high q_i. Together, they suggest which features to preserve, change, or introduce.

An atom environment comprises a center atom and surrounding atoms and bonds within zero, one, or two bond steps. For each present fingerprint bit, we retain all environments that generate it, linking the bit to the candidate's atoms and bonds. In Figure 1(b), bit 3480 is present in Seed 5 and corresponds to a radius-one environment centered on its carbonyl carbon, but has predicted probability below 0.001. The policy uses this localized evidence to score edits. Missing bits have no corresponding environment in the candidate, but still provide target evidence through their fingerprint tokens.

The policy receives 128 fingerprint tokens, initially allocated as 48 likely missing, 48 unsupported present, and 32 supported present bits (Appendix C). Each encodes bit identity, role, q_i, x_i, and uncertainty and discrepancy features.

4.3 Fingerprint-Conditioned Navigation Policy

A transformer encodes the fingerprint tokens, and a Graphormer-style molecular encoder (Ying et al., 2021) processes the candidate. Atom representations attend to fingerprint tokens with learned biases distinguishing environment centers, other members, and atoms outside the environment. Directed-bond representations are then constructed from atom representations and bond features. They attend to fingerprint tokens using separate biases that describe whether a bond and its endpoints belong to an environment. These connections let the policy associate predicted fingerprint support with the atoms and bonds involved in an edit. Appendix D.2 gives the layer arrangement.

Each valid edit is scored from the representations of the atoms and bonds it changes. We concatenate the relevant representations with action-type and global molecular features and combine them using multilayer perceptrons. The same edit can be written with its endpoints in different orders. To make its score independent of these equivalent orderings, we average their action representations before projecting to a logit. A softmax over the valid edit set $\mathcal{A}(m)$ gives

$$p_\theta(a \mid m, q, \mathcal{F}) = \mathrm{softmax}_{a \in \mathcal{A}(m)} \, \ell_\theta(a; m, q, \mathcal{F}). \tag{3}$$

One forward pass scores all valid edits. Product molecules are constructed and fingerprinted only for actions selected by search. The Markovian policy conditions on the current molecule, fingerprint probabilities, and formula, without trajectory history or a depth index.

4.4 Training the Navigation Policy

During training, the chosen target graph t is known. We identify edits that move a source molecule m one step closer to t along a shortest valid path. Several edits may be equally good: if three edits are required, every first edit that leaves a two-edit path to t is optimal. Let $\mathcal{A}^+(m; t)$ contain all such optimal first edits. We maximize their total probability by minimizing

$$\mathcal{L}_{\mathrm{nav}} = -\log \sum_{a \in \mathcal{A}^+(m;t)} p_\theta(a \mid m, q, \mathcal{F}). \tag{4}$$

This objective does not require choosing a single optimal edit as the label. In Figure 1(b), Seed 5's four highest-ranked edits each begin a shortest three-edit path and jointly receive 96.4% of policy probability. Dk denotes sources whose shortest valid path to the chosen target graph t has k edits.[1]

Structure-only pretraining first teaches the policy to navigate with exact target fingerprints. Fine-tuning then adapts it to fingerprints predicted by a frozen MIST encoder from experimental spectra.

PubChem pretraining. We mine PubChem isomer pairs one or two edits apart and keep all optimal first edits for each. The input is the exact target fingerprint, $q = x(t)$, so no spectra are needed. If each source had a single target, the policy could learn a molecule's typical edits and ignore the fingerprint. Half the examples therefore come in groups that share a source but have different targets, preferably with different optimal edits, encouraging the policy to use the target fingerprint. The other half selects one pair per sampled source group to cover more formulas and molecules. Both benchmark models start from one shared checkpoint, selected on validation after 37.5 million examples (Appendix D).

Fine-tuning with predicted fingerprints. Fine-tuning pairs benchmark training targets with same-formula PubChem sources one to three edits away, using fingerprint probabilities predicted from the target's spectrum by the frozen MIST encoder. It varies the source for a fixed target, complementing the fixed-source, varying-target groups in pretraining. We find shortest paths by meeting in the middle: the target is expanded by up to two edits, each source by one, and the neighborhoods are matched. Targets above 35 heavy atoms use D1 and D2 only, because two-step expansion is costly. We keep all available D1 and D2 observations and a capped number of D3 sources per spectrum, selected by MIST Bernoulli agreement to approximate how starting candidates are selected at inference (Appendix B).

[1] Training paths exclude held-out structures and match the chosen target graph up to atom renumbering. Evaluation accepts any structure with the target's InChIKey14, so evaluation distances can be shorter (Appendix A.2).

4.5 Multi-Start Policy Beam Search

GEMS runs an independent beam search from each of K initial structures, with proposal width P, beam width B, and maximum depth D. Our main setting uses $K = 5$, $P = 16$, $B = 16$, and $D = 3$. For a trajectory $\pi_d = (m_0, a_1, m_1, \ldots, a_d, m_d)$, the score is the cumulative policy log probability:

$$L_\theta(\pi_d; q, \mathcal{F}) = \sum_{t=1}^{d} \log p_\theta(a_t \mid m_{t-1}, q, \mathcal{F}). \tag{5}$$

At each depth, the policy scores all valid edits from each beam state. We construct product molecules and compute their fingerprints only for the top P selected actions. Within each start's search, previously visited products are removed, and duplicate products retain their highest-scoring trajectory. The top B trajectories under L_θ form the next beam. The visited set records all proposed products, including those outside the next beam (Algorithm 1, Appendix E).

The final candidate pool contains the initial structures and beam states retained at every depth across all starts. With our main setting, search uses at most $K[1 + B(D - 1)] = 165$ state-level policy evaluations and retains at most $K(1 + BD) = 245$ states before final deduplication.

4.6 Final Ranking by Fingerprint Evidence and Consensus

Searches from different seeds can reach the same molecule. This agreement is an additional ranking signal. Let $c(m)$ count the distinct seeds whose retained search pools contain m, including the seeds themselves. Each seed contributes at most one vote. In Figure 1(c), candidate A receives five votes and candidate C receives three, including a vote because C is Seed 1's initial structure. Support merges atom renumberings and alternative Kekulé assignments while keeping tautomers distinct (Appendix A.3).

The final score is $R(m; q) = s(m; q) + \lambda \log(1 + c(m))$, where $s(m; q)$ measures fingerprint agreement and λ weights consensus. We use $\lambda = 16$ for both benchmarks, selected on MassSpecGym validation. We retain each InChIKey14 group's highest-scoring candidate and rank these representatives by $R(m; q)$.

5 Experiments

5.1 Setup and Evaluation

Datasets and metrics. We use the same NPLIB1 split as FRIGID, with 803 protonated test spectra (Dührkop et al., 2021; Bohde et al., 2026), and the structurally held-out MassSpecGym split with all 17,556 test spectra, 14,066 protonated and 3,490 sodiated (Bushuiev et al., 2024). One MassSpecGym policy serves both adducts. Hit@k measures correctness among the first k ranked connectivities. All queries contribute to accuracy and reach, including 17 NPLIB1 and 200 MassSpecGym queries without starts. MassSpecGym hit counts are pooled across adducts.

Comparisons. Published baselines come from FRIGID Table 1 and *MassSpecGym in the Wild* Table 9 (Bohde et al., 2026; Liu et al., 2026). Appendix F.2 reports MS-GPT's published results and our best-effort reproduction with corrected masking and retraining on the same benchmark splits.

5.2 Main Results

GEMS achieves the highest Hit@1 and Hit@10 in Table 1 on both benchmarks. Relative to the strongest baseline, FRIGID on NPLIB1 and our MS-GPT reproduction on MassSpecGym, the gains are 4.61/17.44 and 6.99/13.63 percentage points. Because PubChem starts exclude the target connectivity, every hit requires at least one edit.

GEMS also has the highest Top-10 Tanimoto similarity among these methods: 0.71 versus 0.64 for FRIGID on NPLIB1, and 0.61 versus 0.53 (FOAM) on MassSpecGym. MCES results are mixed across benchmarks and ranks, with full comparisons in Appendix F.2.

With starts of Tanimoto similarity ≥ 0.7 to the target also excluded, GEMS still exceeds every baseline in Table 1 in Hit@1 and Hit@10 on both benchmarks (Appendix H.2).

Table 1: Known-formula structure elucidation (H@k in %). GEMS rows use $K = 5$, $P = B = 16$, $D = 3$, and $\lambda = 16$. PubChem starts exclude the target connectivity, and Tani. < 0.7 also excludes starts with target similarity ≥ 0.7 (Appendix H.2). Generator $\rightarrow$ GEMS starts from MIST-selected outputs, which may contain the target. †: our same-split reproduction. Bold marks the best value per dataset. Lower maximum common edge subgraph (MCES) distance and higher Tanimoto similarity are better. Empty outputs receive MCES 100. Baselines are from Liu et al. (2026) and Bohde et al. (2026).

Method	H@1	MCES@1	Tani.@1	H@10	MCES@10	Tani.@10
NPLIB1						
MIST + MolForge	2.24	14.16	0.40	5.11	10.96	0.47
DiffMS	8.34	10.46	0.35	15.44	8.54	0.47
FRIGID-base	19.80	7.88	0.54	24.41	6.20	0.61
FRIGID	25.03	**7.10**	0.58	33.37	**5.56**	0.64
MS-GPT†	21.05	18.49	0.50	28.64	17.79	0.54
MS-GPT† $\rightarrow$ GEMS	26.28	20.61	0.52	42.22	19.54	0.60
FRIGID-base $\rightarrow$ GEMS	28.02	11.38	0.56	43.34	10.23	0.64
GEMS (PubChem, Tani. < 0.7)	27.90	8.59	0.56	46.45	6.97	0.67
GEMS	**29.64**	7.73	**0.59**	**50.81**	6.08	**0.71**
MassSpecGym						
DiffMS	2.30	13.96	0.28	4.25	11.68	0.39
FOAM (DiffMS + PubChem)	1.50	12.21	0.35	10.28	**6.06**	0.53
MIST + MolForge (corrected)	10.73	22.15	0.37	14.48	17.88	0.41
FRIGID-base	16.09	15.02	0.41	18.19	11.95	0.46
FRIGID	18.29	13.49	0.43	22.00	11.65	0.47
MS-GPT†	23.23	25.15	0.48	27.70	24.71	0.50
MS-GPT† $\rightarrow$ GEMS	25.56	26.66	0.48	34.07	25.91	0.52
GEMS (PubChem, Tani. < 0.7)	23.93	10.95	0.50	31.75	9.61	0.55
GEMS	**30.22**	**9.68**	**0.55**	**41.33**	8.26	**0.61**

5.3 Refining Generator Initializations

GEMS refines generated candidates without retraining (Table 10). From the NPLIB1 FRIGID-base pool, it raises Hit@1/Hit@10 of the selected starts from 20.42%/23.91% to 28.02%/43.34%, above 24.03%/33.13% for our FRIGID run after 24 ICEBERG refinement rounds, and those of FRIGID's final pool from 22.54%/30.14% to 28.27%/46.82%. For our MS-GPT reproduction, it raises Hit@1 from 21.05% to 26.28% on NPLIB1 and from 23.23% to 25.56% on MassSpecGym (Appendix H.3).

5.4 Learned Proposals Enable Wider Search

We compare learned proposals with Enumerate & Score (E&S) from the same five MIST-ranked, target-excluded starts. E&S constructs and fingerprints every valid one-edit neighbor, retaining the best B trajectories per start by fingerprint log likelihood. Both methods search three edits deep and rank retained candidates with $\lambda = 16$. We test E&S widths $B \in \{1, 2, 4\}$ and GEMS widths $P = B \in \{4, 16\}$ (Table 6). Every tested E&S width exceeds the published baselines in Table 1 in Hit@1 and Hit@10 on both benchmarks.

At width sixteen, GEMS exceeds all tested E&S widths in Hit@1, Hit@10, and reach on both benchmarks. Against E&S $B = 4$, NPLIB1 Hit@1/Hit@10 rises from 26.03%/49.07% to 29.64%/50.81%, and reach from 54.17% to 57.78%. MassSpecGym Hit@1 rises from 28.16% to 30.22% and reach from 44.24% to 46.88%. Learned proposals make the wider beam practical: GEMS takes 200.12 s on NPLIB1 versus 2,524.59 s for E&S ($B = 4$), a 12.62-fold difference. The MassSpecGym ratio is 21.36 (Appendix K).

Random-weight policies achieve $4.61 \pm 2.90\%$ NPLIB1 Hit@1 under the same proposal and beam budget (Appendix G.1).

5.5 Initialization, Search Width, and Depth

At depth three, up to five MIST-ranked starts from the same-formula PubChem pool reach 57.78% of NPLIB1 targets. Ranking within a random 250-isomer subset reaches 50.44%, and using five random

starts reaches 22.17% (Table 7). At width sixteen, extending search from three to ten rounds adds 18.68 points of reach from up to five random starts but 2.12 points from full-pool MIST-ranked starts.

We next vary width and depth from up to five MIST-ranked starts or up to 100 random starts, keeping the checkpoint and each starting set fixed (Figure 4, Appendix I.1). From $P = B = D = 3$, widening to $P = B = 10$ at $D = 3$ and deepening to $D = 10$ at $P = B = 3$ each retain at most 30 search states plus the initial structure per start. Under MIST guidance, widening gives higher Hit@1 than deepening from ranked starts (29.89% versus 29.02%), while deepening gives higher Hit@1 from up to 100 random starts (29.39% versus 25.90%). With exact fingerprints, deepening helps more from both starting sets. Hit@10 and per-seed reach follow the same ordering. The relative benefits of width and depth therefore depend on initialization and fingerprint evidence.

From up to five MIST-ranked starts, most hits need one edit: the second and third add 2.37 and 0.25 Hit@1 points on NPLIB1 and 0.34 and -0.21 on MassSpecGym. Under one-edit exclusion, every hit needs two or three edits, yet GEMS still reaches 25.53% and 22.44% Hit@1 (Appendix H.2).

5.6 Fingerprint Errors Limit Search and Ranking

MIST fingerprints guide retrieval, navigation, and ranking. We keep the MIST-selected starts and trained checkpoint fixed and replace the navigation fingerprint, ranking fingerprint, or both with the exact target fingerprint. A ranking-only substitution preserves the retained pool and consensus counts.

