COST-AWARE HIERARCHICAL EVIDENCE ROUTING FOR MOLECULAR RETRIEVAL FROM MASS SPECTRA

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular retrieval from tandem mass spectra aims to rank candidate structures based on their compatibility with observed fragmentation patterns. Existing approaches based on global spectrum–molecule representations do not adequately distinguish candidates with subtle structural differences and therefore struggle to rank structurally similar molecules accurately. Fine-grained structural evidence can improve discrimination among these candidates, but uniformly evaluating such evidence for all candidates incurs substantial computational costs and overlooks query-specific refinement needs. To address this challenge, we propose **MS-CHER**, a two-stage molecular retrieval framework with **C**ost-aware **H**ierarchical **E**vidence **R**outing. The key idea is to tailor fine-grained evidence computation to the refinement needs of each query. We obtain an initial candidate ranking through global spectrum–molecule matching using structure-aware molecular representations. Building on this ranking, a hierarchical routing policy determines a query-dependent refinement scope and selectively activates structural evidence within this scope according to predicted ranking benefits and computational costs. Experiments on multiple MS/MS retrieval benchmarks demonstrate competitive retrieval performance across diverse candidate settings. Notably, compared with exhaustive evidence evaluation on MassSpecGym, MS-CHER reduces structural evidence evaluation runtime by approximately 79.2% and 82.4% in the mass and formula settings, respectively, while maintaining comparable retrieval accuracy.

1 INTRODUCTION

Tandem mass spectrometry (MS/MS) records the mass-to-charge ratios and intensities of fragment ions from precursor fragmentation, providing observations for small-molecule structure identification (Dührkop et al., 2019; Bushuiev et al., 2026). However, collision-induced fragmentation reveals partial information about the parent molecule, while experimental spectra often contain missing peaks, noise, and incomplete fragmentation patterns. Inferring molecular structure from incomplete observations remains challenging in computational metabolomics (Bushuiev et al., 2024).

Spectral library matching, widely used for molecular identification, compares query spectra against references from known compounds (Tsugawa et al., 2015; Wang et al., 2016). Despite its effectiveness with high-quality references, experimental availability and library size inherently limit coverage (Bushuiev et al., 2026). **Molecule retrieval** provides a more open alternative for ranking ground-truth molecules highly among candidates constrained by precursor mass, formula, or a chemical database, using MS/MS spectrum compatibility (Dührkop et al., 2015; Bushuiev et al., 2024).

Recent deep learning approaches have substantially improved molecular retrieval by learning representations capturing spectrum–structure relationships (Goldman et al., 2023; Kalia et al., 2025; Zhou Chen & Hassoun, 2026). While global similarity efficiently captures overall spectrum–molecule compatibility, it may not fully exploit the structural evidence required to distinguish closely competing candidates. When candidate molecules share similar scaffolds, functional groups, or closely related fragmentation patterns, the evidence determining the final ranking can reside in substantially finer-grained structural relationships (Zhou Chen et al., 2026). Moreover, queries may require different degrees of structural refinement: some are already well separated by global similarity, whereas others require additional structural evidence **(Figure 1a)**. Uniformly evaluating such evidence for

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.

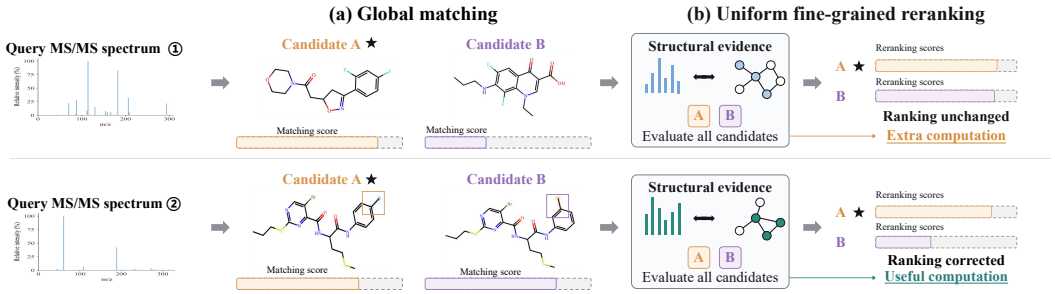

Figure 1: **Motivation for query-dependent evidence allocation.** **(a)** Global matching clearly separates candidates for one query but leaves structurally similar candidates ambiguous for another. **(b)** Uniform fine-grained reranking evaluates structural evidence for both queries, preserving the first ranking while correcting the second. **Differing ranking gains motivate query-dependent allocation of evidence computation.** Examples are schematic, and stars mark ground-truth candidates.

every query and candidate can therefore introduce unnecessary computation **(Figure 1b)**, motivating query-dependent allocation of evidence computation (Jeong et al., 2024; Genc et al., 2026).

Based on these observations, we propose **MS-CHER**, a two-stage molecular retrieval framework with cost-aware hierarchical evidence routing. We obtain an initial candidate ranking through global spectrum–molecule matching using structure-aware molecular representations. For fine-grained refinement, we develop two structural evidence modules: *LOCAL* captures spectrum–substructure correspondence through local structural matching, while *FRAG* evaluates spectral support for candidate-specific fragmentation paths. Before expert execution, a cost-aware hierarchical routing policy selects a query-dependent refinement scope and then determines which evidence modules to activate within this scope according to predicted ranking benefits and computational costs. By tailoring structural evidence computation to query-specific refinement needs, MS-CHER reduces unnecessary evidence evaluation while maintaining strong retrieval performance.

Our main contributions are summarized as follows:

- We formulate fine-grained refinement in molecular retrieval as a **query-dependent evidence allocation problem** and develop two structural evidence modules for *local spectrum–substructure matching* and *fragmentation-path matching*.
- We introduce a **cost-aware hierarchical routing mechanism** that sequentially determines the refinement scope and evidence activation before expert execution, enabling adaptive allocation of structural evidence computation.
- We evaluate MS-CHER on multiple MS/MS retrieval benchmarks and demonstrate competitive retrieval performance across diverse candidate settings. Moreover, on MassSpecGym, MS-CHER substantially reduces structural evidence evaluation runtime compared with exhaustive evidence evaluation while maintaining comparable retrieval accuracy.

2 Related Work

Molecular retrieval from tandem mass spectra. Existing molecular retrieval methods rank candidate structures through fingerprint prediction or learned spectrum–molecule representations. Fingerprint-based methods, including CSI:FingerID and MIST, infer structural properties or molecular fingerprints from spectra and use them for candidate ranking (Dührkop et al., 2015; Goldman et al., 2023). Representation-based approaches learn spectrum–molecule correspondences through joint or cross-modal embedding spaces, including JESTR, CMSSP, MVP, CSU-MS^2, and FLARE (Kalia et al., 2025; Chen et al., 2024; Zhou Chen & Hassoun, 2026; Xie et al., 2025; Zhou Chen et al., 2026). Together, these approaches motivate combining efficient global retrieval with more detailed structural evidence for challenging candidate discrimination.

Self-supervised molecular representation learning. Self-supervised learning has become an effective strategy for molecular representation learning. SMILES-based and graph-based approaches learn representations through sequence modeling, graph pretraining, and masked reconstruction (Chithrananda et al., 2020; Ross et al., 2022; Hu et al., 2020; Rong et al., 2020; Wang et al., 2022; Hou et al., 2022). Recent joint-embedding predictive methods capture structural dependencies by predicting masked regions from contextual information (Assran et al., 2023; Skenderi et al., 2025; Rottach et al., 2026). Motivated by these studies, MS-CHER employs structure-aware latent prediction to obtain molecule-level and region-level representations for retrieval and structural refinement.

Adaptive retrieval and re-ranking. Multi-stage retrieval systems combine efficient initial retrieval with more expressive candidate refinement, as demonstrated by retrieve-and-rerank and late-interaction approaches (Nogueira & Cho, 2019; Khattab & Zaharia, 2020). In molecular reasoning, multi-stage systems such as MADGEN and GLMR similarly decompose an expensive prediction problem into an initial retrieval stage followed by more expressive refinement (Wang et al., 2025; Zhang et al., 2026). Recent adaptive retrieval methods further allocate retrieval or reranking computation according to query complexity (Jeong et al., 2024; Genc et al., 2026). Inspired by this perspective, MS-CHER treats fine-grained structural evidence as a query-dependent computational resource and dynamically allocates evidence according to its expected ranking utility and computational cost.

3 METHODS

3.1 PROBLEM DEFINITION AND FRAMEWORK OVERVIEW

Given an MS/MS spectrum q and its protocol-defined candidate set $\mathcal{C}_q = \{m_i\}_{i=1}^{N_q}$, molecular retrieval assigns each candidate a compatibility score $S_i = S(q, m_i)$ and sorts candidates in descending score order to obtain the ranking π_q, aiming to rank the ground-truth molecule $m^\star$ highly:

$$\pi_q = \mathrm{argsort}_{m_i \in \mathcal{C}_q} S_i. \tag{1}$$

As illustrated in Figure 2, MS-CHER uses a two-stage retrieval framework. Stage-1 computes global spectrum–molecule compatibility scores over $\mathcal{C}_q$ and produces an initial ranking. Stage-2 performs query-dependent structural refinement with two complementary evidence modules: local structural matching (*LOCAL*) models correspondence between spectral peaks and molecular regions, while fragmentation-path matching (*FRAG*) evaluates spectral support for candidate-specific fragmentation paths. A hierarchical router first selects a refinement prefix through the Scope Router and then activates *LOCAL*, *FRAG*, both, or neither through the Evidence Router according to their predicted ranking contributions and execution costs. The activated evidence is fused with the global scores to obtain the final scores over $\mathcal{C}_q$.

