TRANSFERABLE MASS SPECTRUM PREDICTION VIA REFERENCE-GUIDED TEST-TIME SPECIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Tandem mass spectrum prediction supports compound identification across metabolomics, natural-product discovery, and environmental analysis. However, pretrained predictors often degrade under shifts in chemical space and acquisition conditions, while retraining domain-specific models from scratch is costly. We introduce SPARC, a retrieval-guided test-time specialization framework that adapts a pretrained predictor using a spectral reference library without accessing test-query spectra. For each target query, SPARC retrieves chemically related reference spectra to recalibrate fragment intensities within the learned fragmentation space. During Transfer, SPARC combines reference-guided spectral adaptation with reliability-aware consistency, using reconstruction behavior on retrieved spectra to selectively preserve trustworthy predictions during continual specialization. Across MassSpecGym, NPLIB1 and application-specific GNPS libraries, SPARC improves spectral prediction under multiple transfer settings. These results establish retrieval-guided test-time specialization as a practical strategy for extending pretrained MS/MS predictors to specific chemical and acquisition domains, with continual test-time training providing further refinement during deployment.

1 INTRODUCTION

Tandem mass spectrometry is central to molecular annotation, natural-product discovery (Dührkop et al., 2019; Wang et al., 2016) and environmental analysis (Schymanski et al., 2014). By estimating fragment-ion masses and intensities from molecular structures, molecule-to-spectrum predictors extend spectral reference coverage beyond available experimental measurements (Goldman et al., 2024; Murphy et al., 2023; Nowatzky et al., 2025; Young et al., 2024a). Their performance, however, depends on the chemical and acquisition distributions represented during training (Bremer et al., 2022; Young et al., 2024a). Large spectral libraries expose a model to broader chemical space and fragmentation behavior than a single study can provide (Gupta et al., 2026; Murphy et al., 2023). Real-world deployment rarely queries this space uniformly. Each application emphasizes a particular molecular distribution together with its own adducts, ionization modes, instruments, collision energies, and metadata. These factors can change spectral similarity and prediction performance (Bremer et al., 2022; Hoang et al., 2024; Liu et al., 2025). Consequently, broad training coverage does not ensure accurate predictions for the chemical space and acquisition condition of a particular study.

This mismatch produces a practical generalization problem. Changes in acquisition conditions alter the mapping from molecular structure to fragment intensities (Bremer et al., 2022). Study-defined molecular distributions can also concentrate predictions in regions that were sparse during source training (Bushuiev et al., 2024). Adduct-dependent fragmentation provides a concrete example. Compounds retained under one precursor ion may lack paired spectra under another condition. The same molecule can also fragment differently across adducts (Schmid et al., 2021), creating coverage and chemical-composition gaps. When measured spectra are unavailable for the target molecules, chemically related reference spectra provide an alternative source of domain-specific information. Prior work on spectral representation learning and analogue search has shown that MS/MS neighborhoods can recover chemically related molecules (de Jonge et al., 2023; Huber et al., 2021a). This motivates test-time specialization that adapts the predictor to the queried domain.

Test-time training specializes a pretrained model during inference. In classification, TENT usesprediction entropy as a confidence signal, while CoTTA combines teacher averaging with stochastic

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.

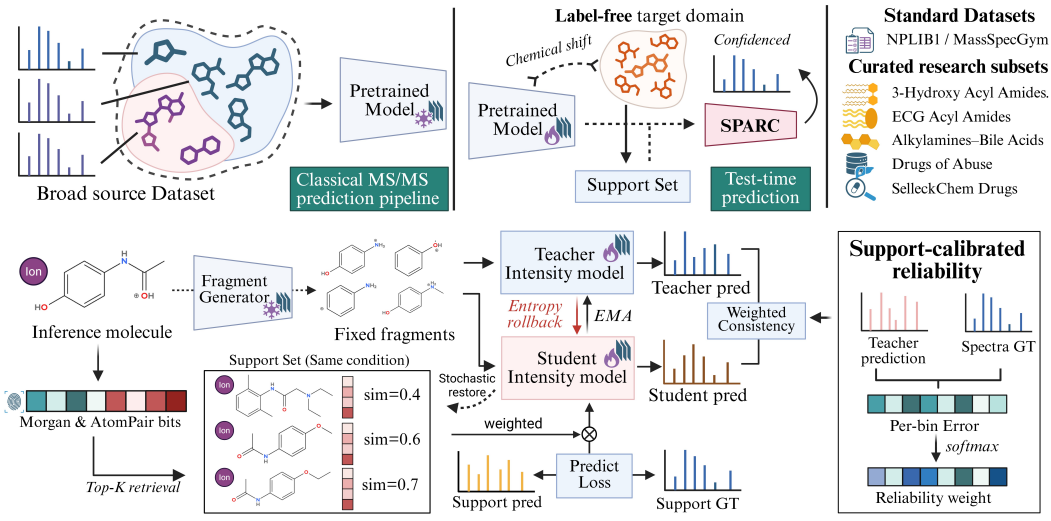


Figure 1: Overview of SPARC. **Top:** A predictor pretrained on a broad source collection is specialized to a target research domain. **Bottom:** Target molecules retrieve acquisition-compatible and chemically similar support spectra. The fragment generator remains frozen, while support reconstruction errors calibrate bin-wise reliability of an exponential moving average teacher. The student intensity model optimizes similarity-weighted support reconstruction and reliability-weighted teacher–student consistency, with stochastic restoration and rollback constraining adaptation drift.

restoration to limit error accumulation and forgetting (Wang et al., 2020; 2022). Transferring these ideas to MS/MS prediction, however, requires complementing confidence-based adaptation with an external spectral signal indicating how predictions change.