On fixed candidate pools, exact-fingerprint ranking raises Hit@1 from 29.64% to 56.66% on NPLIB1 and from 30.22% to 45.12% on MassSpecGym (Figure 3), placing the target first for nearly all recovered queries. Exact navigation raises reach from 57.78% to 72.85% and from 46.88% to 56.07%, respectively. With MIST ranking, however, Hit@1 rises only to 33.25% and 33.82%. Fingerprint prediction error thus limits both recovery and final ranking (Appendix J). Reranking the top ten with an ICEBERG spectral score adds 2.24 and 1.38 Hit@1 points on the two benchmarks (Appendix J.3).

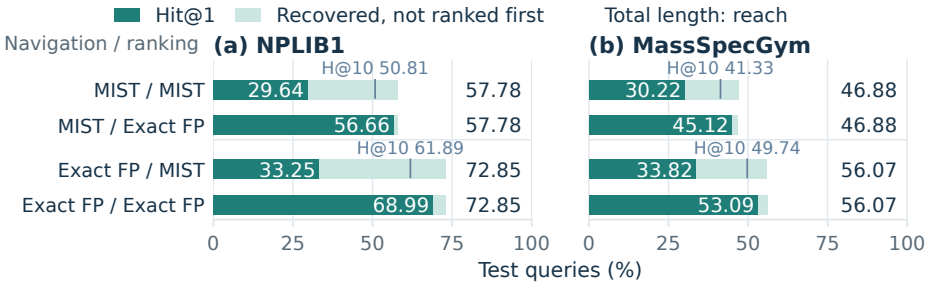


Figure 3: **Finding the target and ranking it first.** Dark bars show Hit@1, total bar length gives reach, and blue markers show Hit@10 for MIST-ranking rows. Labels specify navigation/ranking fingerprints. Exact FP is computed from the target as an oracle. Each dataset uses fixed starts and checkpoint, $K = 5$, $P = B = 16$, $D = 3$, and $\lambda = 16$. All 803 NPLIB1 and 17,556 MassSpecGym queries count, including queries without starts. Table 15 gives the values.

6 Discussion and Limitations

GEMS shows that local correction of complete candidates is effective for known-formula structure elucidation. Needing only same-formula structures and a predicted fingerprint, one policy refines retrieved candidates and outputs of generators with different architectures. It makes wider search practical and remains useful at greater depths with weak starts. The exact-fingerprint interventions point to better spectrum-to-fingerprint prediction as a route to better recovery and ranking.

The method assumes a known formula and evaluates connectivity without stereochemistry. Formula uncertainty would require searching and ranking across multiple formula hypotheses. The policy reflects its PubChem training data, PubChem initialization depends on available same-formula candidates, and valence-preserving edits restrict the states reachable from each start.

Ethics Statement

This work develops computational methods for molecular annotation in metabolomics and natural-product research. More accurate structure proposals may accelerate discovery, but predictions require experimental validation and should not be the sole evidence for safety-critical chemical identification.

AI Use Statement

Generative AI assisted implementation and debugging, experiment planning and coordination, analysis, literature search, and drafting and editing manuscript text. The authors tested the code, checked numerical results against experiment outputs and bibliographic claims against primary sources, and take responsibility for the final content.

Reproducibility Statement

Appendices A–E specify molecular states, supervision construction, model inputs, training, and search. Appendices F–K give full results, evaluation conventions, and timing boundaries. Upon publication, we will publicly release our code, trained GEMS checkpoints, and supporting artifacts, including experiment configurations, starting-candidate lists, and data-preparation scripts, to enable reproduction of the reported results.

References

Utku U. Acikalin, Dieqiao Feng, Aaron M. Ferber, Goncalo J. Gouveia, Tyler J. Schwertfeger, Di Chen, Delia Qu, Marissa A. Fontaine, Yingheng Wang, Richard A. Bernstein, Haofan Wang, Tae-Hyung Won, Christopher N. Parkhurst, Bart Selman, David Artis, Frank C. Schroeder, and Carla P. Gomes. Charting the small-molecule universe from mass spectra with neuro-symbolic AI. *bioRxiv*, 2026. doi: 10.64898/2026.08.05.743095. URL https://www.biorxiv.org/content/10.64898/2026.08.05.743095v1. Preprint, version 1.

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.

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. 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, pp. 110010–110027. Curran Associates, Inc., 2024. doi: 10.52202/079017-3491.

Thomas Butler, Abraham Frandsen, Rose Lightheart, Brian Bargh, Thomas Kerby, Kiana West, Joseph Davison, James Taylor, Christoph Krettler, TJ Bollerman, Gennady Voronov, Kevin Moon, Tobias Kind, Pieter Dorrestein, August Allen, Viswa Colluru, and David Healey. MS2Mol: A transformer model for illuminating dark chemical space from mass spectra. *ChemRxiv*, 2023. doi: 10.26434/chemrxiv-2023-vsmpx-v4. Preprint, version 4.

Kien Do, Truyen Tran, and Svetha Venkatesh. Graph transformation policy network for chemical reaction prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 750–760. Association for Computing Machinery, 2019. doi: 10.1145/3292500.3330958.

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, Louis-Félix Nothias, Markus Fleischauer, Raphael Reher, Marcus Ludwig, Martin 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(4):462–471, 2021. doi: 10.1038/s41587-020-0740-8.

Rosário Fernandes. On the switch-length of two connected graphs with the same degree sequence. *Australasian Journal of Combinatorics*, 83(1):87–100, 2022. URL https://ajc.maths.uq.edu.au/pdf/83/ajc_v83_p087.pdf.

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.

Samuel Goldman, Janet Li, and Connor W. Coley. Generating molecular fragmentation graphs with autoregressive neural networks. *Analytical Chemistry*, 96(8):3419–3428, 2024. doi: 10.1021/acs.analchem.3c04654.

S. L. Hakimi. On realizability of a set of integers as degrees of the vertices of a linear graph. I. *Journal of the Society for Industrial and Applied Mathematics*, 10(3):496–506, 1962. doi: 10.1137/0110037.

Robert K. Lindsay, Bruce G. Buchanan, Edward A. Feigenbaum, and Joshua Lederberg. DENDRAL: A case study of the first expert system for scientific hypothesis formation. *Artificial Intelligence*, 61(2):209–261, 1993. doi: 10.1016/0004-3702(93)90068-M.

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*, 6:132, 2023. doi: 10.1038/s42004-023-00932-3.

Hongxuan Liu, Roman Bushuiev, Ivy Lightheart, Mrunali Manjrekar, Anton Bushuiev, Magdalena Lederbauer, Filip Jozefov, Yinkai Wang, Soha Hassoun, Josef Sivic, James Taylor, Runzhong Wang, David Healey, Tomáš Pluskal, and Connor W. Coley. MassSpecGym in the wild: Uncovering and correcting evaluation pitfalls in AI-driven molecule discovery. *arXiv preprint arXiv:2606.19624*, 2026. URL https://arxiv.org/abs/2606.19624.

Mrunali Manjrekar, Runzhong Wang, Samuel Goldman, Jenna C. Fromer, and Connor W. Coley. Generative structural elucidation from mass spectra as an iterative optimization problem. *arXiv preprint arXiv:2602.07709*, 2026. URL https://arxiv.org/abs/2602.07709.

Brendan D. McKay, Mehmet Aziz Yirik, and Christoph Steinbeck. Surge: A fast open-source chemical graph generator. *Journal of Cheminformatics*, 14(1):24, 2022. doi: 10.1186/s13321-022-00604-9.

Jianan Nie and Peng Gao. FlowMS: Flow matching for de novo structure elucidation from mass spectra. *arXiv preprint arXiv:2603.18397*, 2026. URL https://arxiv.org/abs/2603.18397.

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.

James K. Senior. Partitions and their representative graphs. *American Journal of Mathematics*, 73(3): 663–689, 1951. doi: 10.2307/2372318.

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.

Xichen Sun, Wentao Wei, Jiahua Rao, Jiancong Xie, and Yuedong Yang. De novo molecular generation from mass spectra via many-body enhanced diffusion. *arXiv preprint arXiv:2602.01643*, 2026. URL https://arxiv.org/abs/2602.01643.

R. Taylor. Constrained switchings in graphs. In Kevin L. McAvaney (ed.), *Combinatorial Mathematics VIII*, volume 884 of *Lecture Notes in Mathematics*, pp. 314–336. Springer, 1981. doi: 10.1007/BFb0091828.

Rui-Xi Wang, Runzhong Wang, and Connor W. Coley. GLACIER: Rethinking mass spectrum prediction as an object detection problem. *arXiv preprint arXiv:2606.29161*, 2026. doi: 10.48550/arXiv.2606.29161. URL https://arxiv.org/abs/2606.29161.

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.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do Transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems*, volume 34, pp. 28877–28888. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f1c1592588411002af340cbaedd6fc33-Abstract.html.

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

Adamo Young, Fei Wang, David S. Wishart, Bo Wang, Russell Greiner, and Hannes Röst. FraGNNet: A deep probabilistic model for tandem mass spectrum prediction. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=UsqeHx9Mbx.

Xin Zhao, Yumin Liu, Zhuo Li, Weichu Zheng, Feng Zhu, Xiaokang Yang, Yaohui Jin, and Yanyan Xu. MS-GPT: Rethinking MS/MS de novo structure elucidation as spectrum-induced posterior querying of a molecule-language model. *arXiv preprint arXiv:2607.23607*, 2026. URL https://arxiv.org/abs/2607.23607.

A Molecular States and Target Equivalence

A.1 Molecular Representation and Transition Invariants

A molecular state contains an ordered tuple of heavy-atom symbols and an integer adjacency matrix $\mathbf{A}$. For $i \neq j$, $A_{ij} \in \{0, 1, 2, 3\}$ gives the bond multiplicity between heavy atoms i and j. The diagonal entry A_{ii} gives the number of hydrogens attached to atom i. Atom indices specify the endpoints affected by an individual edit. Every retained state is connected, neutral, non-isotopic, and non-radical. Stereochemistry is removed.

Proposition 1 (Transition invariants)**.** *For every retained transition $m' = T(m, a)$ and aligned atom i,*

$$\mathsf{form}(m') = \mathsf{form}(m), \qquad A'_{ii} + \sum_{j \neq i} A'_{ij} = A_{ii} + \sum_{j \neq i} A_{ij}. \tag{6}$$

Thus each action preserves the molecular formula and the total bond order incident to every aligned atom.

The equalities follow from the balanced bond-unit subtraction and addition in both action types. The transition filter separately enforces supported bond orders and connectedness.

A.2 Training and Evaluation Target Equivalence

Let $\mathcal{T}_{\text{train}}(t)$ contain the states isomorphic to the chosen target graph, preserving heavy-atom symbols, integer bond orders, and attached hydrogens. Atom numbering may differ, but a match must preserve the complete literal graph. The matching key preserves the chosen Kekulé assignment and tautomer. The evaluation target class is

$$\mathcal{T}_{\text{eval}}(t) = \{u \in \mathcal{V}_{\mathcal{F}} : \text{IK14}(u) = \text{IK14}(t)\}. \tag{7}$$

The training class is contained in the evaluation class.

Write $\mathcal{G}_{\mathcal{F}}^{\text{train}}$ for the edit graph restricted to states admitted by the training exclusions, including removal of held-out intermediates. For admitted sources and targets, define

$$\begin{aligned} d_{\text{train}}(s, t) &= \min_{u \in \mathcal{T}_{\text{train}}(t)} d_{\mathcal{G}_{\mathcal{F}}^{\text{train}}}(s, u), \\ d_{\text{eval}}(s, t) &= \min_{u \in \mathcal{T}_{\text{eval}}(t)} d_{\mathcal{G}_{\mathcal{F}}}(s, u). \end{aligned} \tag{8}$$

An unreachable target class has infinite distance. Every admissible training path is also an admissible evaluation path, and the evaluation goal is broader. Consequently,

$$d_{\text{eval}}(s, t) \leq d_{\text{train}}(s, t). \tag{9}$$

A source can require three edits to reach the chosen training target and only two to reach another evaluation-equivalent structure.

A.3 Path Matching and Search Deduplication

Training joins use a canonical key for the literal integer-bond graph without aromatizing or changing its Kekulé assignment. A proposed match must admit an atom mapping that preserves the complete adjacency matrix and heavy-atom symbols. Target-side actions are reversed and transported through that mapping into the source coordinates. Each resulting forward action is checked against the valid edit set, intermediate exclusions are enforced, and the final state must match the reference target graph.

Online beam search and consensus use canonical stereo-free aromatic SMILES as a structural key. This key identifies alternative Kekulé assignments and keeps tautomers distinct. Search retains one concrete state for each selected key and applies subsequent edits to that state. Retaining one representation per key can discard distinct edit continuations. InChIKey14 is used for target exclusion and connectivity evaluation. Search paths consist of valid edits between concrete molecular states.

Table 2: Identity conventions. Training exclusions additionally restrict admissible intermediate states.

Identity	Representations merged	Use
Literal integer-bond graph	Atom renumberings preserving elements, bond orders, and attached hydrogens. The chosen Kekulé assignment is retained.	Training path joins and distance labels
Canonical stereo-free aromatic SMILES	Atom renumberings and alternative Kekulé assignments. Tautomers remain distinct.	Search deduplication and consensus votes
InChIKey14	Structures assigned the same connectivity block by InChI, including some tautomeric forms	Target exclusion, recovery, and final connectivity ranking

B Supervision Construction

B.1 Identity, Validity, and Held-Out Exclusion

Supervision and online search use the same action enumeration, application, and validation code. Appendix A defines the molecular states, matching rules, and identity conventions.

Before PubChem pretraining pairs are mined, every validation and test connectivity from both evaluation benchmarks is excluded from source and target structures. Generated intermediates with an excluded connectivity are also removed.