3.2 STRUCTURE-AWARE MOLECULAR PRETRAINING

The molecular encoder provides the molecule-level and region-level representations used by global retrieval and local structural matching. We pretrain it from molecular graphs alone with a structure-aware joint-embedding predictive objective (Assran et al., 2023; Skenderi et al., 2025).

Structural target construction. For a molecular graph $G = (V, E)$, we select a connected atom region $T \subseteq V$ with reference to its Murcko scaffold. Removing T produces the context graph $C = G[V \setminus T]$. The encoder is therefore trained to infer a missing substructure from the remaining molecular context and its attachment sites, which provides a shared representation basis for molecule-level retrieval and region-level matching.

Attachment-conditioned latent prediction. The online encoder f_θ summarizes the context graph, while an exponential-moving-average (EMA) target encoder $f_{\bar{\theta}}$ encodes the complete graph. The predictor combines the context summary $\mathbf{z}_C$, attachment representation $\mathbf{z}_A$, cut-bond representation $\mathbf{z}_{\partial E}$, and target-type embedding $\mathbf{e}_T$ to estimate the latent representation of the masked region:

$$\widehat{\mathbf{z}}_T = g_\psi(\mathbf{z}_C, \mathbf{z}_A, \mathbf{z}_{\partial E}, \mathbf{e}_T), \qquad \mathbf{z}_T = \mathrm{sg}[\mathrm{MeanPool}(f_{\bar{\theta}}(G)[T])], \tag{2}$$

where $\mathbf{z}_C$ is derived from $f_\theta(C)$ and $\mathrm{sg}[\cdot]$ stops gradients through the target branch. The target encoder is updated by EMA, and the pretraining objective is:

$$\mathcal{L}_{\mathrm{pre}} = \mathcal{L}_{\mathrm{target}} + \mathcal{L}_{\mathrm{aux}}, \qquad \mathcal{L}_{\mathrm{target}} = 1 - \cos(\widehat{\mathbf{z}}_T, \mathbf{z}_T), \tag{3}$$

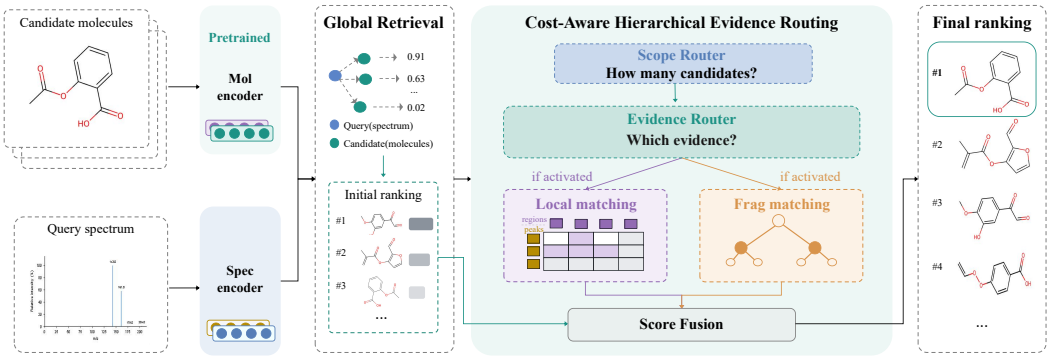

Figure 2: **Overview of MS-CHER**. Stage-1 performs global spectrum–molecule retrieval to obtain an initial candidate ranking. Stage-2 performs query-dependent structural refinement: the Scope Router selects a refinement prefix, and the Evidence Router selectively activates *LOCAL* and *FRAG* based on predicted ranking contributions and execution costs. The activated evidence is fused with global scores to produce the final ranking.

where $\mathcal{L}_{\text{aux}}$ supplies additional structural supervision; its detailed terms are given in the appendix A.

After pretraining, the frozen EMA encoder produces atom-level representations $\mathbf{H}_i = f_{\bar{\theta}}(G_i)$. Global mean pooling yields the molecule-level representation $\mathbf{m}_i$ for Stage-1, while mean pooling over structural regions $\mathcal{R}_{i,v}$ followed by ℓ_2 normalization yields region tokens $\{\mathbf{u}_{i,v}\}_{v=1}^{M_i}$ for *LOCAL*:

$$\mathbf{m}_i = \text{MeanPool}(\mathbf{H}_i), \qquad \mathbf{u}_{i,v} = \text{Norm}_2\left(\text{MeanPool}\left(\mathbf{H}_i[\mathcal{R}_{i,v}]\right)\right), \quad v = 1, \ldots, M_i. \tag{4}$$

3.3 Global Spectrum–Molecule Retrieval

Stage-1 combines a pretrained DreaMS representation $\mathbf{d}_q$ (Bushuiev et al., 2026) with a 1-Da-binned spectrum $\mathbf{b}_q$. Separate projection towers map the two spectral representations and the molecular representation $\mathbf{m}_i$ into a shared retrieval space:

$$\mathbf{h}_q = \text{Norm}_2\left(T_D(\mathbf{d}_q) + \tanh(\alpha) T_B(\mathbf{b}_q)\right), \qquad \mathbf{h}_i = \text{Norm}_2\left(T_M(\mathbf{m}_i)\right). \tag{5}$$

Their scaled inner product defines the global compatibility score:

$$s_i^G = \eta\, \mathbf{h}_q^\top \mathbf{h}_i, \qquad \pi_q^G = \text{argsort}_{m_i \in \mathcal{C}_q}\, s_i^G, \tag{6}$$

where α is a learned scalar controlling the contribution of the binned spectrum, $\eta > 0$ is a learned logit scale, and candidates in π_q^G are ordered by decreasing score. We denote by $\rho_q(i)$ the position of candidate m_i in π_q^G. This initial ranking provides the pre-execution retrieval state for Stage-2 routing.

3.4 Cost-Aware Hierarchical Evidence Routing

Figure 3 summarizes the two routing decisions made before structural evidence execution. The Scope Router selects a query-dependent refinement prefix, after which the Evidence Router estimates the marginal ranking contribution of each evidence source and determines its activation by considering the corresponding execution cost.

Query-dependent refinement scope. For rank position r in the Stage-1 ranking, let $\mathbf{f}_{q,r}$ denote the routing features of the candidate occupying that position and $\mathbf{c}_q$ the query representation. A position-wise scoring network with shared parameters combines the candidate features, the query representation, and the mean context of the valid candidate set:

$$\ell_{q,r} = H_s\left(C_s(\mathbf{f}_{q,r}) + Q_s(\mathbf{c}_q) + \frac{1}{|\mathcal{V}_q|}\sum_{j \in \mathcal{V}_q} C_s(\mathbf{f}_{q,j})\right), \qquad \mathbf{p}_q = \text{MaskedSoftmax}_{\mathcal{V}_q}(\boldsymbol{\ell}_q), \tag{7}$$

here, $\mathcal{V}_q$ indexes the valid candidate positions in the complete Stage-1 ranking, and $p_{q,r}$ estimates the probability that the ground-truth molecule occupies position r. The refinement depth K_q is the length

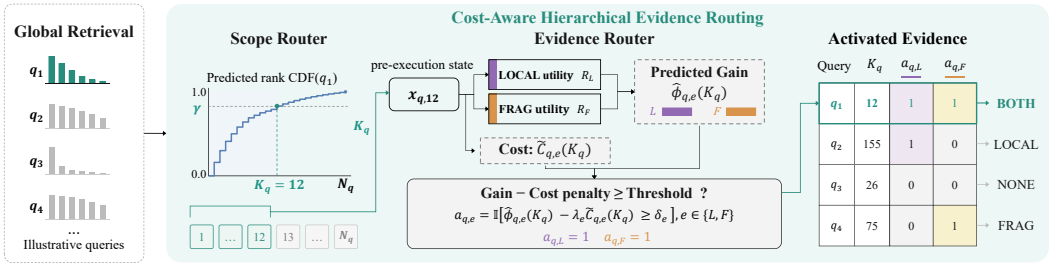

Figure 3: **Cost-aware hierarchical evidence routing.** The Scope Router selects a query-dependent refinement prefix, after which the Evidence Router activates *LOCAL* and *FRAG* according to their predicted marginal utilities and execution costs.

of the shortest prefix whose cumulative probability reaches a threshold γ. This prefix determines the candidates eligible for structural refinement, while all candidates remain in the final ranking:

$$K_q = \min\left\{k \in \{1, \ldots, N_q\} : \sum_{r=1}^{k} p_{q,r} \geq \gamma\right\}. \tag{8}$$

Scope-conditioned evidence utility. Let $\mathcal{E} = \{L, F\}$ denote *LOCAL* and *FRAG*. For an expert subset $\mathcal{S} \subseteq \mathcal{E}$ evaluated within a prefix of length K, let $r_q(\mathcal{S}; K)$ denote the ground-truth molecule's rank over the complete candidate set and define $U_q(\mathcal{S}; K) = 1/r_q(\mathcal{S}; K)$. The target utility of expert e averages its contribution when used alone and when added to the other expert:

$$\phi_{q,e}(K) = \frac{1}{2}\big[U_q(\{e\}; K) - U_q(\varnothing; K) + U_q(\mathcal{E}; K) - U_q(\mathcal{E} \setminus \{e\}; K)\big]. \tag{9}$$

An expert-specific router predicts this utility from the retrieval state available before structural evidence execution:

$$\widehat{\phi}_{q,e}(K) = R_e(\mathbf{x}_{q,K}), \qquad e \in \mathcal{E}, \tag{10}$$

where $\mathbf{x}_{q,K}$ summarizes the query representation, the Stage-1 ranking state, and the evidence availability within the prefix.

Cost-aware expert selection. Let $v^e_{q,i} \in \{0, 1\}$ indicate whether candidate m_i has valid inputs for expert e. For a prefix of length K, the number of eligible candidates and the corresponding normalized execution cost are:

$$n_{q,e}(K) = \sum_{i:\rho_q(i) \leq K} v^e_{q,i}, \qquad \widetilde{C}_{q,e}(K) = \frac{n_{q,e}(K) c_e}{N_q(c_L + c_F)}, \tag{11}$$

where c_e is the measured per-pair execution cost of expert e. The denominator defines a query-specific reference for applying both expert costs to the complete candidate set. Expert activation follows:

$$a_{q,e} = \mathbb{I}\left[\widehat{\phi}_{q,e}(K_q) - \lambda_e \widetilde{C}_{q,e}(K_q) \geq \delta_e\right]. \tag{12}$$

Here, λ_e controls the cost penalty, while δ_e is the activation threshold on the cost-adjusted utility.

Local structural matching. *LOCAL* establishes a soft correspondence between spectral peaks and molecular structural regions. Let $\mathbf{p}'_{q,u}$ and $\mathbf{u}'_{i,v}$ denote the projected representations of peak u and structural region v, respectively. Their pairwise affinity is:

$$A_{i,uv} = (\mathbf{p}'_{q,u})^\top \mathbf{u}'_{i,v}. \tag{13}$$

Dustbin entries augment the affinity matrix, allowing unmatched peaks and molecular regions under the partial-matching formulation common in optimal-transport matching (Sarlin et al., 2020). An entropy-regularized optimal-transport layer (Peyré & Cuturi, 2019) produces the transport plan:

$$\mathbf{P}_i = \mathrm{Sinkhorn}_\varepsilon\left(\widetilde{\mathbf{A}}_i; \boldsymbol{\mu}_q, \boldsymbol{\nu}_i\right), \tag{14}$$

where $\widetilde{\mathbf{A}}_i$ is the dustbin-augmented affinity matrix and $\boldsymbol{\mu}_q$ and $\boldsymbol{\nu}_i$ are the peak and region marginals. Matched and unmatched transport mass and affinity statistics yield the candidate-level score:

$$s_i^L = \frac{\mathbf{w}_L^\top \mathbf{t}(\mathbf{P}_i, \widetilde{\mathbf{A}}_i) + b_L}{\tau_L}. \tag{15}$$

$\mathbf{t}$ collects transport and affinity statistics; $\mathbf{w}_L$ and b_L are readout parameters; τ_L controls score scale.

Fragmentation-path matching. *FRAG* evaluates whether candidate-specific fragmentation paths are supported by the observed spectrum. Theoretical fragment and neutral-loss masses are matched to observed peaks under a combined absolute and relative tolerance:

$$|m_{\text{obs}} - m_{\text{theo}}| \le \max\left(\epsilon_{\text{abs}}, \epsilon_{\text{rel}} |m_{\text{theo}}|\right), \tag{16}$$

where ϵ_{abs} and ϵ_{rel} denote the absolute and relative mass tolerances. The resulting matching features are first aggregated within each fragmentation path and then pooled across valid paths:

$$s_i^F = F_{\text{cand}}\left(\text{TopPathPool}\left\{F_{\text{path}}\left(\mathbf{x}_{i,p}, \text{Pool}\{\mathbf{x}_{i,p,c}\}_c\right)\right\}_p\right), \tag{17}$$

where $\mathbf{x}_{i,p}$ describes the parent-fragment evidence of path p and $\mathbf{x}_{i,p,c}$ describes the corresponding descendant-fragment evidence.

Selective score fusion. For each query, global scores are standardized over all candidates in $\mathcal{C}_q$. Expert scores are standardized over the evaluated candidates using query-specific moments shrunk toward frozen Train-OOF reference statistics. The final score is:

$$S_i = \widetilde{s}_i^G + \mathbb{I}[\rho_q(i) \le K_q] \sum_{e \in \mathcal{E}} a_{q,e} v_{q,i}^e \beta_e \widetilde{s}_i^e, \tag{18}$$

where β_e is the fusion weight of expert e. Evaluated candidate–expert pairs contribute score corrections, and all candidates remain in the final ranking.

3.5 Training and Calibration

MS-CHER is trained sequentially. We first pretrain the molecular encoder using the structure-aware objective in Section 3.2. The Stage-1 retriever is then optimized with listwise cross-entropy and a hard-negative margin loss, while *LOCAL* and *FRAG* are trained independently with their retrieval objectives. These modules are frozen for router training. Molecule-grouped Train-OOF expert scores provide normalization statistics and, combined with validation-selected fusion weights, yield the counterfactual rankings used to compute the marginal-utility targets in Eq. (9). The Scope Router is trained with cross-entropy on the ground-truth Stage-1 position, while the Evidence Routers regress these utility targets using SmoothL1 loss. Routing thresholds and cost weights are calibrated on a validation subset. The models and calibration parameters remain fixed during testing.

4 Experiments

4.1 Experimental Settings

Datasets and evaluation metrics. We evaluate MS-CHER on three public MS/MS benchmarks: MassSpecGym (Bushuiev et al., 2024), NPLIB1 (Dührkop et al., 2021), and SpectraVerse (Gupta et al., 2026). On MassSpecGym, a general MS/MS-based molecular discovery benchmark, we evaluate precursor-mass- and molecular-formula-constrained candidate retrieval. NPLIB1 is a natural-product-oriented MS/MS dataset introduced for CANOPUS (Dührkop et al., 2021); we use its complete, untruncated formula-based candidate lists. SpectraVerse is a large-scale curated MS/MS collection integrating spectra from multiple public repositories; we evaluate precursor-mass- and molecular-formula-constrained retrieval.

We evaluate retrieval performance using Hit@k for $k \in \{1, 5, 20\}$ and mean reciprocal rank (MRR). Given the ground-truth rank r_q of query q among its candidate set and the total number of test queries Q, the metrics are defined as:

$$\text{Hit@}\,k = \frac{100}{Q} \sum_{q=1}^{Q} \mathbb{I}[r_q \le k], \qquad \text{MRR} = \frac{100}{Q} \sum_{q=1}^{Q} \frac{1}{r_q}. \tag{19}$$

Baselines. We compare MS-CHER with representative MS/MS retrieval methods. We include Random, FFN, DeepSets, and DeepSets with Fourier features from MassSpecGym (Bushuiev et al., 2024), as well as JESTR (Kalia et al., 2025), MVP (Zhou Chen & Hassoun, 2026), MIST (Goldman et al., 2023), and FLARE (Zhou Chen et al., 2026). Following the leakage-aware protocol of MassSpecGym v1.5 (Liu et al., 2026), methods in the mass setting do not use the query's ground-truth molecular formula. Therefore, methods relying on formula information are evaluated only under the formula setting. MIST and FLARE are reported for formula retrieval, while MVP adopts BinnedSpec and FormSpec for mass and formula retrieval, respectively. All baselines are evaluated under the same candidate-set construction and evaluation protocol per benchmark. We estimate 95% confidence intervals using 20,000 bootstrap resamples; complete results for three datasets are provided in Appendix B.

Training setup and efficiency measurement. The molecular encoder is initialized through structure-aware pretraining on the filtered molecule set from MassSpecGym v1.5 (Liu et al., 2026). Model selection and cost calibration use validation data only, and parameters are frozen before test evaluation. Evidence runtime is measured after GPU synchronization, following one warm-up run and three repetitions; we report the median wall-clock time per query. The timed interval covers *LOCAL* and *FRAG* evidence evaluation and excludes Stage-1 retrieval, routing, and score fusion. Experiments use two NVIDIA GeForce RTX 4090 GPUs with 24 GB memory each.

4.2 Main Retrieval Results

Table 1 shows that MS-CHER achieves the best results across all reported metrics on MassSpecGym. Under the mass setting, MS-CHER reaches 30.88% Hit@1 and 44.36% MRR, exceeding MVP by 18.65 and 23.12 percentage points, respectively. The consistent improvements in Hit@5 and Hit@20 indicate better overall candidate ordering. Under the formula setting, MS-CHER achieves 23.47% Hit@1 and 36.69% MRR, improving over FLARE by 1.26 and 1.37 percentage points.