B.2 Exact-Fingerprint PubChem Pairs

For each formula, every valid one-edit product of every remaining PubChem structure is generated and identical exact products are merged. Two PubChem structures connected directly define a directed D1 pair. Two structures whose expansions meet at a shared product define a D2 pair when no direct transition exists. All valid first actions that realize a shortest direct connection or enter a shortest shared intermediate are retained as positives.

Pairs are grouped by source molecule. Half of the sampled examples come from contrast groups containing up to four D1 or D2 targets for the same source, prioritizing different sets of optimal first edits and a mixture of distances. The other half selects one pair from each sampled group to broaden coverage across formulas and source molecules, preferring a D1 target when present and otherwise using a D2 target. Each example uses the exact target fingerprint.

The mined corpus contains 986,681,148 directed source–target examples. Pretraining samples 125 million examples with a 50:50 contrast/coverage allocation. Contrast groups contribute at most four examples and coverage groups one.

B.3 Target-Conditioned D1–D3 Supervision

A distance shell contains states at a given shortest-path distance from the target. For a fixed target t, define these shells and optimal first-action sets as

$$\begin{aligned} \mathcal{D}_k(t) &= \{m : d_{\text{train}}(m,t) = k\}, \\ \mathcal{A}^+(m;t) &= \{a \in \mathcal{A}(m) : T(m,a) \in \mathcal{D}_{k-1}(t)\}, \quad m \in \mathcal{D}_k(t),\ k \geq 1. \end{aligned} \tag{10}$$

Here d_{train} uses the target class and training exclusions defined in Appendix A.2. Thus $\mathcal{D}_0(t)$ contains the atom-renumbered versions of the reference target graph.

For the D1–D3 corpus, containing molecules with at most 35 heavy atoms, each chosen training target is expanded through two edits. States are matched by the literal bond-and-hydrogen graph up to atom renumbering. Held-out intermediates are excluded during construction, so the resulting shells are complete in the admissible training graph. Reusing these target neighborhoods avoids expanding every PubChem source through two steps.

Same-formula, target-excluded PubChem sources matching the target's D1 or D2 shell receive the corresponding distance. Every source is expanded once and its products are matched to the target shells. For a source outside the two-step neighborhood, a valid join to the D2 shell gives a three-step path after reversing the target-side path and mapping its atom indices to those of the source-side meeting state. Shorter shell membership is checked first, so the joined path certifies distance three to

the chosen training goal. Every retained path is replayed forward under the edit rules and exclusions described in Appendix A.3.

For a source $m \in \mathcal{D}_k(t)$, every valid first action entering the complete $\mathcal{D}_{k-1}(t)$ shell is positive. All such first actions are retained for each selected training pair, with a replayed path witnessing each one.

Larger molecules. For molecules with more than 35 heavy atoms, we construct a separate D1–D2 corpus to avoid the cost of exhaustive two-step target expansion. We enumerate the one-edit neighbors of each target and each same-formula PubChem source. A direct source–target connection gives D1 supervision. If no direct connection exists, a shared one-edit neighbor gives a shortest two-edit path. All source edits entering such shared neighbors are retained as positive actions, and the corresponding paths are replayed and verified. Thus, one-step expansion on each side supplies supervision at distances one and two without a target D2 shell.

Training observations. D3 sources are capped independently for each target spectrum by

$$N_{\mathrm{D3}}^{\mathrm{keep}} = \min(N_{\mathrm{D3}}, \max(50, N_{\mathrm{D2}})) \,. \tag{11}$$

The D3 sets contain 378,563 source–target pairs for NPLIB1 and 1,677,732 for MassSpecGym. After pairing these with the target spectra, the cap retains 155,625 and 6,486,972 training observations, respectively. Membership is selected by descending MIST Bernoulli agreement for each spectrum, so spectra of the same target can select different D3 sources. All available D1 and D2 observations are retained.

MIST predicts the target fingerprint from the spectrum, formula, and formula-based peak annotations. The target structure supplies the shortest-path labels. Training excludes source–target pairs at distance zero and mixes D1–D3 examples for molecules with at most 35 heavy atoms with D1–D2 examples for molecules with more than 35 heavy atoms, using the batch schedules in Appendix D.

C Fingerprint Correspondence Details

The candidate fingerprint is a binary radius-two Morgan fingerprint with 4,096 hashed bits, bond types enabled, and chirality disabled (Rogers & Hahn, 2010). Each of the 128 fingerprint tokens encodes a bit's identity, evidence type, target probability, candidate presence, and uncertainty and discrepancy features. We initially allocate 48 tokens to high-probability bits missing from the candidate, 48 to present bits with low predicted support, and 32 to present bits with high predicted support. Unfilled slots are filled from the remaining bits by confidence ranking.

The numerical features for a selected bit are

$$(q_i, x_i, q_i(1-x_i), (1-q_i)x_i, q_i x_i, q_i - x_i, h(q_i)),$$

where $h(q) = -q\log q - (1-q)\log(1-q)$ is Bernoulli entropy, with $h(0) = h(1) = 0$. The discrepancy feature is signed.

The occurrence map retains every environment that generates each present bit. Radius-zero membership contains only the center atom. Radius-one and radius-two membership expand through the candidate adjacency and include the corresponding heavy bonds. These relations connect fingerprint evidence to atom and directed-bond tokens in the policy. Atom relations distinguish centers and other members at each radius from atoms outside the environment. Bond relations record how many endpoints belong to an environment and whether the bond itself belongs. Missing bits have a separate relation code because they have no occurrence in the candidate.

D Training Protocol

D.1 MIST Encoders and Input Processing

We use the released FRIGID MIST encoders `mist_canopus.pt` for NPLIB1 and `mist_msg.pt` for MassSpecGym. Each benchmark uses one fixed MIST encoder for all training and evaluation

spectra. NPLIB1 uses public formula-only peak assignments. MassSpecGym training uses formula-only subformula assignment with 15 ppm tolerance. Protonated evaluation spectra use 20 ppm on both benchmarks. Sodiated MassSpecGym evaluation uses 15 ppm. Peak assignment and fingerprint prediction use the spectrum and formula. Sodiated inference uses the same MassSpecGym encoder with corrected padding and pairwise formula masks, a single square-root intensity transform, the raw Top-60 peaks, and all peaks retained after annotation. Fingerprint prediction uses the sodium adduct. The navigation policy does not take adduct as an input. Adduct-dependent evidence enters through the MIST fingerprint probabilities.

D.2 Network Configuration

All reported fingerprint policies use the same architecture. The embedding width is 256, the feed-forward width is 512, and every attention block uses eight heads. The target fingerprint encoder has two self-attention layers. The molecular encoder has six layers before the two atom–fingerprint cross-attention layers and two layers afterward. Directed bond representations use two self-attention and two fingerprint cross-attention layers. Residual and attention dropout are both 0.1. Atom and bond cross-attention gates and the global target gate are initialized with logit -2. Collision energy is not used.

Action heads concatenate directed-bond representations, action-type embeddings, and global molecular features before an MLP. Type A averages four equivalent endpoint orderings. Type B first constructs a hydrogen-transfer representation from donor and recipient atom states and the global state, combines it with directed-bond representations, and averages two orderings. The averaged action representation is projected to a logit.

D.3 PubChem Pretraining

Pretraining uses four H100 GPUs with distributed data parallelism, global batch size 256, and AdamW with learning rate 4×10^{-4}, weight decay 10^{-2}, and $\beta = (0.9, 0.999)$. The learning rate follows cosine decay to zero, starting at the specified rate. Gradient norm is clipped at 1. We use full-precision training, one batch per optimizer step, and random seed 42. The run processes 125 million examples in 488,282 optimizer steps, saving checkpoints at 20 evenly spaced training milestones. Downstream training begins from the checkpoint after 37,499,904 examples, selected by validation Hit@1. We initialize the policy weights from this checkpoint and reset the optimizer, learning-rate schedule, and random states before fine-tuning.

D.4 Target-Conditioned Fine-Tuning

Fine-tuning minimizes Equation 4 and alternates batches from the D1–D3 set (at most 35 heavy atoms) and the D1–D2 set (more than 35 heavy atoms). NPLIB1 uses a 63:1 batch ratio, and MassSpecGym uses a 15:1 ratio. After removing distance-zero pairs, the combined sets contain 314,220 training observations for NPLIB1 and 13,412,519 for MassSpecGym. NPLIB1 trains for three epochs with 5,746 steps per epoch, and MassSpecGym trains for one epoch with 221,855 steps. The reported NPLIB1 checkpoint is from step 8,619, halfway through epoch two. The MassSpecGym checkpoint is from the end of its training epoch.

Both models use one H100, batch size 64, learning rate 5×10^{-5}, AdamW with weight decay 10^{-2} and $\beta = (0.9, 0.999)$, and cosine decay to zero from the initial learning rate. Gradient norm is clipped at 1. We use full precision, one batch per optimizer step, seed 42, and two data loader workers with unpinned batches.

E Search Algorithm

Valid-action enumeration checks bond-order bounds and heavy-atom connectivity using temporary adjacency arrays. Search materializes product states and constructs RDKit molecules for fingerprinting only for selected top-P actions. These proposals can later be removed by duplicate/visited checks or beam pruning. Trajectories use policy log probability to preserve sequences of learned edit choices, which need not monotonically improve the current molecule's fingerprint agreement.

Algorithm 1 Multi-start policy beam search

Require: Fingerprint probabilities q, formula $\mathcal{F}$, initial isomers $\{m_0^k\}_{k=1}^K$, policy p_θ, proposal width P, beam width B, depth D

$\mathcal{C} \leftarrow \{m_0^1, \ldots, m_0^K\}$
for $k = 1, \ldots, K$ **do**
  $\mathcal{H}_k \leftarrow \{(m_0^k, 0)\}$. $\mathcal{V}_k \leftarrow \{m_0^k\}$
end for
for $d = 1, \ldots, D$ **do**
  for $k = 1, \ldots, K$ **do**
    $\mathcal{P}_k \leftarrow \emptyset$
    for $(m, L) \in \mathcal{H}_k$ **do**
      enumerate valid edits $\mathcal{A}(m)$
      for top-P actions a under $p_\theta(\cdot \mid m, q, \mathcal{F})$ **do**
        $m' \leftarrow T(m, a)$. $L' \leftarrow L + \log p_\theta(a \mid m, q, \mathcal{F})$
        add (m', L') to $\mathcal{P}_k$
      end for
    end for
    remove members of $\mathcal{V}_k$ and deduplicate $\mathcal{P}_k$, retaining the largest L' per structure
    $\mathcal{H}_k \leftarrow$ top-B members of $\mathcal{P}_k$ by L'
    $\mathcal{V}_k \leftarrow \mathcal{V}_k \cup \{m' : (m', L') \in \mathcal{P}_k\}$
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{m' : (m', L') \in \mathcal{H}_k\}$
  end for
end for
compute multi-start support $c(m)$ for every $m \in \mathcal{C}$, including each seed's depth-zero visit
return unique InChIKey14 classes ranked using the score in Section 4.6

F Evaluation Protocol and Baseline Comparisons

F.1 Conventions

Unless stated otherwise, GEMS uses up to $K = 5$ MIST-ranked, target-excluded PubChem starts, $P = B = 16$, $D = 3$, $\lambda = 16$, and the reported checkpoint for each benchmark. Query-level results use all 803 NPLIB1 and 17,556 MassSpecGym test queries. MassSpecGym results pool 14,066 protonated and 3,490 sodiated spectra. Queries without starts count as failures. With PubChem initialization, these are 17 NPLIB1 and 200 MassSpecGym queries (129 protonated and 71 sodiated). Appendix H.3 gives the corresponding counts for generated pools.

Reach is the fraction of queries whose retained candidate pool contains a structure with the target's InChIKey14. It does not depend on ranking and is an upper bound on Hit@k. Per-seed reach, used in Appendix I, is the fraction of seed–target pairs whose own retained lineage contains the target. Queries without starts contribute no pairs, and queries with more starts receive more weight.

Results at a depth D below the searched maximum are prefixes of longer searches, with candidates and consensus votes restricted to states retained by depth D. A target first recovered at depth D shows a path of D edits, but the beam may have discarded a shorter one.

Labels such as MIST and Exact FP name the fingerprint used for navigation, ranking, or both. Exact FP is computed from the target structure. It is an oracle diagnostic that is unavailable at inference. Starting sets are fixed across corresponding MIST and exact-fingerprint conditions. Ranked starting sets are selected with MIST. Each training setting uses one random seed. The random-five comparison in Appendix H.1 reports repeated starting-set selections.

F.2 Structural Metrics and Published Baselines

Tables 3 and 4 give the complete published comparisons. NPLIB1 values come from FRIGID Table 1. MassSpecGym values come from *MassSpecGym in the Wild* Table 9, with FRIGID-base from FRIGID Table 1. We use the corrected MolForge results. *MS-GPT (published)* provides contextual comparisons (Zhao et al., 2026), with a different NPLIB1 split from ours. For numerical comparisons

and refinement experiments, we use *MS-GPT (ours)*, our best-effort reproduction retrained on the same benchmark splits after correcting the MIST attention masking issue described below. This reproduction has not been validated by the original authors. The reproduction uses the same full test sets and FRIGID-compatible structural metric implementation as GEMS. The NPLIB1 adapter checkpoint was selected on its 396 validation spectra, and the MassSpecGym checkpoint by validation cross-entropy.

MIST masking in MS-GPT. The released MS-GPT implementation retains the MIST padding-mask issue documented in *MassSpecGym in the Wild* (Liu et al., 2026, Section 6.1.1).[2] Padded tokens can contribute to attention, making fingerprint predictions depend on batch padding. GEMS excludes padded tokens from attention. The effect of the released masking code on MS-GPT's reported accuracy has not been measured.

GEMS structural metrics use the ranked top-ten unique InChIKey14 candidates and include every test query. We reuse FRIGID's implementation with RDKit 2026.3.5, `myopic_mces` 1.2.2, and PuLP 3.3.2. Molecules are converted through InChI, tautomer-canonicalized using FRIGID's two-transform canonicalizer, and stripped of stereochemistry. Tanimoto uses binary radius-two Morgan fingerprints with 2,048 bits. For each query, Top-k Tanimoto is the maximum similarity among its first k candidates, averaged over all queries.