Table 1: **MassSpecGym retrieval results under mass and formula settings.** Each column's best result is bolded; the second-best is underlined. Dashes indicate protocol-incompatible results.

	mass				formula			
Method	Hit@1	Hit@5	Hit@20	MRR	Hit@1	Hit@5	Hit@20	MRR
Random	0.43	1.94	8.06	2.43	3.02	11.44	27.96	8.40
FFN	1.56	7.29	21.54	5.76	4.85	14.51	30.54	10.57
DeepSets	1.65	6.45	19.96	5.46	4.55	14.75	30.15	10.41
DeepSets + Fourier	3.49	11.45	27.34	8.65	5.86	15.64	34.40	11.91
JESTR	11.87	27.96	<u>51.23</u>	20.77	12.50	32.27	57.60	22.77
MVP	<u>12.23</u>	<u>29.27</u>	50.97	<u>21.24</u>	11.33	31.43	62.30	22.05
MIST	–	–	–	–	10.53	24.66	47.04	18.35
FLARE	–	–	–	–	<u>22.21</u>	<u>49.95</u>	<u>75.12</u>	<u>35.32</u>
MS-CHER	**30.88**	**59.90**	**81.82**	**44.36**	**23.47**	**51.54**	**75.90**	**36.69**

To assess cross-dataset transfer, we directly evaluate the trained models on SpectraVerse and NPLIB1 without additional training or fine-tuning. Despite the distribution shift, MS-CHER maintains strong performance on SpectraVerse(Table 2). MS-CHER achieves the highest scores across all metrics in the mass setting. Its Hit@1 and MRR exceed the strongest baseline, DeepSets with Fourier features, by 10.12 and 13.50 percentage points. In the formula setting, MS-CHER obtains the highest MRR with competitive Hit@1. These results indicate MS-CHER maintains strong retrieval performance under a shifted candidate distribution.

4.3 Ablation Studies

Effect of Stage-2 Refinement. We evaluate the overall contribution of the Stage-2 refinement pipeline. Stage-1 ranks candidates using the global compatibility score s_i^G, whereas the complete MS-CHER model performs query-dependent structural refinement through hierarchical evidence

Table 2: **SpectraVerse retrieval results under mass and formula settings.** Each column's best result is bolded; the second-best is underlined. Dashes indicate protocol-incompatible results.

	mass				formula			
Method	Hit@1	Hit@5	Hit@20	MRR	Hit@1	Hit@5	Hit@20	MRR
Random	0.67	2.95	9.44	2.72	1.07	4.58	12.56	3.71
FFN	3.79	11.35	23.87	8.35	3.94	13.13	26.52	9.28
DeepSets	4.94	13.40	27.46	10.03	5.19	15.26	29.91	11.06
DeepSets + Fourier	<u>19.25</u>	<u>31.78</u>	<u>47.21</u>	<u>25.92</u>	19.41	32.93	48.73	26.53
JESTR	3.03	7.80	19.76	6.62	3.30	8.88	21.94	7.32
MVP	6.98	19.56	35.75	13.67	3.04	11.76	28.46	8.38
MIST	–	–	–	–	**31.17**	<u>39.18</u>	46.75	<u>35.16</u>
FLARE	–	–	–	–	16.65	37.71	<u>56.29</u>	26.68
MS-CHER	**29.37**	**50.79**	**66.11**	**39.42**	<u>29.67</u>	**49.91**	**66.52**	**39.36**

routing and selective *LOCAL/FRAG* evaluation. Both variants share the same candidate sets and frozen Stage-1 model, allowing the comparison to isolate the contribution of the refinement pipeline.

As shown in Table 3, adding Stage-2 structural refinement improves Hit@1 by 12.96 and 8.80 percentage points under the mass and formula settings, respectively, while MRR increases by 14.65 and 11.51 points. Consistent improvements in Hit@5 and Hit@20 show that the benefits extend beyond first-position identification to recovering correct molecules within larger candidate shortlists. These gains demonstrate that Stage-2 provides discriminative structural evidence beyond global compatibility, substantially improving the initial candidate ordering under both retrieval settings.

Table 3: **Effect of Stage-2 refinement on MassSpecGym.** Bold values indicate better performance.

	mass				formula			
Method	Hit@1	Hit@5	Hit@20	MRR	Hit@1	Hit@5	Hit@20	MRR
Stage-1	17.92	42.50	65.77	29.71	14.67	35.47	59.83	25.18
MS-CHER	**30.88**	**59.90**	**81.82**	**44.36**	**23.47**	**51.54**	**75.90**	**36.69**

Analysis of Hierarchical Routing. This ablation examines how candidate-scope and evidence-type selection contribute to the trade-off between retrieval performance and evidence-execution cost. By removing each router separately, we assess its role. All variants share the same frozen Stage-1 model, structural experts, and score-fusion parameters. *w/o Scope Router* sets $K_q = N_q$ while retaining the Evidence Router. *w/o Evidence Router* retains the query-dependent scope but evaluates all eligible *LOCAL* and *FRAG* evidence pairs within that scope. Full-Eval removes both routing decisions by setting $K_q = N_q$ and evaluating all eligible candidate–expert pairs.

Table 4: **Ablation of hierarchical routing policies on MassSpecGym.** Bold indicates the best retrieval score or lowest evidence-execution cost. Second-best values are underlined.

Setting	Method	Hit@1	Hit@5	Hit@20	MRR	Relative Runtime	Evidence s/query
Mass	w/o Scope Router	29.89	58.72	81.62	43.29	<u>0.2502</u>	<u>0.2085</u>
	w/o Evidence Router	<u>30.71</u>	<u>60.13</u>	<u>82.22</u>	**44.38**	0.5405	0.4504
	Full-Eval	30.50	**60.14**	**82.60**	44.30	1.0000	0.8334
	MS-CHER	**30.88**	59.90	81.82	<u>44.36</u>	**0.2076**	**0.1730**
Formula	w/o Scope Router	23.07	51.10	76.34	36.32	<u>0.2140</u>	<u>0.0954</u>
	w/o Evidence Router	23.15	<u>51.62</u>	<u>76.46</u>	**37.36**	0.7988	0.3561
	Full-Eval	**23.70**	**51.77**	**76.92**	<u>36.99</u>	1.0000	0.4458
	MS-CHER	<u>23.47</u>	51.54	75.90	36.69	**0.1760**	**0.0785**

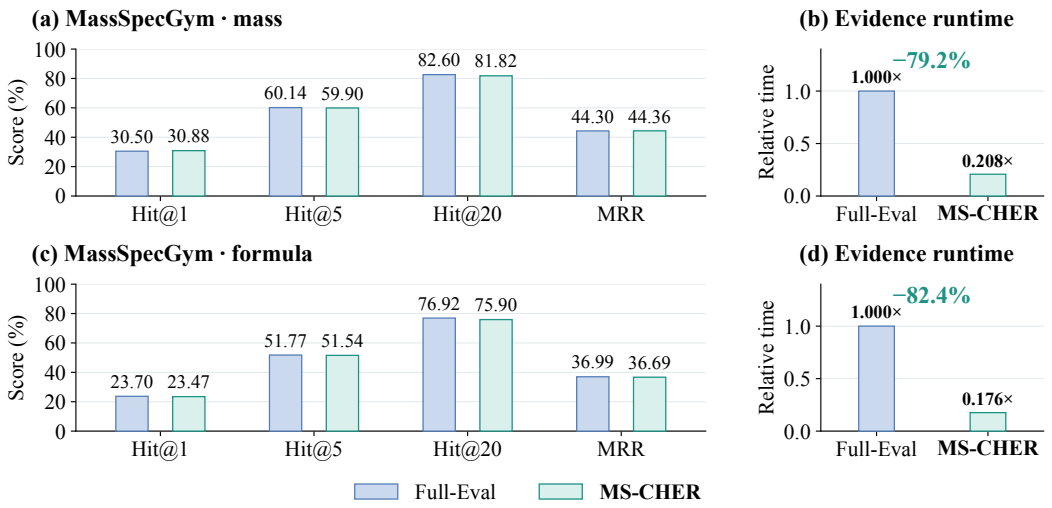

Figure 4: **Full-Eval and MS-CHER retrieval performance and measured evidence-execution runtime on MassSpecGym.** Runtime is normalized to setting-specific Full-Eval.

Table 4 shows that removing the Scope Router lowers Hit@1 and MRR while increasing evidence runtime in both settings, supporting query-dependent scope selection within the joint routing policy. Removing the Evidence Router substantially increases runtime and improves Hit@5, Hit@20, and MRR, but slightly lowers Hit@1. Thus, selective activation reduces evidence runtime at the expense of some ranking metrics, rather than uniformly improving retrieval. The Scope Router determines where refinement occurs, while the Evidence Router selects which structural evidence to evaluate. Together, they yield the lowest evidence runtime among the compared variants in both settings, while retaining retrieval performance close to Full-Eval.

Figure 4 compares MS-CHER with exhaustive evidence evaluation. Relative to Full-Eval, MS-CHER changes Hit@1 by $+0.38$ and -0.23 percentage points under the mass and formula settings, respectively, while using $0.2076\times$ and $0.1760\times$ of its measured evidence-execution runtime. The MRR differences are $+0.06$ and -0.30 points. Thus, MS-CHER retains retrieval performance close to Full-Eval while substantially reducing the time spent evaluating structural evidence. These findings suggest that the ranking benefits of structural evidence can largely be retained without exhaustive evaluation of every candidate–expert pair. By jointly adapting the refinement scope and expert activation to each query, MS-CHER allocates structural computation according to expected ranking benefits and execution costs, enabling more efficient use of fine-grained evidence for molecule retrieval.

5 Conclusion

We presented MS-CHER, a two-stage molecular retrieval framework that enhances global spectrum–molecule matching with query-adaptive structural evidence refinement. By integrating complementary evidence from local structural matching and fragmentation-path matching, MS-CHER improves retrieval accuracy beyond global similarity modeling. Cost-Aware Hierarchical Evidence Routing further enables efficient evidence utilization by adaptively allocating refinement computation according to ranking benefits and costs.

Experiments across multiple benchmarks demonstrate strong retrieval performance, while the MassSpecGym analysis shows that hierarchical routing retains performance close to exhaustive refinement with substantially lower evidence-execution cost. These results highlight the effectiveness of combining structural evidence with adaptive computation for scalable molecular retrieval. Future extensions may incorporate richer molecular evidence sources and further improve robustness under challenging retrieval scenarios with incomplete or ambiguous structural information.

AI USE STATEMENT

During the preparation of this manuscript, we used generative AI tools to assist with language editing and translation, literature discovery, manuscript organization, and the presentation of the proposed methodology. We also used these tools to obtain feedback on experimental design and result interpretation, and to assist with code debugging and implementation verification. All such outputs were treated as suggestions rather than authoritative results.

The final methodological choices, experimental protocols, and scientific claims were determined by the authors. All reported experiments and numerical results were produced from the authors' implementations, and AI-assisted interpretations, citations, and manuscript content were reviewed and verified before inclusion. We take full responsibility for the accuracy, originality, and integrity of the submitted work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the proposed framework, training procedures, and evaluation protocols in the main paper and supplementary material. The appendix includes additional implementation details, dataset processing information, experimental settings, and ablation configurations. All experiments are conducted using publicly available datasets, including MassSpecGym, NPLIB1, and SpectraVerse, following the corresponding evaluation protocols described in the paper. The released resources will include the necessary components for reproducing the reported results, including model training, inference, and performance evaluation.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499. URL https://doi.org/10.1109/CVPR52729.2023.01499.

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*, 2024. URL https://arxiv.org/abs/2410.23326.

Roman Bushuiev, Anton Bushuiev, Raman Samusevich, Corinna Brungs, Josef Sivic, and Tomáš Pluskal. Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. *Nature Biotechnology*, 44(4):630–640, 2026. doi: 10.1038/s41587-025-02663-3. URL https://doi.org/10.1038/s41587-025-02663-3.

Lu Chen, Bing Xia, Yu Wang, Xia Huang, Yucheng Gu, Wenlin Wu, and Yan Zhou. CMSSP: A Contrastive Mass Spectra-Structure Pretraining Model for Metabolite Identification. *Analytical Chemistry*, 96(42):16871–16881, 2024. doi: 10.1021/acs.analchem.4c03724.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. arXiv preprint arXiv:2010.09885, 2020. URL https://arxiv.org/abs/2010.09885.

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

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

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. URL https://doi.org/10.1038/s41587-020-0740-8.

Ata Cinar Genc, Emir Kaan Korukluoglu, and James Allan. Adaptive Re-Ranking. arXiv preprint arXiv:2606.25249, 2026. URL https://arxiv.org/abs/2606.25249.

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(9):965–979, 2023. doi: 10.1038/s42256-023-00708-3. URL https://doi.org/10.1038/s42256-023-00708-3.

Vishu Gupta, Hantao Qiang, Hsin-Hsiang Chung, Ehud Herbst, and Michael A. Skinnider. Comprehensive Curation and Harmonization of Small-Molecule MS/MS Libraries in Spectraverse. *Analytical Chemistry*, 98(5):3934–3943, 2026. doi: 10.1021/acs.analchem.5c06256. URL https://doi.org/10.1021/acs.analchem.5c06256.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. GraphMAE: Self-Supervised Masked Graph Autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 594–604, 2022. doi: 10.1145/3534678.3539321. URL https://doi.org/10.1145/3534678.3539321.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for Pre-training Graph Neural Networks. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1905.12265.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, 2024. doi: 10.18653/v1/2024.naacl-long.389. URL https://doi.org/10.18653/v1/2024.naacl-long.389.

Apurva Kalia, Yan Zhou Chen, Dilip Krishnan, and Soha Hassoun. JESTR: Joint Embedding Space Technique for Ranking candidate molecules for the annotation of untargeted metabolomics data. *Bioinformatics*, 41(7):btaf354, 2025. doi: 10.1093/bioinformatics/btaf354. URL https://doi.org/10.1093/bioinformatics/btaf354.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

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.

Rodrigo Nogueira and Kyunghyun Cho. Passage Re-ranking with BERT. arXiv preprint arXiv:1901.04085, 2019. URL https://arxiv.org/abs/1901.04085.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. doi: 10.1561/2200000073. URL https://doi.org/10.1561/2200000073.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-Supervised Graph Transformer on Large-Scale Molecular Data. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2007.02835.

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

Florian Rottach, Sebastian Schieferdecker, William Rudman, Randall Balestriero, and Carsten Eickhoff. Mol-JEPA: A multimodal Joint Embedding Predictive Architecture for Molecules. arXiv preprint arXiv:2608.22642, 2026. URL https://arxiv.org/abs/2608.22642.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning Feature Matching With Graph Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4937–4946, 2020. doi: 10.1109/CVPR42600.2020.00499. URL https://doi.org/10.1109/CVPR42600.2020.00499.

Geri Skenderi, Hang Li, Jiliang Tang, and Marco Cristani. Graph-level Representation Learning with Joint-Embedding Predictive Architectures. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=v47f4DwYZb.

Hiroshi Tsugawa, Tomas Cajka, Tobias Kind, Yan Ma, Brendan Higgins, Kazutaka Ikeda, Mitsuhiro Kanazawa, Jean VanderGheynst, Oliver Fiehn, and Masanori Arita. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods*, 12(6):523–526, 2015. doi: 10.1038/nmeth.3393. URL https://doi.org/10.1038/nmeth.3393.

Mingxun Wang, Jeremy J. Carver, Vanessa V. Phelan, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology*, 34(8):828–837, 2016. doi: 10.1038/nbt.3597. URL https://doi.org/10.1038/nbt.3597.

Yinkai Wang, Xiaohui Chen, Liping Liu, and Soha Hassoun. MADGEN: Mass-Spec Attends to De Novo Molecular Generation. In *International Conference on Learning Representations*, 2025.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3): 279–287, 2022. doi: 10.1038/s42256-022-00447-x. URL https://doi.org/10.1038/s42256-022-00447-x.

Ting Xie, Hailiang Zhang, Qiong Yang, Jinyu Sun, Yue Wang, Jia Long, Zhimin Zhang, and Hongmei Lu. CSU-MS2: A Contrastive Learning Framework for Cross-Modal Compound Identification from MS/MS Spectra to Molecular Structures. *Analytical Chemistry*, 97(25):13350–13360, 2025. doi: 10.1021/acs.analchem.5c01594.

Yiwen Zhang, Keyan Ding, Yihang Wu, Xiang Zhuang, Yi Yang, Qiang Zhang, and Huajun Chen. Breaking the modality barrier: Generative modeling for accurate molecule retrieval from mass spectra. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1561–1569, 2026. doi: 10.1609/aaai.v40i2.37132.

Yan Zhou Chen and Soha Hassoun. Learning from All Views: A Multiview Contrastive Framework for Metabolite Annotation. *Analytical Chemistry*, 98(9):6598–6606, 2026. doi: 10.1021/acs.analchem.5c05675. URL https://doi.org/10.1021/acs.analchem.5c05675.

Yan Zhou Chen, Blake R. Rushing, and Soha Hassoun. FLARE: Fine-grained Learning for Alignment of spectra-molecule REpresentation Enhances Metabolite Annotation. bioRxiv preprint, 2026. URL https://doi.org/10.64898/2026.01.27.702086.

A Additional Method and Evaluation Details

A.1 Molecular Pretraining

Structure-aware target construction. For each molecular graph, one connected target region is sampled at each training epoch using a structure-aware priority. The molecular core is defined by the Murcko scaffold; when no valid scaffold is available, a compact graph-center region is used as a fallback. We first sample an eligible non-core branch connected to the core through a single boundary bond. If no such branch exists, a valid ring system is selected after merging rings that share atoms. When neither structural target is available, a connected region is generated by breadth-first expansion from a randomly selected atom. The resulting target is required to be a proper subset of the molecular graph. The sampling policy is summarized in Table 5.

Table 5: Structural target sampling used for molecular pretraining.

Priority	Target type	Construction
1	Branch	Eligible single-attachment non-core branch
2	Ring	Valid merged ring system
3	Connected region	Random connected 2–6 atom region
4	None	Regional prediction objective is skipped

After sampling target T, its atoms and incident edges are removed to form the context graph $C = G[V \setminus T]$. Context-side atoms adjacent to T provide the attachment representation, while the removed boundary bonds provide complementary bond information. These representations are combined with the global context representation and the target-type embedding by the predictor described in the main paper. Unlike random atom masking, this construction requires prediction of a coherent missing structural region from the remaining molecular context.

Pretraining objectives. In addition to the regional latent-prediction objective $\mathcal{L}_{\text{target}}$ defined in the main paper, we employ global latent prediction, masked atom recovery, masked bond recovery, and variance–covariance regularization. The global objective predicts the EMA representation pooled over the complete molecule. Atom and bond recovery randomly mask 15% of the corresponding inputs and reconstruct their original classes using cross-entropy supervision. The variance and covariance terms regularize batch-level global predictions by encouraging sufficient variation across representation dimensions while suppressing redundant cross-dimensional correlation.

The complete molecular pretraining objective is

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{target}} + \lambda_g \mathcal{L}_{\text{global}} + \lambda_a \mathcal{L}_{\text{atom}} + \lambda_b \mathcal{L}_{\text{bond}} + \lambda_v \mathcal{L}_{\text{var}} + \lambda_c \mathcal{L}_{\text{cov}}. \tag{20}$$

The fixed loss weights used throughout pretraining are reported in Table 6.