MCES (maximum common edge subgraph) distance measures differences in molecular bonding under the best correspondence between atoms of the same element, accounting for bond order. Lower values indicate closer structural agreement, including when a candidate is not an exact match.

We report thresholded MCES with threshold 15. FRIGID's evaluator removes invalid or disconnected structures and duplicate InChIs from the first ten ranked candidates, preserving order without back-filling. Thresholded scoring can return lower bounds instead of exact distances. If canonicalization or MCES computation fails, the evaluator substitutes a bond-count score: the sum of target and candidate bond counts, or twice the target bond count. For each query, Top-k MCES is the minimum returned score among the first k remaining candidates, capped at 100 and averaged over all test queries. Queries without eligible candidates receive MCES 100 and Tanimoto zero. In our best-effort MS-GPT reproduction, 106 NPLIB1 and 3,295 MassSpecGym queries have no eligible candidates after structural preprocessing. Their MCES-100 penalties contribute 13.20 and 18.77, respectively, to the reported mean MCES.

The full results show the same pattern as Table 1: GEMS improves exact-match accuracy and Top-10 Tanimoto over the baselines evaluated on the same splits. FRIGID has lower MCES on NPLIB1. On MassSpecGym, GEMS reduces both MCES measures relative to FRIGID, while FOAM has lower Top-10 MCES (6.06 versus 8.26).

[2] Code inspected at MS-GPT commit `f2c844e`.

Table 3: NPLIB1 known-formula baselines and GEMS variants. Non-MS-GPT published values come from FRIGID Table 1 (Bohde et al., 2026). †: our best-effort same-split MS-GPT reproduction with corrected masking and retraining. ‡: published MS-GPT values are contextual comparisons with protocol qualifications in this section (Zhao et al., 2026). All GEMS variants use up to five starts, $P = B = 16$, $D = 3$, and $\lambda = 16$. PubChem starts exclude the target connectivity. The one-edit row additionally excludes every target-specific one-edit neighbor connectivity group, without a Tanimoto filter. Generator and Tanimoto-filtered starts follow Table 1. Our evaluations retain all test queries. Best displayed values are bold, including ties.

Method	H@1	MCES@1	Tani.@1	H@10	MCES@10	Tani.@10
Spec2Mol	0.00	48.57	0.12	0.00	17.58	0.16
MADGEN	2.10	20.56	0.22	2.39	12.69	0.27
MIST + Neuraldecipher	2.32	10.26	0.35	6.11	8.53	0.43
MIST + MolForge	2.24	14.16	0.40	5.11	10.96	0.47
MIST + MSNovelist	5.40	13.10	0.34	11.04	8.57	0.44
MS-BART	7.45	9.66	0.44	10.99	8.31	0.51
DiffMS	8.34	10.46	0.35	15.44	8.54	0.47
FRIGID-base	19.80	7.88	0.54	24.41	6.20	0.61
FRIGID	25.03	**7.10**	0.58	33.37	5.56	0.64
MS-GPT (published)‡	**29.76**	7.38	**0.61**	41.07	**5.51**	0.67
MS-GPT†	21.05	18.49	0.50	28.64	17.79	0.54
MS-GPT† → GEMS	26.28	20.61	0.52	42.22	19.54	0.60
FRIGID-base → GEMS	28.02	11.38	0.56	43.34	10.23	0.64
GEMS (PubChem, Tani. < 0.7)	27.90	8.59	0.56	46.45	6.97	0.67
GEMS (one-edit starts excluded)	25.53	8.69	0.55	42.22	7.06	0.66
GEMS	29.64	7.73	0.59	**50.81**	6.08	**0.71**

Table 4: MassSpecGym known-formula baselines and GEMS variants. Non-MS-GPT published values come from MassSpecGym in the Wild Table 9 (Liu et al., 2026), with FRIGID-base from FRIGID Table 1 (Bohde et al., 2026). †: our best-effort same-split MS-GPT reproduction with corrected masking and retraining. ‡: published MS-GPT values are contextual comparisons with protocol qualifications in this section (Zhao et al., 2026). All GEMS variants use up to five starts, $P = B = 16$, $D = 3$, and $\lambda = 16$. PubChem starts exclude the target connectivity. The one-edit row additionally excludes every target-specific one-edit neighbor connectivity group, without a Tanimoto filter. Generator and Tanimoto-filtered starts follow Table 1. Our evaluations retain all test queries. Best displayed values are bold, including ties.

Method	H@1	MCES@1	Tani.@1	H@10	MCES@10	Tani.@10
MIST + MSNovelist	0.00	39.84	0.06	0.00	18.83	0.15
Spec2Mol	0.00	36.78	0.12	0.00	36.02	0.16
MIST + Neuraldecipher	0.00	22.93	0.14	0.00	21.76	0.16
MADGEN	1.31	27.47	0.20	1.54	16.84	0.26
MS-BART	1.07	16.47	0.23	1.11	15.12	0.28
DiffMS	2.30	13.96	0.28	4.25	11.68	0.39
FOAM (DiffMS + PubChem)	1.50	12.21	0.35	10.28	**6.06**	0.53
MIST + MolForge (corrected)	10.73	22.15	0.37	14.48	17.88	0.41
FRIGID-base	16.09	15.02	0.41	18.19	11.95	0.46
FRIGID	18.29	13.49	0.43	22.00	11.65	0.47
MS-GPT (published)‡	23.91	15.50	0.52	28.65	11.74	0.54
MS-GPT†	23.23	25.15	0.48	27.70	24.71	0.50
MS-GPT† → GEMS	25.56	26.66	0.48	34.07	25.91	0.52
GEMS (PubChem, Tani. < 0.7)	23.93	10.95	0.50	31.75	9.61	0.55
GEMS (one-edit starts excluded)	22.44	11.02	0.49	29.57	9.65	0.55
GEMS	**30.22**	**9.68**	**0.55**	**41.33**	8.26	**0.61**

G Learning the Navigation Policy

G.1 Pretraining and Supervision

Table 5 compares random-weight and trained policies under the same MIST-guided search and $\lambda = 16$ ranking. The random-weight policy uses the reported architecture without training and gives $4.61 \pm 2.90\%$ Hit@1 and $8.30 \pm 5.29\%$ reach on NPLIB1 across three weight initializations (mean $\pm$ sample standard deviation), well below every trained NPLIB1 policy. Training improves performance under this fixed proposal and beam budget.

PubChem pretraining alone gives 28.89% Hit@1 on NPLIB1 without downstream fine-tuning, exceeding supervised training from scratch at 28.02%, although its Hit@10 and reach are lower. Fine-tuning adds 0.75 percentage points in Hit@1, 3.99 in Hit@10, and 7.72 in reach. It therefore improves identification as well as recovery. This contrast combines downstream training with adaptation from exact target fingerprints to MIST predictions.

Pretraining followed by fine-tuning outperforms supervised training from scratch in Hit@1, Hit@10, and reach on both benchmarks. The supervised models use the same benchmark-specific navigation examples and validation-selected checkpoints. From-scratch training uses six epochs on NPLIB1 and two on MassSpecGym, with an initial learning rate of 10^{-4} and cosine decay to zero without warmup on both benchmarks.

Table 5: Evaluation of random-weight and trained policies with $K = 5$, $P = B = 16$, $D = 3$, MIST fingerprints for navigation and ranking, and $\lambda = 16$. All 803 NPLIB1 and 17,556 MassSpecGym queries count. Pretrained only uses the PubChem checkpoint without downstream fine-tuning. Pretraining uses exact target fingerprints as inputs. REINFORCE uses MIST inputs and exact target fingerprints for its training reward. The random-weight row reports mean $\pm$ sample standard deviation across three weight initializations. Each trained setting uses one training random seed. Training schedules and selection details are given in the text.

Dataset	Policy	H@1	H@10	Reach
NPLIB1	Random weights	4.61 ± 2.90	7.97 ± 4.98	8.30 ± 5.29
NPLIB1	RL from scratch (exact-FP reward)	24.78	42.59	47.20
NPLIB1	Supervised from scratch	28.02	47.45	52.80
NPLIB1	Pretrained only	28.89	46.82	50.06
NPLIB1	Pretrained + fine-tuned	29.64	50.81	57.78
MassSpecGym	Supervised from scratch	29.14	39.46	44.20
MassSpecGym	Pretrained + fine-tuned	30.22	41.33	46.88

Shortest-path supervision versus REINFORCE. We compare training from scratch with shortest-path supervision against REINFORCE using exact-target-fingerprint rewards on NPLIB1. The matched recipe uses the same source–spectrum training examples, source-batch schedule, global batch size of 64, and optimizer schedule. REINFORCE samples 32 three-edit trajectories per source. Rewards measure improvement in the best Bernoulli agreement with the exact target fingerprint along a trajectory, divided by 100, with a leave-one-out baseline across trajectories. Policy inputs remain MIST predictions. The training schedule specifies 34,476 updates over six epochs, with four validation evaluations per epoch. Both selected checkpoints are at step 27,293, with supervised checkpoint selection based on $\lambda = 16$ validation Hit@1.

Shortest-path supervision exceeds REINFORCE in Hit@1, Hit@10, and reach (Table 5). Matching the update and source-exposure schedules does not match GPU-hours or establish variability across training runs. The comparison does not isolate the effect of retaining multiple optimal actions rather than one.

G.2 Learned Search versus Enumerate & Score

Table 6 reports the accuracy and cost of each depth-three configuration. At $P = B = 4$, GEMS gives 27.65% Hit@1 on NPLIB1 and 29.18% on MassSpecGym, with processing times of 81.11 and 1,456.71 seconds. This setting costs less than width sixteen, but recovers fewer targets. At the same beam width four, E&S reaches 54.17% of NPLIB1 targets and 44.24% of MassSpecGym targets, compared with 49.07% and 41.35% for GEMS. GEMS has higher Hit@1 on both benchmarks at this width and reduces processing time by factors of 31.1 and 60.0, respectively. E&S has higher NPLIB1 Hit@10.

Increasing the learned beam to sixteen raises reach to 57.78% and 46.88%. At this width, GEMS exceeds every tested E&S width in Hit@1, Hit@10, and reach. Compared with the widest tested E&S beam ($B = 4$), GEMS reduces reported runtime by factors of 12.6 on NPLIB1 and 21.4 on MassSpecGym.

Table 6: GEMS versus Enumerate & Score (E&S), with $K = 5$, $D = 3$ and $\lambda = 16$. Each search task uses 32 physical CPU cores and 32 workers; GEMS also uses one H100. NPLIB1 GEMS at $P = B = 16$ reports end-to-end wall time, including loading and output writing. Other runtime entries report accumulated processing time for preprocessing, retrieval, search, and final ranking (Appendix K.2). MassSpecGym search covers 16 tasks.

	NPLIB1					MassSpecGym				
Search (Params.)	H@1	H@10	Reach	Search (s)	Runtime (s)	H@1	H@10	Reach	Search (s)	Runtime (s)
E&S, $B = 1$	25.40	43.84	44.21	770.01	802.88	28.03	37.98	38.57	25,250.45	25,351.17
E&S, $B = 2$	25.65	47.57	50.19	1,364.69	1,397.98	28.18	39.25	41.55	47,022.92	47,131.94
E&S, $B = 4$	26.03	49.07	54.17	2,491.07	2,524.59	28.16	39.05	44.24	87,247.68	87,372.32
GEMS, $P = B = 4$	27.65	47.07	49.07	47.14	81.11	29.18	39.39	41.35	1,333.48	1,456.71
GEMS, $P = B = 16$	29.64	50.81	57.78	129.28	200.12	30.22	41.33	46.88	3,869.90	4,091.00

H Where Search Starts

H.1 PubChem Starting Sets

Table 7 compares query-level identification and recovery across four NPLIB1 starting sets at depths three, five, and ten. The full-catalog and sampled-250 settings select up to five starts by MIST agreement, using either the complete same-formula PubChem pool or a uniform 250-candidate draw. Random starts are the first up to 100 candidates of that draw before fingerprint sorting. The fixed random-five condition uses their first five starts. The same trained checkpoint and $P = B = 16$ are used throughout. MIST and exact-fingerprint conditions share the same starts, and each uses its stated fingerprints for both navigation and ranking. Candidates and consensus votes are restricted to the corresponding depth and starting set.

Table 7: NPLIB1 query-level Hit@1, Hit@10, and reach (%) across initialization and depth, with $P = B = 16$, $\lambda = 16$, and the same trained checkpoint. All 803 queries count, including 17 without starts. MIST and Exact FP specify both navigation and ranking. Starts are fixed across guidance conditions. Random, up to 5 uses the first five starts of the random-100 searches. The ten-repeat random-five comparison is reported separately in the text.

Initialization	D	MIST			Exact FP (oracle)		
		H@1	H@10	Reach	H@1	H@10	Reach
Full-catalog MIST, up to 5	3	29.64	50.81	57.78	68.99	72.73	72.85
	5	29.76	50.56	58.78	73.23	77.58	77.96
	10	29.64	50.81	59.90	79.45	84.43	84.93
Sampled-250 MIST, up to 5	3	26.15	43.46	50.44	61.15	64.38	64.51
	5	27.90	46.20	53.80	68.99	73.23	73.60
	10	27.90	48.19	56.54	77.58	82.57	83.06
Random, up to 5 (fixed)	3	11.83	18.93	22.17	29.76	32.50	32.63
	5	17.43	29.27	32.88	48.69	52.43	52.80
	10	22.29	35.24	40.85	68.62	72.23	72.60
Random, up to 100	3	25.53	43.96	57.66	66.63	69.74	69.74
	5	26.40	46.58	62.39	76.96	80.32	80.57
	10	28.89	49.07	66.00	85.43	89.54	89.91

Start quality. Under MIST guidance, full-catalog MIST selection gives the strongest D3 results among the tested initializations. Ranking within a 250-candidate sample lowers Hit@1 from 29.64% to 26.15% and reach from 57.78% to 50.44%. Up to 100 random starts nearly match full-catalog reach at 57.66%, but identify fewer targets, with 25.53% Hit@1.