Table 6: Loss weights used for molecular pretraining.

Term	Target	Global	Atom	Bond	Var.	Cov.
Weight	1.00	0.20	0.25	0.25	0.05	0.01

A.2 Global Retrieval and Structural Evidence

Stage-1 training. For query q, let $i^\star$ denote the ground-truth candidate index. Stage-1 combines candidate-level listwise cross-entropy with a hard-negative margin loss:

$$\begin{aligned} \mathcal{L}^G_{\text{list}} &= -\log \frac{\exp(s^G_{i^\star})}{\sum_{j=1}^{N_q} \exp(s^G_j)}, \\ \mathcal{L}^G_{\text{margin}} &= \frac{1}{|\mathcal{H}_q|} \sum_{j \in \mathcal{H}_q} \max\left(0, m - s^G_{i^\star} + s^G_j\right). \end{aligned} \tag{21}$$

Here, $\mathcal{H}_q$ contains the indices of the 16 highest-scoring negative candidates, and $m = 0.1$. The training objective is $\mathcal{L}_G = \mathcal{L}^G_{\text{list}} + 0.2\,\mathcal{L}^G_{\text{margin}}$.

LOCAL evidence. *LOCAL* uses a deterministic dictionary of overlapping structural regions comprising the whole molecule, the molecular core, boundary anchors, connected non-core branches, ring components, and single-atom fallback regions. Regions are retained in a fixed priority order up to the specified capacity. Each token is obtained by mean-pooling the corresponding atom representations from the frozen molecular encoder and applying ℓ_2 normalization.

The dustbin-augmented transport plan is summarized by matched affinity, matched assignment mass, unmatched peak mass, and unmatched region mass. These four statistics form $\mathbf{t}_i = \mathbf{t}(\mathbf{P}_i, \widetilde{\mathbf{A}}_i)$ used by the *LOCAL* scoring function. *LOCAL* is trained with bidirectional spectrum–molecule contrastive supervision and candidate-level listwise ranking:

$$\mathcal{L}_L = \frac{1}{2}\left[\mathrm{CE}(\mathbf{S}, \mathbf{I}) + \mathrm{CE}(\mathbf{S}^\top, \mathbf{I})\right] + 0.5\,\mathcal{L}^L_{\mathrm{list}}, \tag{22}$$

where $\mathbf{S}$ is the batch spectrum–molecule score matrix, $\mathbf{I}$ specifies the matched pairs, and $\mathcal{L}^L_{\mathrm{list}}$ is the listwise loss over query-specific candidates.

FRAG evidence. *FRAG* constructs candidate fragmentation paths through two successive bond cleavages. Eligible bonds are non-aromatic single bonds outside rings, and each cleavage must divide the current graph into exactly two connected components. Both components from the first cleavage are considered as first-step fragments, within which a second cleavage generates descendant fragments. Bonds involving heteroatoms and terminal substituents receive higher priority, followed by stable structural tie-breaking.

For each fragment, matching descriptors summarize direct fragment-ion support, neutral-loss support, their joint evidence, matching coverage, mass accuracy, and high-intensity peak support. Mass matching follows Eq. (16), using an absolute tolerance of 0.02 Da and a relative tolerance of 20 ppm. Descendant evidence is aggregated within each first-step branch and combined with parent-fragment evidence and acquisition information. A learned scorer then selects and pools the highest-scoring paths to obtain the candidate-level *FRAG* score.

The *FRAG* ranking objective is listwise cross-entropy over candidates with valid fragmentation evidence:

$$\mathcal{L}_F = -\log \frac{\exp(s^F_{i^\star})}{\sum_{j:\, m_j \in \mathcal{C}^F_q} \exp(s^F_j)}, \tag{23}$$

where $\mathcal{C}^F_q$ denotes the valid candidate set for *FRAG*.

A.3 Router Supervision and Calibration

The training split is partitioned into five folds grouped by molecular identity, with all spectra associated with the same molecule assigned to the same fold. *LOCAL* and *FRAG* are trained on four folds and evaluated on the held-out fold to obtain Train-OOF expert scores for router supervision. The final experts used for validation and testing are refitted on the complete training split. Stage-1 remains frozen during router training.

For each training query and sampled refinement depth, we construct counterfactual rankings using no expert, *LOCAL* only, *FRAG* only, and both experts. Ground-truth ranks are computed over the complete candidate set and converted into marginal-utility targets using Eq. (9). The Scope Router uses cross-entropy supervision on the ground-truth Stage-1 position, while the Evidence Routers regress these utility targets using SmoothL1 loss. Routing inputs comprise query features, Stage-1 ranking statistics, prefix statistics, and evidence availability. These inputs are available before expert execution and contain neither expert scores nor ground-truth ranks.

Table 7 summarizes the principal training settings, and Table 8 provides the router architecture and additional optimization settings. The mass and formula settings use identical router architectures, loss weights, and training configurations. *LOCAL* and *FRAG* have separate utility-prediction networks with the same architecture.

The validation split is divided into non-overlapping selection and calibration subsets. The selection subset is used to select fusion weights and router checkpoints. Fusion weights are fixed before constructing counterfactual training targets. Expert-score normalization uses query-specific moments shrunk toward frozen Train-OOF reference statistics with a pseudocount of 4.

Table 7: Principal training settings. Router training settings are identical for mass and formula.

Component	Learning rate	Batch size	Training duration
Stage-1	3×10^{-4}	256	$\leq$ 20,000 steps
LOCAL	2×10^{-4}	8	1,500 steps
FRAG	2×10^{-4}	4	1,500 steps
Routers	3×10^{-4}	64	6 epochs

Table 8: Router architecture and additional training settings shared by mass and formula.

Configuration	Value
Candidate / query / evidence-state dimensions	5 / 27 / 61
Scope candidate encoder	$5 \to 128 \to 128$
Scope query projection	$27 \to 128$
Scope position scorer	$128 \to 128 \to 1$
Each expert-utility network	$61 \to 256 \to 128 \to 1$
Dropout rate	0.1
Scope / LOCAL-utility / FRAG-utility loss weights	1.0 / 1.0 / 1.0
Weight decay	0.01
Gradient clipping threshold	5.0
Training precision	bfloat16

The calibration subset is used to measure per-pair expert costs and select an accuracy–cost operating point. Among the evaluated routing policies, we choose the one with the lowest mean normalized cost of activated evidence subject to calibration decreases of at most 0.25 percentage points in Hit@1 and MRR and 0.50 percentage points in Hit@5 relative to Full-Eval. These tolerances govern policy selection on the calibration subset and do not guarantee test performance. Model weights, fusion weights, reference normalization statistics, measured cost coefficients, and routing-policy parameters are frozen before test evaluation.

A.4 Inference and Runtime Measurement

Algorithm 1 summarizes inference. The refinement scope and expert activations are determined before structural evidence execution. Activated experts evaluate only candidates with valid inputs inside the selected prefix, and their standardized scores contribute corrections according to Eq. (18). All remaining candidates retain their standardized global scores and participate in the final ranking. If neither expert is activated, the Stage-1 ordering is preserved.

Evidence runtime is measured using wall-clock timing with GPU synchronization before and after the timed interval. Following one warm-up run, we perform three timed repetitions and report the median duration divided by the number of evaluated queries. The timed interval covers *LOCAL* and *FRAG* evidence evaluation. Relative evidence runtime is the ratio of the measured per-query evidence time to that of Full-Eval under the same candidate setting. Full-Eval evaluates both experts over all valid candidate–expert pairs in the complete candidate set. This runtime ratio is computed from actual execution times, whereas the normalized cost in Eq. (11) is an estimate used for routing and policy selection. The reported runtime savings therefore concern structural-evidence evaluation.

B Additional Experimental Results

B.1 Dataset Details

MassSpecGym and SpectraVerse are evaluated under both precursor-mass and molecular-formula constraints. For NPLIB1, we use the complete formula-based candidate lists without truncation and do not evaluate a separate mass setting. Table 9 summarizes dataset statistics and average candidate-set sizes.

Algorithm 1 MS-CHER inference

Require: Query q, candidate set $\mathcal{C}_q$, frozen models and calibration parameters
Ensure: Final ranking over $\mathcal{C}_q$

```
Compute global scores and Stage-1 rank positions ρ_q(i)
Standardize global scores; initialize S_i ← s̃_i^G
Select prefix length K_q using Eq. (8)
Construct the pre-execution state x_{q,K_q}
Predict expert utilities φ̂_{q,e}(K_q) for e ∈ {L, F}
for all e ∈ {L, F} do
    J_{q,e} ← {i : ρ_q(i) ≤ K_q, v^e_{q,i} = 1}
    Compute C̃_{q,e}(K_q) using Eq. (11)
    Determine a_{q,e} using Eq. (12)
end for
for all e ∈ {L, F} do
    if a_{q,e} = 1 and J_{q,e} ≠ ∅ then
        Evaluate and standardize expert scores on J_{q,e}
        S_i ← S_i + β_e s̃^e_i for i ∈ J_{q,e}
    end if
end for
return All candidates sorted by decreasing S_i
```

Table 9: **Dataset statistics and candidate-set configurations.** Dashes indicate settings not evaluated.

Dataset	# Spectra	# Structures	Avg. cand. (Mass)	Avg. cand. (Formula)
MassSpecGym	231,104	28,936	252.85	212.06
NPLIB1	8,203	7,130	–	2,207.70
SpectraVerse	488,797	44,304	3,075.41	1,440.05

B.2 Complete Results

Evaluation and bootstrap confidence intervals. Tables 10, 11, and 12 report the complete results for 41 method–setting configurations across MassSpecGym, NPLIB1, and SpectraVerse. Each entry contains a point estimate and a 95% confidence interval obtained from 20,000 bootstrap resamples. Hit@1, Hit@5, Hit@20, and MRR are reported as percentages, with higher values indicating better retrieval performance. These metrics include queries with empty retrieval outputs, which are counted as failures. Bold and underlined retrieval scores indicate the best and second-best point estimates within each setting; these markings do not denote statistical significance. Across the five dataset–setting combinations, MS-CHER achieves the highest point estimates on 15 of the 20 retrieval metrics. The gains are particularly pronounced under mass constraints, while performance remains competitive under formula constraints: MS-CHER leads all four metrics on MassSpecGym and three on SpectraVerse, and ranks second in all three Hit metrics on NPLIB1. These results support the effectiveness of MS-CHER across multiple benchmarks and candidate-set configurations.

B.3 Ablation Studies

Structural Evidence Ablation. Table 13 compare Stage-1 with *LOCAL* only, *FRAG* only, and both experts to examine the contribution of structural evidence. These comparisons use the same frozen Stage-1 model, expert parameters, candidate sets, and fusion weights. Evidence is evaluated over the complete candidate scope without routing; Full-Eval applies both experts to all valid candidate–expert pairs. This separates the ranking contribution of the evidence sources from the execution policy examined in the main paper.

Both evidence sources improve Stage-1 retrieval, with *FRAG* providing the larger gains. Combining *LOCAL* with *FRAG* further increases Hit@1 by 0.95 and 0.58 percentage points under the mass and formula settings, respectively, and increases MRR by 1.29 and 0.67 points. Full-Eval also obtains the highest Hit@5 and Hit@20 among these configurations. These results identify *FRAG* as the main source of the structural-refinement gains, with *LOCAL* providing additional ranking information.

Table 10: **Complete results on MassSpecGym.** Each cell reports a point estimate with its 95% bootstrap confidence interval in brackets.

Method	Hit@1 (%)	Hit@5 (%)	Hit@20 (%)	MRR (%)
		mass		
Random	0.43 [0.34, 0.53]	1.94 [1.74, 2.15]	8.06 [7.66, 8.46]	2.43 [2.32, 2.55]
FFN	1.56 [1.38, 1.75]	7.29 [6.90, 7.67]	21.54 [20.94, 22.15]	5.76 [5.55, 5.97]
DeepSets	1.65 [1.46, 1.84]	6.45 [6.09, 6.82]	19.96 [19.37, 20.55]	5.46 [5.24, 5.67]
DeepSets + Fourier	3.49 [3.22, 3.75]	11.45 [10.98, 11.92]	27.34 [26.68, 28.00]	8.65 [8.36, 8.94]
JESTR	11.87 [11.38, 12.35]	27.96 [27.28, 28.62]	51.23 [50.49, 51.98]	20.77 [20.29, 21.24]
MVP	12.23 [11.75, 12.72]	29.27 [28.60, 29.94]	50.97 [50.23, 51.70]	21.24 [20.75, 21.72]
MS-CHER	**30.88** [30.20, 31.58]	**59.90** [59.16, 60.61]	**81.82** [81.24, 82.39]	**44.36** [43.77, 44.94]
		formula		
Random	3.02 [2.77, 3.28]	11.44 [10.98, 11.92]	27.96 [27.30, 28.63]	8.40 [8.12, 8.68]
FFN	4.85 [4.54, 5.18]	14.51 [14.00, 15.03]	30.54 [29.87, 31.22]	10.57 [10.23, 10.91]
DeepSets	4.55 [4.25, 4.86]	14.75 [14.23, 15.28]	30.15 [29.48, 30.83]	10.41 [10.09, 10.74]
DeepSets + Fourier	5.86 [5.52, 6.21]	15.64 [15.11, 16.19]	34.40 [33.69, 35.10]	11.91 [11.56, 12.28]
JESTR	12.50 [12.01, 12.99]	32.27 [31.58, 32.96]	57.60 [56.87, 58.33]	22.77 [22.29, 23.24]
MVP	11.33 [10.86, 11.80]	31.43 [30.75, 32.12]	62.30 [61.58, 63.00]	22.05 [21.59, 22.50]
MIST	10.53 [10.08, 10.98]	24.66 [24.04, 25.30]	47.04 [46.30, 47.78]	18.35 [17.90, 18.80]
FLARE	22.21 [21.60, 22.82]	49.95 [49.21, 50.69]	75.12 [74.48, 75.74]	35.32 [34.77, 35.87]
MS-CHER	**23.47** [22.84, 24.10]	**51.54** [50.80, 52.28]	**75.90** [75.27, 76.53]	**36.69** [36.13, 37.25]

Table 11: **Complete results on SpectraVerse.** Each cell reports a point estimate with its 95% bootstrap confidence interval in brackets.

Method	Hit@1 (%)	Hit@5 (%)	Hit@20 (%)	MRR (%)
		mass		
Random	0.67 [0.60, 0.74]	2.95 [2.81, 3.09]	9.44 [9.19, 9.69]	2.72 [2.64, 2.80]
FFN	3.79 [3.63, 3.95]	11.35 [11.09, 11.62]	23.87 [23.52, 24.22]	8.35 [8.18, 8.52]
DeepSets	4.94 [4.77, 5.13]	13.40 [13.12, 13.69]	27.46 [27.09, 27.83]	10.03 [9.84, 10.22]
DeepSets + Fourier	19.25 [18.92, 19.58]	31.78 [31.40, 32.17]	47.21 [46.79, 47.63]	25.92 [25.60, 26.23]
JESTR	3.03 [2.89, 3.17]	7.80 [7.58, 8.02]	19.76 [19.42, 20.09]	6.62 [6.47, 6.78]
MVP	6.98 [6.77, 7.19]	19.56 [19.22, 19.88]	35.75 [35.35, 36.15]	13.67 [13.45, 13.89]
MS-CHER	**29.37** [28.99, 29.74]	**50.79** [50.37, 51.20]	**66.11** [65.72, 66.51]	**39.42** [39.08, 39.77]
		formula		
Random	1.07 [0.99, 1.16]	4.58 [4.41, 4.76]	12.56 [12.28, 12.84]	3.71 [3.61, 3.81]
FFN	3.94 [3.78, 4.10]	13.13 [12.85, 13.41]	26.52 [26.16, 26.88]	9.28 [9.11, 9.46]
DeepSets	5.19 [5.01, 5.38]	15.26 [14.96, 15.56]	29.91 [29.53, 30.30]	11.06 [10.86, 11.26]
DeepSets + Fourier	19.41 [19.07, 19.73]	32.93 [32.54, 33.32]	48.73 [48.32, 49.16]	26.53 [26.21, 26.84]
JESTR	3.30 [3.15, 3.45]	8.88 [8.64, 9.11]	21.94 [21.59, 22.28]	7.32 [7.16, 7.48]
MVP	3.04 [2.90, 3.19]	11.76 [11.50, 12.03]	28.46 [28.08, 28.83]	8.38 [8.23, 8.54]
MIST	**31.17** [30.78, 31.56]	39.18 [38.78, 39.59]	46.75 [46.34, 47.16]	35.16 [34.79, 35.53]
FLARE	16.65 [16.34, 16.96]	37.71 [37.31, 38.12]	56.29 [55.88, 56.71]	26.68 [26.39, 26.98]
MS-CHER	29.67 [29.29, 30.05]	**49.91** [49.49, 50.33]	**66.52** [66.13, 66.92]	**39.36** [39.01, 39.72]

Table 12: **Complete results on NPLIB1 under the formula setting.** Each cell reports a point estimate with its 95% bootstrap confidence interval in brackets.

Method	Hit@1 (%)	Hit@5 (%)	Hit@20 (%)	MRR (%)
Random	1.83 [0.98, 2.81]	7.81 [6.10, 9.65]	16.85 [14.29, 19.41]	5.35 [4.32, 6.45]
FFN	1.71 [0.85, 2.69]	8.91 [6.96, 10.87]	20.15 [17.46, 22.83]	5.60 [4.61, 6.68]
DeepSets	1.83 [0.98, 2.81]	9.52 [7.57, 11.60]	18.68 [16.00, 21.37]	5.82 [4.77, 6.96]
DeepSets + Fourier	10.74 [8.67, 12.94]	24.05 [21.25, 26.98]	38.10 [34.80, 41.39]	17.39 [15.27, 19.57]
JESTR	8.42 [6.59, 10.38]	25.76 [22.71, 28.82]	43.96 [40.66, 47.37]	17.06 [15.12, 19.07]
MVP	13.43 [11.11, 15.75]	36.63 [33.33, 39.93]	61.42 [58.00, 64.71]	24.97 [22.74, 27.26]
MIST	**35.41** [32.11, 38.71]	59.83 [56.41, 63.13]	76.68 [73.75, 79.49]	46.61 [43.68, 49.44]
FLARE	31.38 [28.33, 34.55]	**70.45** [67.28, 73.50]	**88.64** [86.45, 90.72]	**48.55** [46.00, 51.13]
MS-CHER	32.36 [29.18, 35.53]	62.39 [59.10, 65.69]	81.68 [79.00, 84.25]	46.11 [43.40, 48.82]

Table 13: Structural evidence ablation on MassSpecGym. All metrics are reported in percent. Bold values indicate the best result within each setting.