Depth and identification. Deeper search partly compensates for weaker starts. At D10, the fixed random-five setting achieves 22.29% Hit@1 and 40.85% reach, still below the full-catalog ranked setting at D3. At D10, the random-100 setting reaches more targets than the ranked setting (66.00% versus 59.90%), but identifies fewer (28.89% versus 29.64% Hit@1).

Exact guidance. With exact fingerprints, the fixed random-five setting at D10 nearly matches full-catalog ranked reach at D3 (72.60% versus 72.85%). Exact-guided search from up to 100 random starts recovers the target for 89.91% of queries by D10. These starting sets therefore contain useful edit paths that MIST-guided search does not fully exploit.

Repeated random-five selection. Across ten runs drawing up to five starts from the same fixed random-100 pool, D3 Hit@1 is $11.41 \pm 0.73\%$, Hit@10 is $20.40 \pm 0.74\%$, and reach is $23.79 \pm 1.00\%$ (mean $\pm$ sample standard deviation). These runs use different subsets from the fixed five-start prefix in Table 7.

H.2 Starts Filtered by Similarity to the Target

Target-excluded PubChem starts can still include close analogs of the target. Motivated by the analysis of close analogs in pretraining data by Liu et al. (2026), we test sensitivity to similar structures among inference-time starts. We exclude candidates whose Tanimoto similarity to the target is at or above each cutoff, using radius-two, 2,048-bit Morgan fingerprints after the molecular standardization described in Appendix F.2. In a separate condition, we exclude all target-specific connectivity groups identified as one-edit neighbors by exhaustive enumeration. Both conditions retain target-connectivity exclusion. We then select up to five remaining candidates from the full same-formula catalog by MIST agreement. The checkpoint, search settings, and final ranking remain unchanged.

Results. Accuracy falls as stricter similarity cutoffs are applied (Table 8). At a cutoff of 0.7, the threshold used for close analogs in *MassSpecGym in the Wild* (Liu et al., 2026), GEMS gives 27.90%/46.45% Hit@1/Hit@10 on NPLIB1 and 23.93%/31.75% on MassSpecGym, above FRIGID's 25.03%/33.37% and 18.29%/22.00% (Table 1). MassSpecGym is more sensitive to the filter, losing about 6.3 Hit@1 points at 0.7 compared with 1.7 on NPLIB1. On NPLIB1, Hit@1 falls below FRIGID's only at a cutoff of 0.5 (21.92%), while Hit@10 remains higher (36.36% versus 33.37%). At this cutoff, additional edit rounds raise NPLIB1 Hit@1 from 9.34% at $D = 1$ to 19.05% at $D = 2$ and 21.92% at $D = 3$.

With up to five MIST-ranked starts and $P = B = 16$, the D1/D2/D3 Hit@1 values are 27.02%/29.39%/29.64% on NPLIB1 and 30.09%/30.43%/30.22% on MassSpecGym without Tanimoto filtering. At cutoff 0.7, they are 22.79%/27.27%/27.90% and 18.75%/22.76%/23.93%, respectively. The D1–D2 gains are therefore 4.48 and 4.02 percentage points after filtering, compared with 2.37 and 0.34 without it; the D2–D3 gains are 0.62 and 1.17, compared with 0.25 and -0.21. Gains are computed from hit counts before rounding.

With every one-edit neighbor removed, no target is recovered at $D = 1$. Every target recovered by $D = 3$ is therefore reached through two or three edits. GEMS gives 25.53%/42.22% Hit@1/Hit@10 on NPLIB1 and 22.44%/29.57% on MassSpecGym, still above FRIGID on both benchmarks. Increasing depth from two to three raises Hit@1 from 24.16% to 25.53% on NPLIB1 and from 20.31% to 22.44% on MassSpecGym. With target exclusion alone, the corresponding Hit@1 values change

Table 8: Hit@1, Hit@10, and reach (%) at $D = 3$, with $P = B = 16$, $\lambda = 16$, and up to five MIST-ranked starts after exclusion. Tanimoto rows keep only starts below the cutoff. One-edit exclusion removes every starting connectivity one valid edit from the target. Target-connectivity exclusion applies throughout. MassSpecGym Tanimoto filtering was evaluated at cutoffs of 0.7 and above. All test queries count, including those without starts.

	NPLIB1			MassSpecGym		
Starts removed	H@1	H@10	Reach	H@1	H@10	Reach
Target only (reported)	29.64	50.81	57.78	30.22	41.33	46.88
Tanimoto ≥ 0.9	29.39	50.44	57.04	29.94	40.62	46.07
Tanimoto ≥ 0.8	29.27	49.07	55.79	28.49	38.29	42.61
Tanimoto ≥ 0.7	27.90	46.45	52.18	23.93	31.75	34.64
Tanimoto ≥ 0.6	25.16	42.84	47.95	–	–	–
Tanimoto ≥ 0.5	21.92	36.36	40.10	–	–	–
One edit from target	25.53	42.22	46.58	22.44	29.57	32.03

from 29.39% to 29.64% on NPLIB1 and from 30.43% to 30.22% on MassSpecGym. These depth comparisons use up to five starts and $P = B = 16$.

Additional starts give larger identification gains under stricter similarity filtering. Increasing the maximum number of starts from five to twenty raises NPLIB1 Hit@1 by 0.75 points without filtering and by 2.12 points at a cutoff of 0.5, and Hit@10 by 1.99 and 5.35 points, respectively. Reach rises by about eight to nine points at every cutoff.

H.3 Generated Starting Pools and Refinement

For both FRIGID and MS-GPT, we select up to five executable same-formula candidates by MIST Bernoulli score, retaining the target when present. The original-ranking rows in Table 10 preserve each generator's output order. Selected (D0) evaluates the chosen starts, and GEMS (D3) adds states retained during three edit rounds from those same starts. Both use MIST fingerprints and $\lambda = 16$ ranking, with $P = B = 16$ for search. All 803 NPLIB1 and 17,556 MassSpecGym queries remain included, counting queries without executable starts as misses.

FRIGID candidate pools. The NPLIB1 FRIGID-base pool is the initial round of the same run that produces the final FRIGID pool. Its 65,151 candidates precede all ICEBERG refinement. FRIGID selects up to twelve parents by its own ranking in each of 24 refinement rounds, producing eight masked variants per parent and adding 32 fresh samples. Missing refinement slots are filled with fresh samples. GEMS and FRIGID therefore share the initial generated pool, while parent selection, ranking, and computational budgets differ. For further refinement, GEMS selects from the top 128 saved candidates per query after the 24 FRIGID refinement rounds. The original FRIGID row reports identification accuracy only because the full cumulative pool was not saved. There are 45 NPLIB1 queries without executable FRIGID-base starts and 23 without executable refined FRIGID starts. MassSpecGym includes 474 queries for which FRIGID-base outputs are unavailable.

MS-GPT candidate pools. We use outputs from our best-effort corrected and retrained MS-GPT reproduction (Appendix F.2), applying the same MIST-based selection and GEMS refinement procedure. There are 128 NPLIB1 and 3,633 MassSpecGym queries without executable selected starts.

Proximity of generated starts. Table 9 compares the three NPLIB1 starting pools. D0 checks whether a selected start already matches the target, and D1 exhaustively enumerates every valid one-edit product. D2/D3 combine this coverage with saved MIST- and oracle-guided paths, giving lower bounds on the fraction of queries with a start within that many edits. Among queries without a correct selected start, three-edit coverage is at least 292/611 (47.79%) for FRIGID-base, 275/561 (49.02%) for refined FRIGID, and 240/582 (41.24%) for MS-GPT. Queries without starts remain included. These conditional sets differ between generators. The counts establish proximity of the selected starts, independently of final ranking.

Refinement results. Table 10 compares the original generator rankings, selected starts, and GEMS refinement. Three edit rounds improve Hit@1, Hit@10, and reach over the selected-start baseline for every evaluated generator pool. On NPLIB1, GEMS exceeds our FRIGID run's original Hit@1 of 24.03%, reaching 28.02% from its initial pool and 28.27% from its refined pool. GEMS also improves on the original ranking of our best-effort MS-GPT reproduction, raising Hit@1 from 21.05% to 26.28% on NPLIB1 and from 23.23% to 25.56% on MassSpecGym. These results use MIST predictions for both navigation and ranking.

Table 9: NPLIB1 queries with at least one selected start within D edits of the target. Each generator pool supplies up to five MIST-selected starts. All 803 queries count, including those without starts. D0/D1 are exact. D2/D3 are lower bounds combining exhaustive one-edit coverage with saved MIST- and oracle-guided paths. MS-GPT (ours) is our best-effort reproduction. These counts measure proximity, independently of final ranking.

	FRIGID-base		FRIGID after refinement		MS-GPT (ours)	
Maximum edits	Queries	%	Queries	%	Queries	%
D0	192	23.91	242	30.14	221	27.52
D1	319	39.73	377	46.95	344	42.84
D2	$\geq$420	$\geq$52.30	$\geq$453	$\geq$56.41	$\geq$410	$\geq$51.06
D3	$\geq$484	$\geq$60.27	$\geq$517	$\geq$64.38	$\geq$461	$\geq$57.41

Table 10: Generator outputs and GEMS refinement. FRIGID-base and FRIGID are the initial and final pools of the same NPLIB1 run. MS-GPT (ours) is our best-effort reproduction with corrected masking and retraining. Original ranking uses the generator output order. Selected (D0) and GEMS (D3) use the same up-to-five MIST-ranked starts and $\lambda = 16$. All 803 NPLIB1 and 17,556 MassSpecGym queries count. Oracle results are separate in Table 11.

Dataset	Initializer	Stage	H@1	H@10	Reach
NPLIB1	FRIGID-base	Selected (D0)	20.42	23.91	23.91
NPLIB1	FRIGID-base	GEMS (D3)	28.02	43.34	46.20
NPLIB1	FRIGID	Original ranking	24.03	33.13	–
NPLIB1	FRIGID	Selected (D0)	22.54	30.14	30.14
NPLIB1	FRIGID	GEMS (D3)	28.27	46.82	50.31
NPLIB1	MS-GPT (ours)	Original ranking	21.05	28.64	–
NPLIB1	MS-GPT (ours)	Selected (D0)	22.67	27.52	27.52
NPLIB1	MS-GPT (ours)	GEMS (D3)	26.28	42.22	45.08
MassSpecGym	FRIGID-base	Selected (D0)	11.60	12.16	12.16
MassSpecGym	FRIGID-base	GEMS (D3)	18.12	23.43	24.72
MassSpecGym	MS-GPT (ours)	Original ranking	23.23	27.70	–
MassSpecGym	MS-GPT (ours)	Selected (D0)	24.00	26.68	26.68
MassSpecGym	MS-GPT (ours)	GEMS (D3)	25.56	34.07	36.64

H.4 Oracle Guidance from Generated Starts

Table 11 replaces navigation and ranking evidence with exact target fingerprints for the FRIGID and MS-GPT starts. Seed selection still uses MIST, so the starting pools are identical to the corresponding refinement runs. These are oracle diagnostics. Oracle D3 results are prefixes of completed ten-round searches, as are the corresponding MIST results for MS-GPT. The FRIGID MIST comparisons use standalone three-round searches.

With the starting pools fixed, exact fingerprints improve recovery and identification from both FRIGID pools on NPLIB1 and from MS-GPT on both benchmarks. On NPLIB1, Hit@1 rises from 28.02% to 57.91% for FRIGID-base and from 28.27% to 61.52% for refined FRIGID. Navigation and ranking change together in these comparisons. FRIGID-base oracle reach is 60.15%, while the 60.27% proximity bound in Table 9 also includes exhaustive one-edit coverage and paths witnessed by other searches.

Generated starting pools often contain candidates connected to the target by short edit paths. However, proximity alone does not ensure successful refinement: MIST-guided search misses some reachable targets, and final ranking does not consistently place recovered targets first. Exact-fingerprint interventions improve both recovery and identification from the same starting pools.

Table 11: Oracle diagnostics for generated starts. Exact target fingerprints guide navigation and ranking, using the same MIST-selected starts as Table 10. Search uses up to $K = 5$ starts, $P = B = 16$, $D = 3$, and $\lambda = 16$. MS-GPT (ours) is our best-effort reproduction. All 803 NPLIB1 and 17,556 MassSpecGym queries count, including missing starts.

Dataset	Initializer	Stage	H@1	H@10	Reach
NPLIB1	FRIGID-base	GEMS (D3)	57.91	59.90	60.15
NPLIB1	FRIGID	GEMS (D3)	61.52	64.38	64.38
NPLIB1	MS-GPT (ours)	Selected (D0)	27.27	27.52	27.52
NPLIB1	MS-GPT (ours)	GEMS (D3)	55.17	57.41	57.41
MassSpecGym	MS-GPT (ours)	Selected (D0)	26.13	26.68	26.68
MassSpecGym	MS-GPT (ours)	GEMS (D3)	41.05	42.87	42.92

I Search Width and Depth

I.1 Width versus Depth

The NPLIB1 grid varies $P = B$ and depth using the reported model and fixed ordered starting sets: up to five MIST-ranked starts or up to 100 random starts per query. Depths are prefixes of ten-round searches. Table 12 reports query-level identification, while Figure 4 and Table 13 report per-seed reach, the fraction of individual seed–target searches that recover the target. Per-seed reach separates individual search success from the number of opportunities per query, although queries with more starts receive more weight and the starting distributions differ.

Width versus depth under MIST. From the reference setting $P = B = D = 3$, we compare widening to $P = B = 10$ at $D = 3$ with deepening to $D = 10$ at $P = B = 3$. Both expanded settings retain at most $BD = 30$ search states plus the initial structure per seed before final deduplication. This matches the maximum retained-candidate budget, while actual unique candidate counts and computation can differ. From ranked starts, widening raises Hit@1 from 28.02% to 29.89%, compared with 29.02% for deepening. Per-seed reach similarly rises from 32.43% to 38.79% with widening and 37.22% with deepening. From random starts, the ordering reverses: widening raises Hit@1 from 24.91% to 25.90%, while deepening raises it to 29.39%. Per-seed reach rises from 5.58% to 7.10% and 20.38%, respectively. Hit@10 follows the same ordering for both starting sets. Ranked starts therefore benefit more from exploring alternatives within a short search, whereas random starts benefit more from extending trajectories in this comparison.