Setting	Configuration	Hit@1 (%)	Hit@5 (%)	Hit@20 (%)	MRR (%)
Mass	Stage-1	17.92	42.50	65.77	29.71
	Stage-1 + *LOCAL*	18.82	44.76	67.32	30.42
	Stage-1 + *FRAG*	29.55	58.36	81.44	43.01
	Full-Eval	**30.50**	**60.14**	**82.60**	**44.30**
Formula	Stage-1	14.67	35.47	59.83	25.18
	Stage-1 + *LOCAL*	15.47	36.11	60.57	25.98
	Stage-1 + *FRAG*	23.12	51.07	76.16	36.32
	Full-Eval	**23.70**	**51.77**	**76.92**	**36.99**

Molecular Pretraining and Encoder Comparison. We evaluate molecular representations using Stage-1 alone on MassSpecGym-mass. Table 14 compares the proposed pretrained encoder with the same architecture without molecular pretraining and with ChemBERTa-77M-MLM and MoLFormer.

Table 14: Molecular encoder comparison and pretraining ablation for Stage-1 on MassSpecGym-mass. All metrics are reported in percent. The best result in each column is bold.

Encoder	Hit@1 (%)	Hit@5 (%)	Hit@20 (%)	MRR (%)
ChemBERTa-77M-MLM	12.22	31.52	56.54	22.35
MoLFormer	16.42	40.01	**65.99**	28.20
Ours w/o pretraining	10.70	31.99	55.50	21.10
Ours	**17.92**	**42.50**	65.77	**29.71**

Removing molecular pretraining decreases Hit@1 from 17.92% to 10.70% and MRR from 29.71% to 21.10%. Among the compared encoders, the proposed representation achieves the highest Hit@1, Hit@5, and MRR, while MoLFormer obtains a slightly higher Hit@20. These results support the use of molecular pretraining for the initial global retrieval stage.

B.4 Case Studies

Rank recovery across benchmarks. Figure 5 presents three selected examples of successful rank recovery on MassSpecGym, SpectraVerse, and NPLIB1. The ground-truth molecules move from Stage-1 ranks 209, 152, and 14, respectively, to first place after MS-CHER's Stage-2 refinement. In particular, the first two cases demonstrate recovery of correct structures lying far outside the initial top-20 shortlist. Because refinement operates on the existing candidate set, these recoveries highlight its ability to exploit structural evidence that is not fully captured by the global compatibility score.

The displayed molecular structures connect these rank changes to the specific candidates retrieved before and after refinement. Across the three benchmarks, these examples illustrate how Stage-2 can correct substantial initial ranking errors and turn low-ranked ground-truth candidates into successful top-1 identifications.

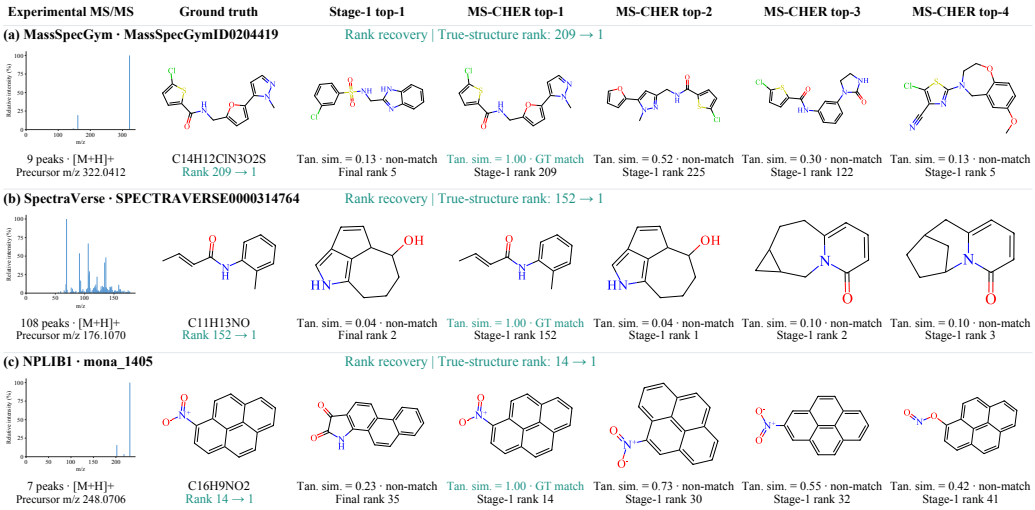

Figure 5: Selected retrieval examples from MassSpecGym, SpectraVerse, and NPLIB1, showing the experimental spectrum, ground truth, Stage-1 top-1 candidate, and MS-CHER top-4 candidates. Tanimoto similarity to the ground truth is computed using 4096-bit ECFP4 fingerprints.

MassSpecGym · mass · Test · Evidence within selected K

	GLOBAL	+ LOCAL	+ FRAG	+ BOTH	Adaptive
MassSpecGymID0051219 Scope K = 122 / 256 GLOBAL	1	6	4	7	1
MassSpecGymID0251345 Scope K = 149 / 256 GLOBAL + LOCAL	6	1	15	4	1
MassSpecGymID0204554 Scope K = 203 / 256 GLOBAL + FRAG	202	211	1	1	1
MassSpecGymID0244372 Scope K = 164 / 256 GLOBAL + LOCAL + FRAG	85	17	4	1	1

Figure 6: **Ground-truth ranks under different evidence configurations.** Alternative configurations are evaluated within each query's selected refinement prefix in the MassSpecGym mass setting. Adaptive denotes the configuration selected by MS-CHER. Lower ranks are better.

Query-dependent evidence selection. Figures 6 and 7 present four selected queries from the MassSpecGym mass setting, covering global retrieval without additional evidence, *LOCAL*-only refinement, *FRAG*-only refinement, and joint refinement. For each query, alternative evidence

configurations are evaluated within the selected refinement prefix. The rank comparison shows how the evidence choices affect retrieval, while the molecular examples display the corresponding candidate structures and final rankings.

In case (a), MS-CHER preserves the correct global rank of 1 without activating either expert, whereas adding *LOCAL*, *FRAG*, or both yields ranks 6, 4, and 7. In case (b), the selected *LOCAL* refinement improves the rank from 6 to 1, while *FRAG* and joint refinement yield ranks 15 and 4. In case (c), *FRAG* alone improves the rank from 202 to 1, matching joint refinement without activating *LOCAL*. In case (d), combining both experts improves the rank from 85 to 1, compared with ranks 17 and 4 using *LOCAL* and *FRAG* individually. These cases highlight how MS-CHER matches structural computation to query-specific needs: preserving correct global results, selecting effective individual experts, and combining complementary evidence when beneficial, while avoiding unnecessary expert execution.

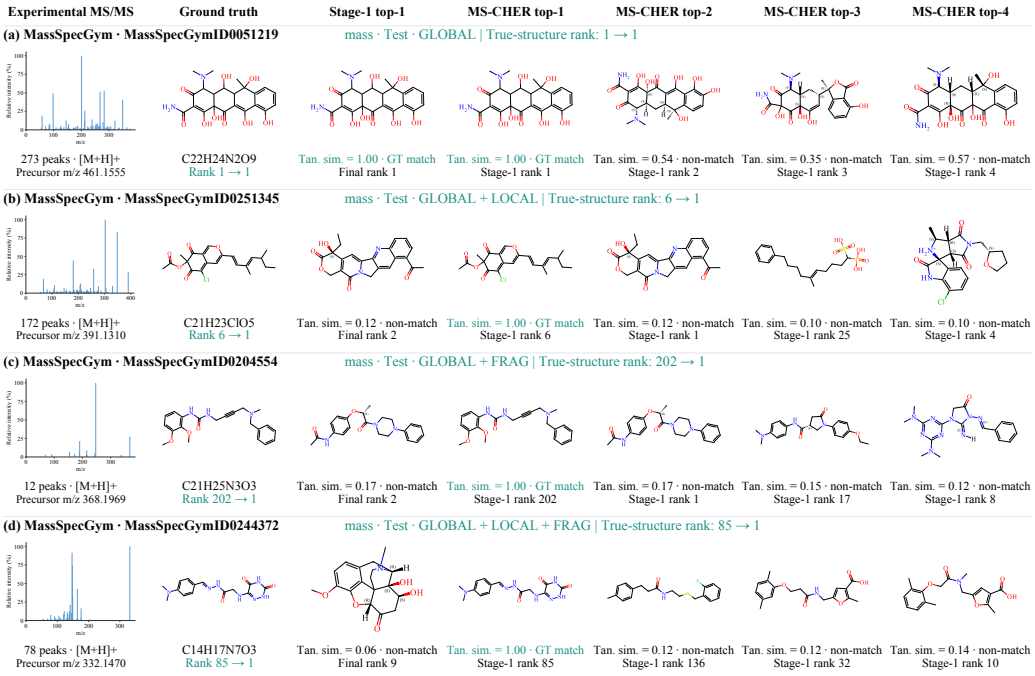

Figure 7: **Molecular retrieval examples with different adaptive evidence choices.** Cases (a)–(d) use global retrieval alone, *LOCAL* refinement, *FRAG* refinement, and joint refinement, respectively. Each row shows the experimental spectrum, ground-truth molecule, Stage-1 top-ranked candidate, and the top four candidates after MS-CHER refinement. Tanimoto similarity to the ground truth is computed using 4096-bit ECFP4 fingerprints.