Exact fingerprint evidence changes the tradeoff. With exact fingerprints, deepening is more beneficial for both starting sets. From ranked starts, reference Hit@1 is 68.99%, compared with 68.87% for widening and 78.46% for deepening. From random starts, the corresponding values are 66.25%, 66.50%, and 84.18%. Per-seed reach and Hit@10 support the same ordering. Deeper search can thus yield substantially greater gains with more informative fingerprint evidence. Exact fingerprints replace both navigation and ranking inputs here, so this comparison does not isolate their individual contributions.

Diminishing returns in identification. From ranked starts under MIST, widths ten and sixteen differ by at most 0.50 percentage points in Hit@1 at the reported depths $D \in \{1, 3, 5, 10\}$. Recovery still improves: at $D = 3$, per-seed reach rises from 38.79% to 41.31% between these widths. At width sixteen, extending $D = 3$ to $D = 10$ further raises per-seed reach to 45.99%, while Hit@1 and Hit@10 remain at 29.64% and 50.81% at the two endpoints. Additional search therefore continues to improve recovery without necessarily improving identification.

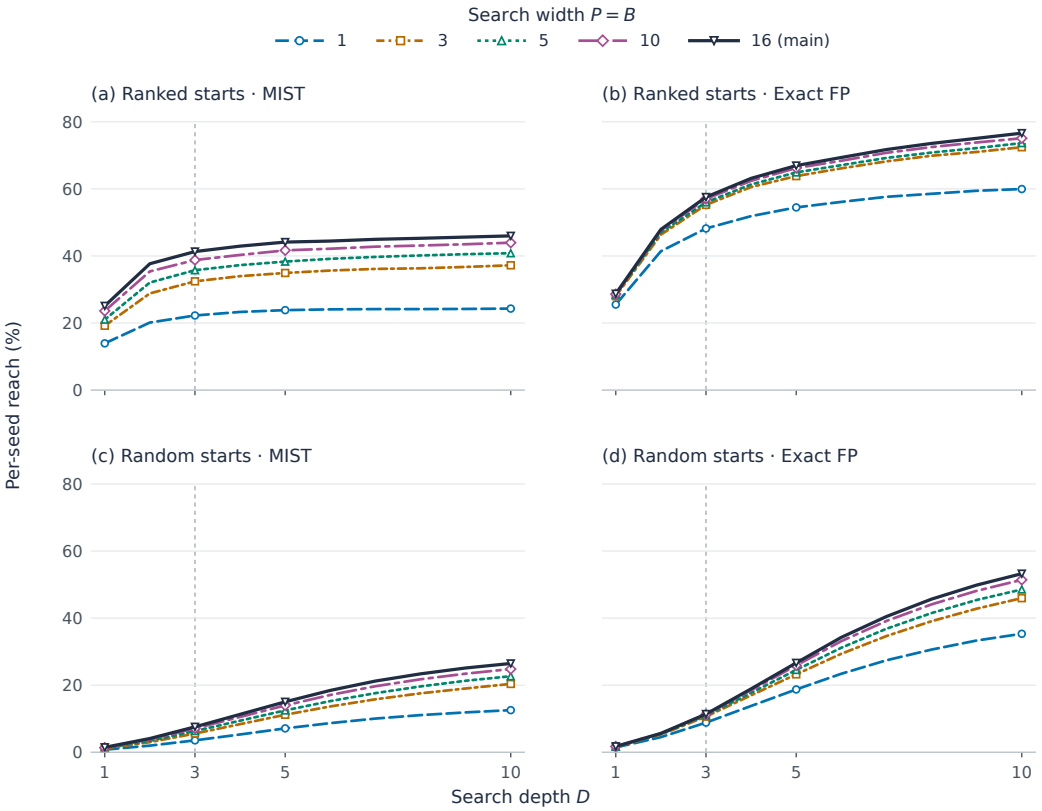


Figure 4: **Width and depth affect recovery differently.** Per-seed reach on NPLIB1 from up to five MIST-ranked starts or up to 100 random starts per query. Each curve uses $P = B$ and the same trained checkpoint. The vertical line marks the main depth $D = 3$. Exact FP uses fingerprints computed from the known targets and is an oracle diagnostic. Denominators are 3,888 ranked and 68,811 random seed–target pairs. Queries without starts contribute no pairs. Table 13 gives the numerical values.

Table 12: Query-level Hit@1 / Hit@10 (%) on NPLIB1 for the width–depth settings in Table 13. All 803 queries count, including 17 without starts. Each run uses $P = B$ and $\lambda = 16$ consensus with initial-seed support. Fingerprints specify both navigation and ranking. Exact FP uses the known target structure and is an oracle diagnostic. Candidates and consensus votes are restricted to depth D.

Initialization	Fingerprints	$P = B$	$D = 1$	$D = 3$	$D = 5$	$D = 10$
MIST-ranked, up to 5	MIST	1	23.54 / 32.38	25.16 / 36.74	25.78 / 37.24	26.03 / 37.48
		3	26.03 / 40.35	28.02 / 45.58	28.39 / 45.95	29.02 / 47.07
		5	26.15 / 43.09	28.52 / 47.95	28.89 / 48.32	28.89 / 48.19
		10	26.65 / 45.70	29.89 / 49.56	29.76 / 50.06	30.14 / 50.56
		16	27.02 / 46.82	29.64 / 50.81	29.76 / 50.56	29.64 / 50.81
MIST-ranked, up to 5	Exact FP	1	54.55 / 55.29	66.00 / 68.12	68.99 / 71.61	74.10 / 76.71
		3	56.91 / 57.66	68.99 / 71.48	72.35 / 75.47	78.46 / 81.82
		5	57.16 / 58.41	68.74 / 71.36	73.10 / 75.97	79.45 / 82.57
		10	57.53 / 58.78	68.87 / 72.23	73.10 / 77.21	78.83 / 83.56
		16	57.66 / 58.90	68.99 / 72.73	73.23 / 77.58	79.45 / 84.43
Random, up to 100	MIST	1	15.57 / 23.91	23.04 / 35.99	25.65 / 39.10	27.27 / 42.47
		3	16.56 / 29.51	24.91 / 41.34	26.90 / 43.84	29.39 / 46.45
		5	16.94 / 30.64	25.40 / 42.59	26.77 / 44.33	28.64 / 46.95
		10	17.93 / 32.38	25.90 / 42.84	26.65 / 46.08	29.14 / 48.19
		16	18.31 / 32.88	25.53 / 43.96	26.40 / 46.58	28.89 / 49.07
Random, up to 100	Exact FP	1	43.21 / 44.33	63.64 / 64.76	71.73 / 73.35	80.57 / 82.44
		3	45.58 / 46.70	66.25 / 68.37	75.22 / 78.08	84.18 / 87.30
		5	45.58 / 47.20	66.25 / 68.74	75.84 / 79.08	84.81 / 88.29
		10	45.83 / 47.32	66.50 / 68.99	76.46 / 80.07	85.06 / 89.17
		16	45.95 / 47.32	66.63 / 69.74	76.96 / 80.32	85.43 / 89.54

Table 13: Per-seed reach (%) on NPLIB1 for the searches in Figure 4. A seed–target pair succeeds when its own retained lineage contains the target by depth D. Denominators are 3,888 pairs for MIST-ranked starts and 68,811 pairs for random starts. Queries without starts contribute no pairs. All runs use $P = B$ and the same trained checkpoint. Exact FP is an oracle diagnostic.

Starts	Guidance	$P = B$	$D = 1$	$D = 3$	$D = 5$	$D = 10$
MIST-ranked, up to 5	MIST	1	13.94	22.25	23.84	24.31
		3	19.19	32.43	34.93	37.22
		5	21.01	35.75	38.32	40.82
		10	23.61	38.79	41.67	43.96
		16	25.10	41.31	44.14	45.99
MIST-ranked, up to 5	Exact FP	1	25.51	48.20	54.48	59.95
		3	28.01	55.25	63.81	72.43
		5	28.47	55.99	64.92	73.66
		10	28.60	57.02	66.23	75.10
		16	28.73	57.56	66.95	76.62
Random, up to 100	MIST	1	0.76	3.55	7.12	12.56
		3	1.08	5.58	11.18	20.38
		5	1.21	6.33	12.49	22.67
		10	1.36	7.10	13.97	24.81
		16	1.46	7.57	15.04	26.47
Random, up to 100	Exact FP	1	1.44	8.84	18.73	35.33
		3	1.62	10.64	23.23	45.96
		5	1.67	10.98	24.55	48.55
		10	1.72	11.26	25.89	51.42
		16	1.74	11.39	26.65	53.25

I.2 Recovery by Depth on Both Benchmarks

Table 14 reports query-level recovery by depth on both benchmarks, using the reported trained policies, up to five MIST-ranked PubChem starts, and $P = B = 16$. The retained paths show that many target-excluded retrieved isomers are close to an accepted target. With MIST guidance, at least one start reaches the target within a single edit for 51.56% of NPLIB1 queries and 43.30% of MassSpecGym queries. Recovery rises to 57.78% and 46.88% by depth three. These measurements depend on the candidate pool, independently of final ranking.

On both benchmarks, additional depth yields substantially larger recovery gains with exact fingerprint evidence. Extending search from D3 to D10 adds 2.12 percentage points of reach on NPLIB1 and 0.94 on MassSpecGym under MIST guidance. With exact fingerprints, the same extension adds 12.08 and 8.96 points.

Additional recovery does not consistently improve top-ranked identification. We compare depth-three and depth-five prefixes of the same searches, using the candidates and lineage votes retained through each depth. Under MIST guidance, extending search from D3 to D5 raises NPLIB1 reach from 57.78% to 58.78%, while Hit@1 changes from 29.64% to 29.76%. On MassSpecGym, reach rises from 46.88% to 47.30%, while Hit@1 decreases from 30.22% to 29.88%.

Table 14: Queries with a witnessed path of at most D edits from at least one MIST-ranked starting isomer. All queries remain in the denominator, including queries without starts. Percentages are recovery, independent of final ranking. These provide lower bounds on the fraction of queries having a seed within D edits. Failures do not establish a larger shortest distance. Exact FP is an oracle diagnostic.

Dataset	Guidance	Queries	D1	D2	D3	D5	D10
NPLIB1	MIST	803	51.56	56.66	57.78	58.78	59.90
NPLIB1	Exact FP	803	59.03	69.49	72.85	77.96	84.93
MassSpecGym	MIST	17556	43.30	45.94	46.88	47.30	47.82
MassSpecGym	Exact FP	17556	48.18	53.86	56.07	58.40	65.03

I.3 Witnessed Paths and Certified Distances

A recovered target establishes a valid edit path, but the first recovery depth need not be the shortest possible distance. Certifying that distance requires both a verified path and exhaustive exclusion of shorter paths. Figure 1 illustrates a certified three-edit example. Table 14 reports witnessed paths across the benchmarks, where this certification procedure has not been applied to every starting seed. Failure to recover a target with a finite beam leaves its distance unresolved.

We measure $d_{\mathrm{eval}}(s, t)$ from the exact initial seed state, accepting any endpoint whose InChIKey14 matches the reference target. For a target-excluded seed, exhaustive forward search through two edits tests every reachable state for a match. A first match at depth one or two certifies that distance. If neither depth contains a match, the distance is at least three.

To search for a three-step path, we also expand the reference target by one edit. The InChIKey14 values in that neighborhood select promising seed-reachable states at depth two. We then expand those states forward from their actual bond-and-hydrogen configurations. A path is accepted only when its endpoint matches the target InChIKey14 and all three edits replay successfully. Combined with the two-step lower bound, this certifies distance three.

The target-side filter can miss paths ending at other accepted target structures. When no verified path is found, the lower bound remains three and the exact distance is unresolved.

J Fingerprint Evidence in Navigation and Ranking

J.1 Fingerprint Interventions with PubChem Starts

Table 15 reports fingerprint interventions with and without consensus. Its $\lambda = 16$ rows give the numerical results for Figure 3.

Table 15: Fingerprint replacement during navigation and ranking at $K = 5$, $P = B = 16$, and $D = 3$, with pure Bernoulli ranking ($\lambda = 0$) or consensus ($\lambda = 16$). All rows use the same MIST-selected starts. Changing only the ranking fingerprint preserves the candidate pool and reach. Metrics are percentages. MassSpecGym includes both adducts. Rows with $\lambda = 16$ give the values in Figure 3.

			NPLIB1			**MassSpecGym**		
Navigation	Ranking	λ	H@1	H@10	Reach	H@1	H@10	Reach
MIST	MIST	0	28.02	50.56	57.78	29.47	40.94	46.88
MIST	MIST	16	29.64	50.81	57.78	30.22	41.33	46.88
MIST	Exact FP	0	55.54	57.66	57.78	44.76	46.74	46.88
MIST	Exact FP	16	56.66	57.53	57.78	45.12	46.79	46.88
Exact FP	MIST	0	29.39	58.78	72.85	31.19	47.10	56.07
Exact FP	MIST	16	33.25	61.89	72.85	33.82	49.74	56.07
Exact FP	Exact FP	0	67.62	72.60	72.85	51.04	55.87	56.07
Exact FP	Exact FP	16	68.99	72.73	72.85	53.09	55.94	56.07

Navigation. With $\lambda = 16$ and MIST ranking, replacing the navigation fingerprint with the exact target fingerprint raises reach from 57.78% to 72.85% on NPLIB1 and from 46.88% to 56.07% on MassSpecGym. Hit@1 increases from 29.64% to 33.25% and from 30.22% to 33.82%, respectively. Recovery therefore improves more than top-ranked identification.

Ranking. Many recovered targets are not placed first by MIST ranking. With the MIST-guided candidate pools fixed, replacing only the ranking fingerprint raises Hit@1 from 29.64% to 56.66% on NPLIB1 and from 30.22% to 45.12% on MassSpecGym. These values approach the corresponding reach of 57.78% and 46.88%. Candidate pools and consensus counts remain unchanged, so the improvement comes from selecting among already recovered candidates rather than additional search.

Replacing both navigation and ranking fingerprints raises Hit@1 further, to 68.99% on NPLIB1 and 53.09% on MassSpecGym. Together, these oracle interventions show scope to improve both target recovery and selection from the recovered pool, with the starting sets and trained checkpoints held fixed.

J.2 Effect of Consensus

We select $\lambda = 16$ by MassSpecGym validation Hit@1 across both adducts and use it for both benchmarks.

Consensus reorders a fixed candidate pool without changing reach. It raises Hit@1 from 28.02% to 29.64% on NPLIB1 and from 29.47% to 30.22% on MassSpecGym (Table 15). The improvement in NPLIB1 Hit@10 is smaller, from 50.56% to 50.81%.

Among the tested interventions, consensus gives its largest Hit@1 gains with exact navigation and MIST ranking: 3.86 percentage points on NPLIB1 and 2.63 on MassSpecGym.

Removing consensus leaves the main conclusion of the fingerprint experiment unchanged (Table 15). With the MIST-guided pools fixed, exact-fingerprint ranking raises Hit@1 from 28.02% to 55.54% on NPLIB1 and from 29.47% to 44.76% on MassSpecGym. Using exact fingerprints for both navigation and ranking raises these values to 67.62% and 51.04%. These comparisons use the same seeds, checkpoint, and candidate pools as the $\lambda = 16$ conditions. Only the consensus term is removed.

J.3 Reranking with Predicted Spectra

We test whether predicted spectra provide complementary ranking evidence by reranking the fixed top-ten lists from the main GEMS configuration. ICEBERG (Goldman et al., 2024) predicts a spectrum for each candidate using the benchmark-specific checkpoints released with FRIGID. We add a spectral score $S(m)$ to the GEMS score, giving $R(m; q) + \beta S(m)$, while keeping $\lambda = 16$. The metric, energy aggregation, and β are selected by validation Hit@1, with $\beta \in \{0, 4, 8, 16, 32, 64, 128\}$.

NPLIB1 selects $\beta = 128$ and experimental intensity coverage, the fraction of observed intensity matched by predicted peaks. MassSpecGym selects $\beta = 32$ and cosine similarity of square-root intensities. Both use one-to-one peak matching within 20 ppm and exclude precursor and above-

precursor peaks. For NPLIB1, scores use matched recorded collision energies when available and otherwise average over 10–90 eV in 10 eV increments. MassSpecGym uses the nearest grid energy to the recorded value, or 40 eV when energy is unavailable.

Table 16: ICEBERG reranking of the fixed GEMS top-ten lists. Hit@1 is reported in percent. Gains and their paired 95% bootstrap intervals, grouped by target connectivity, are in percentage points.

Ranking	NPLIB1	MassSpecGym
GEMS	29.64	30.22
GEMS + ICEBERG	31.88	31.60
Gain [95% interval]	+2.24 [0.62, 3.96]	+1.38 [0.40, 2.80]

The validation-selected spectral scores improve Hit@1 on both benchmarks (Table 16), showing that forward spectral predictions complement fingerprint-based ranking. Substantial identification error remains within the fixed candidate lists. All test queries remain in the denominators. Hit@10 and full-pool reach are unchanged because reranking preserves candidate membership and cannot rescue targets outside the top ten. Appendix K.2 reports the additional prediction cost.

K Runtime Measurements

At $K = 5$, $P = B = 16$, and $D = 3$, GEMS takes 200.12 seconds end to end on NPLIB1, the median of three runs over all 803 test spectra, or 0.249 seconds per spectrum. This wall time includes process startup, model and data loading, spectrum preprocessing, peak annotation, MIST prediction, fresh fingerprint database construction, retrieval, search, final ranking, and output writing. Post-hoc accuracy evaluation and correctness audits are excluded, as are queue waiting and prior PubChem download and catalog ingestion. The MassSpecGym measurement is accumulated processing time of 4,091.00 seconds, or 0.233 seconds per spectrum, and excludes startup and file I/O. Table 17 reports these runtimes alongside their component costs.

K.1 Search Timing Boundaries

Search timers start after the main-process model has been initialized and include worker initialization, warm-up, and search until results are returned in memory. NPLIB1 processes the full test set in one task. MassSpecGym is divided into sixteen contiguous groups, with at most two tasks running concurrently. Reported time is the sum of all sixteen task durations. Each task uses 32 physical CPU cores, 32 workers, and 192 GiB RAM. Learned search also uses one H100.

K.2 Pipeline Timing Boundaries

The NPLIB1 $P = B = 16$ GEMS timer starts at process launch and stops after ranked predictions have been written and the process exits. It measures one continuous execution from available raw spectra and formula catalogs. All three runs start fresh processes and reload the models; filesystem caches are not flushed. The end-to-end value is used consistently in Tables 6 and 17.

The component rows in Table 17 isolate in-memory processing costs. Preprocessing parses and normalizes spectrum peaks; annotation assigns molecular subformulas; MIST predicts fingerprint probabilities. Database construction freshly fingerprints validated candidate SMILES, and retrieval selects the starting structures. Final ranking includes consensus scoring, molecule validation, connectivity deduplication, and worker startup and shutdown. Input-component and ranking times are medians of three repetitions, as is the NPLIB1 $P = B = 16$ search time; other search times retain the recorded task measurements.

Continuous end-to-end timing is available only for NPLIB1 GEMS at $P = B = 16$. The remaining entries in Table 6 report accumulated processing costs for the shared inputs, each method's search, and final ranking, excluding startup, file I/O, and post-hoc evaluation. Thus the NPLIB1 comparison charges loading and output overhead to GEMS alone. MassSpecGym search costs cover sixteen tasks, with final ranking measured separately over the full cohort on 32 CPU workers.

Table 17: Runtime measurements for $K = 5$, $P = B = 16$, and $D = 3$, using one H100 and 32 physical CPU cores. NPLIB1 reports median end-to-end wall time, including startup, loading, and output writing. MassSpecGym reports accumulated processing time excluding this overhead; its search covers 16 tasks. Component rows show in-memory processing costs. Per-spectrum values divide the reported runtimes by 803 or 17,556 spectra.

Component	NPLIB1 (s)	MassSpecGym (s)
Spectrum preprocessing	0.57	0.40
Peak annotation	0.61	3.04
MIST prediction	0.40	4.91
Build fingerprint database from SMILES	28.01	77.41
Retrieve top-5 seeds	0.32	1.50
GEMS search	129.28	3,869.90
Final ranking	6.06	133.85
Reported runtime	200.12	4,091.00
Per spectrum	0.249	0.233

FRIGID-base generation. A separate FRIGID-base generation run over all 803 NPLIB1 spectra took 5,353 seconds on one H100. This standalone timing is distinct from the run supplying the initial pool for the paired FRIGID/GEMS refinement comparison. FRIGID's published accuracy in Table 1 includes subsequent ICEBERG refinement.

FRIGID with ICEBERG refinement. Our NPLIB1 run takes 94,408 seconds, approximately 26.2 hours, for all 803 test spectra. It uses one H100, 32 allocated CPU cores, and 128 GB host memory, with 128 attempted samples per spectrum per round and initial generation followed by 24 refinement rounds. The final saved ranking gives Hit@1/Hit@10 of 24.03%/33.13%. This benchmark wall time includes model and data loading, MIST prediction, generation, ICEBERG subprocesses, per-round ranking and evaluation, and output writing. GEMS's end-to-end runtime is 200.12 seconds, approximately 3.3 minutes, using one H100, 32 physical CPU cores, and 192 GiB host memory. It includes loading and writing through the final ranked predictions, with post-hoc evaluation excluded as described above.

ICEBERG reranking. Producing the test prediction caches for Appendix J.3 takes 16 minutes 58 seconds on one H100 for NPLIB1 and 9 minutes 59 seconds across eight H100s for MassSpecGym, totaling 17.0 and 79.2 GPU-minutes, respectively. These prediction-job times include model setup and output writing, but exclude queue waiting, input preparation, GEMS search, and coefficient selection. CPU spectral scoring adds approximately 15 and 81 seconds, respectively. The caches support multiple scoring and energy choices. The selected rules require only 60,622 of 69,425 cached conditions on NPLIB1 and 106,676 of 355,912 on MassSpecGym.

K.3 Illustrative Neighborhood Enumeration

For a 32-heavy-atom NPLIB1 molecule, one-edit enumeration produced 1,904 distinct connectivities in 2.10 seconds. Extending to two edits processed 4.57 million valid proposals and found 1.72 million additional connectivities in 41 minutes 39 seconds on one CPU worker.[3] A third exhaustive round would expand 2.20 million atom-aligned states, with an upper bound of 5.68 billion raw edit proposals before validity filtering. This rapid neighborhood growth motivates GEMS's selective edit expansion and bounded beam at each depth.

[3]The example is $C_{24}H_{30}F_2O_6$, InChIKey14 `FEBLZLNTKCEFIT`.

L ILLUSTRATIVE SEARCH TRAJECTORIES

Figures 5 and 6 show MIST-guided paths from selected PubChem seeds at certified distances one, two, and three on each benchmark. Distance refers to the displayed seed, not the closest seed in the query's starting pool. Each sequence follows a retained seed lineage to its first target hit. All displayed edits were replayed from the saved initial state, checking formula and atom-valence preservation and the final target InChIKey14. The shorter-path exclusion procedure is described in Appendix I.3. These examples were selected after evaluation for illustration.

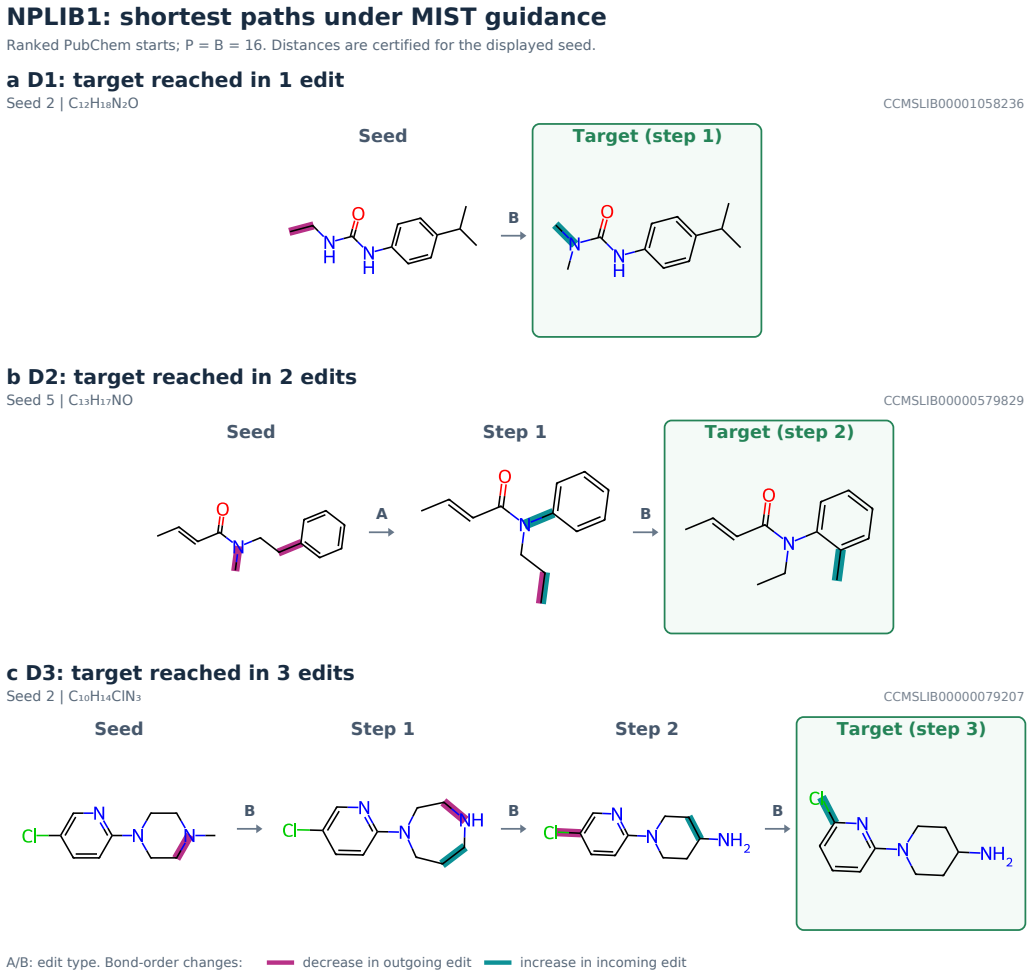


Figure 5: **NPLIB1 paths from seeds at certified distances D1–D3.** MIST-guided search reaches the target in one, two, and three edits, respectively, matching the minimum distance from each displayed seed. The examples show side-chain rearrangement, changes in substituent attachment, and ring expansion followed by contraction. All use the reported checkpoint, ranked PubChem starts, and $P = B = 16$. A/B denote edit types. Magenta highlights mark bond-order decreases in the outgoing edit, and teal marks increases in the incoming edit. These are graph-edit paths, not proposed chemical reactions.

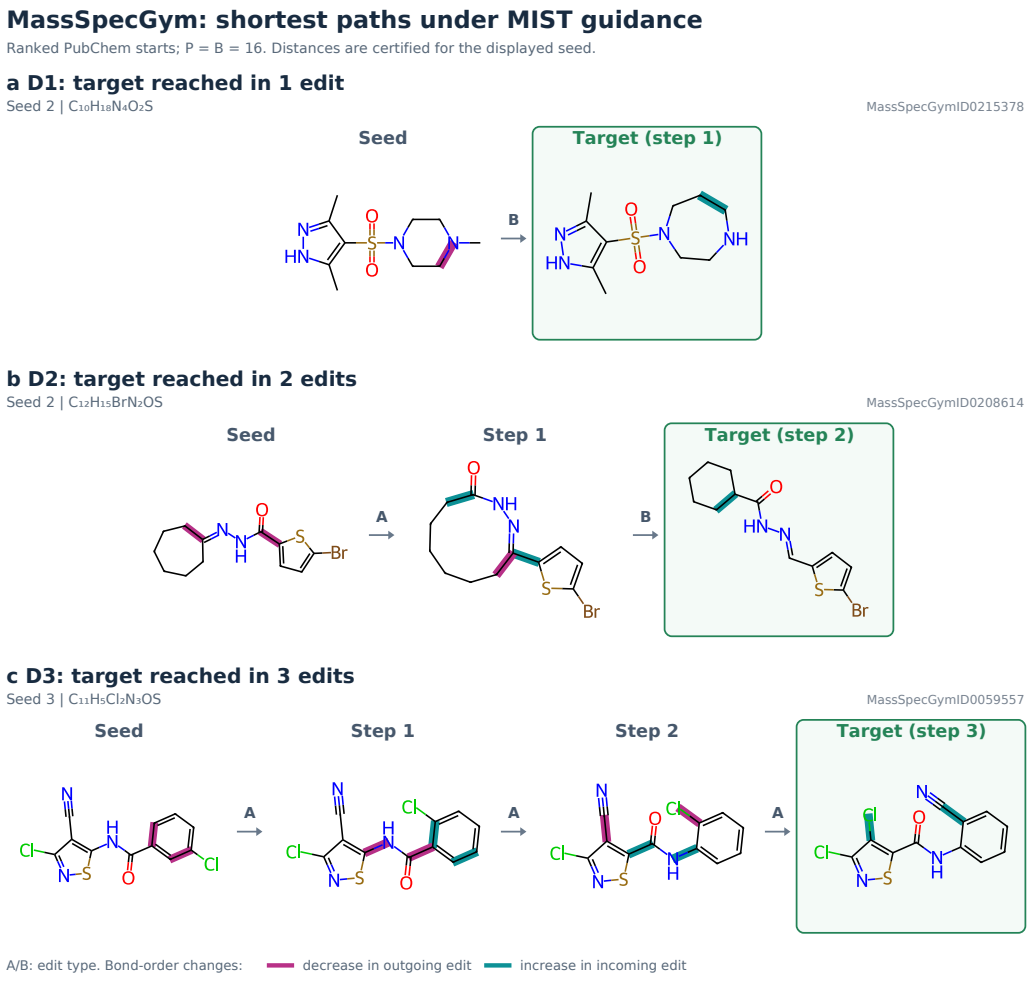


Figure 6: **MassSpecGym paths from seeds at certified distances D1–D3.** MIST-guided search follows a shortest path from each displayed seed. The examples include ring expansion, rearrangement through a larger-ring intermediate, and redistribution of substituents between aromatic rings. All use the reported checkpoint, ranked PubChem starts, and $P = B = 16$. Display conventions follow Figure 5.

Longer paths under predicted fingerprints. Figures 7 and 8 compare guidance from MIST predictions and exact target fingerprints. Within each pair, the initial seed state, trained checkpoint, search widths, and final target structure are identical. The first retained target path from that seed takes five versus three edits on NPLIB1 and six versus three on MassSpecGym. The certified distance is three in both cases, so the MIST-guided paths contain two and three additional edits, respectively. These comparisons concern navigation; they do not depend on the target's final rank or imply that another seed could not reach it sooner.

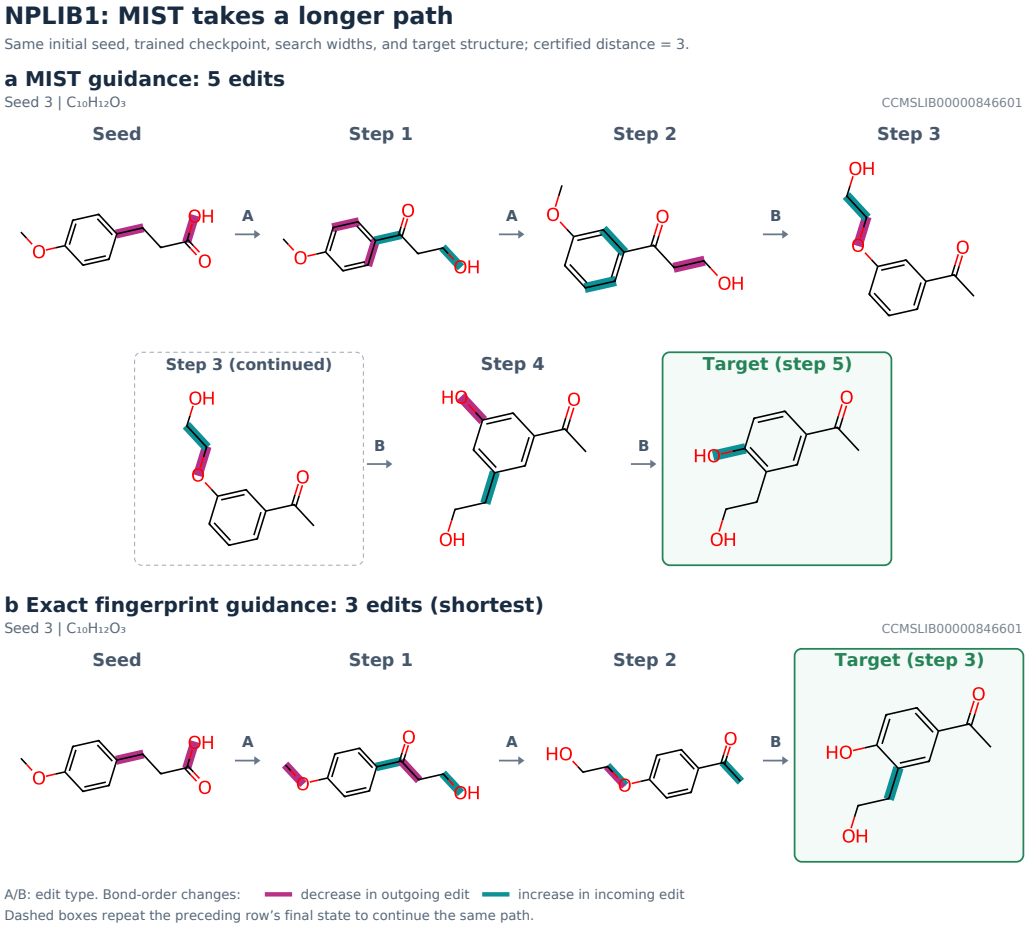


Figure 7: **A five-edit MIST path and a three-edit exact-guided path on NPLIB1.** Both paths start from Seed 3 of `CCMSLIB00000846601`. They share their first edit and then diverge. Exact guidance reaches the target in the certified minimum of three edits, while MIST first reaches it after five. Search uses $P = B = 16$ with a ten-round budget. The displayed paths stop at the first hit from this seed. Dashed boxes repeat the preceding row's final state to continue the same path.

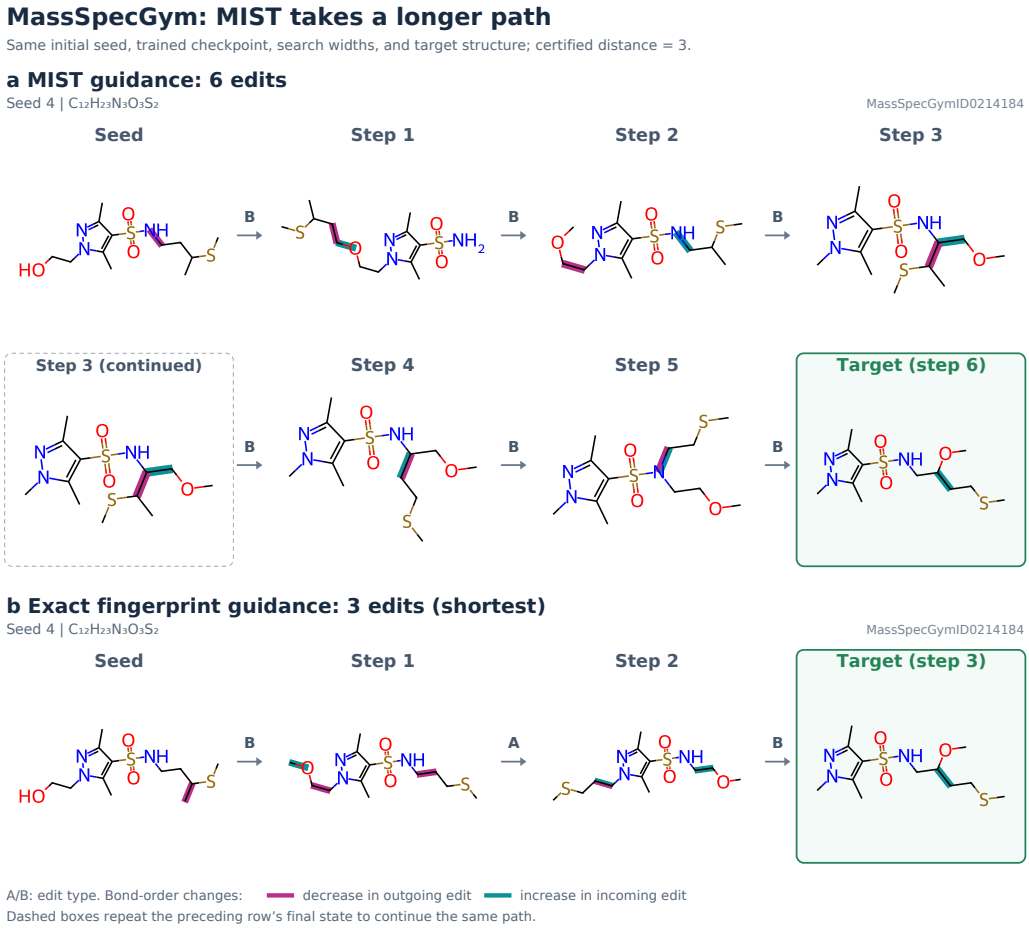


Figure 8: **A six-edit MIST path and a three-edit exact-guided path on MassSpecGym.** Both paths start from Seed 4 of `MassSpecGymID0214184` and end at the same target structure. Exact guidance follows a shortest path. MIST takes three additional edits in this search before its first retained target hit from this seed. Search uses $P = B = 16$ with a ten-round budget. Display conventions follow Figure 7.

A longer path from a random start. Figure 9 follows one of 100 random PubChem starts for an NPLIB1 query. The retained path opens and reconnects rings before reaching the target at its tenth edit. This is an observed path length; its shortest possible distance has not been certified. Other seeds for this query recover the target earlier.

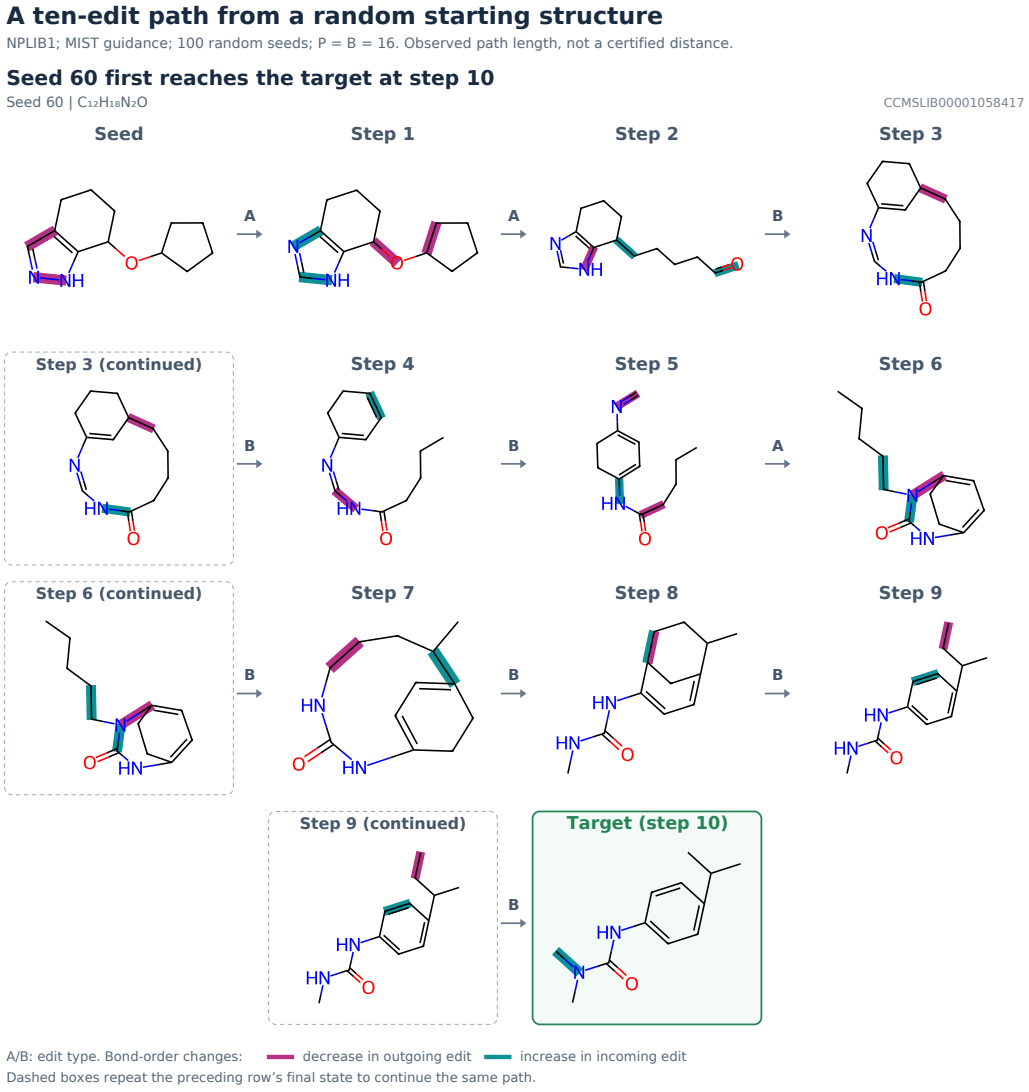


Figure 9: **A ten-edit MIST-guided path from a random seed.** Seed 60 of the 100-seed random pool first reaches the target at depth ten for NPLIB1 query `CCMSLIB00001058417`. Every displayed transition is a replayed retained edge from the $P = B = 16$, $D = 10$ search. The example illustrates substantial structural rearrangement along a successful lineage, rather than a certified ten-edit distance or a requirement for ten rounds across all seeds.