Spectral entropy summarizes the dispersion of normalized peak intensities (Li et al., 2021), but it does not encode whether peaks occur at the correct locations. Entropy can therefore decrease by suppressing weak peaks without improving the predicted fragmentation pattern. Moreover, target molecular structures identify the chemical region being queried but do not specify how the predicted spectrum should change. Chemically related measured reference spectra provide the missing external signal, while teacher predictions regularize the update and preserve reliable source-model structure.

We introduce SPARC (**S**pectral **P**rediction via **A**daptation with **R**etrieval-calibrated **C**onsistency), a retrieval-guided test-time training framework that combines support-guided specialization with drift-aware continual adaptation for pretrained MS/MS predictors (Fig. 1). For each target, SPARC retrieves chemically related labeled spectra using acquisition-aware prioritization and uses their measured intensities to supervise adaptation. In addition, SPARC uses reconstruction errors on retrieved support spectra to weight teacher–student consistency, relaxing teacher constraints in bins with larger support errors. An exponential moving average teacher, stochastic restoration and rollback mechanism constrain continual updates together.

We evaluate SPARC on MassSpecGym and NPLIB1 (Bushuiev et al., 2024; Dührkop et al., 2021), covering specialization to an adduct-defined subset and transfer between spectral libraries. We additionally curate five application-specific GNPS collections using a shared processing pipeline to represent focused scientific domains (Wang et al., 2016). These evaluations show robust improvements in the main benchmark transfer settings and higher entropy similarity across all application libraries. Peak-level analyses further characterize how adaptation improves predictions within the preserved candidate space, revealing increases in precision and F1-score at the peak and intensity level.

2 Related Work

MS/MS Spectrum Prediction. Computational mass spectrometry encompasses both spectrum interpretation and molecule-to-spectrum prediction. SIRIUS combines isotope patterns and fragmentation trees for molecular formula and structure annotation (Dührkop et al., 2019). Earlier deep learning approaches focused on direct spectrum regression, including NEIMS for fingerprint-based

EI prediction (Wei et al., 2019), 3DMolMS for 3D structure-based MS/MS prediction (Hong et al., 2023), and GrAFF-MS for graph-based high-resolution spectrum prediction (Murphy et al., 2023). More recently, explicit fragmentation models have further connected predicted peaks to molecular substructures. Iceberg combines autoregressive fragmentation-graph generation with fragment intensity prediction (Goldman et al., 2024), whereas FIORA estimates fragment-ion probabilities from single-bond cleavages and their local molecular neighborhoods (Nowatzky et al., 2025). MassFormer uses graph transformers to model molecular structure, whereas FraGNNet introduces a structured probabilistic model for high-resolution spectrum prediction (Young et al., 2024a;b).

Test-Time Training. Test-time training updates a trained model using information available during inference. TENT minimizes prediction entropy by updating normalization parameters (Wang et al., 2020), while CoTTA supports non-stationary streams through an exponential-moving-average teacher, augmentation pseudo-labels, and stochastic restoration (Wang et al., 2022). Beyond computer vision, TAIP adapts interatomic potentials to out-of-distribution molecular configurations using global- and local-structure self-supervision (Cui et al., 2025). Test-time learning has also been explored in mass spectrometry for peptide-spectrum prediction (Ye et al., 2024) and *de novo* small-molecule generation from observed spectra (Mismetti et al., 2026).

Mass Spectral Libraries and Benchmarks. Public repositories such as GNPS, MassBank, and MetaboLights aggregate spectra across laboratories, scientific domains, and acquisition conditions, providing broad coverage but substantial heterogeneity (Horai et al., 2010; Wang et al., 2016; Yurekten et al., 2024). To support systematic machine learning evaluation, several curated benchmarks have since been developed. NPLIB1 provides a natural-product-oriented collection, MassSpecGym standardizes prediction and retrieval benchmarks, and MS^{n}Lib and SpectraVerse broaden coverage across compounds, adducts, and ionization modes (Brungs et al., 2025; Bushuiev et al., 2024; Dührkop et al., 2021; Gupta et al., 2026). In addition, application-specific collections such as oxylipin libraries and GNPS community subsets reflect concrete scientific settings (Elloumi et al., 2024), from which we curate five application-grounded target domains for SPARC.

3 METHOD

We consider MS/MS spectrum prediction from a known molecular structure and its acquisition conditions. Let $\mathbf{x} = (\mathcal{G}, \mathbf{c})$, where $\mathcal{G}$ is the molecular graph and $\mathbf{c}$ contains the available metadata, including precursor adduct, instrument type, and collision energy. The output is a non-negative intensity vector $\mathbf{y} \in \mathbb{R}^B_{\geq 0}$ over B fixed m/z resolution bins.

A predictor F_ω estimates the intensity vector as $\hat{\mathbf{y}} = F_\omega(\mathbf{x})$. At deployment, the target molecules and acquisition conditions may differ from those represented during source training. Given a pretrained predictor F_{ω_0}, a target query x_t, and a labeled reference library $S = \{(x_i^s, y_i^s)\}_{i=1}^M$, our goal is to specialize the predictor using the query input and reference spectra without using the measured query spectrum in the adaptation objective.

We consider two deployment settings, both with access to a labeled reference library. In the **target-with-validation** setting, labeled target-domain validation spectra guide update acceptance and checkpoint selection. In the **target-only** setting, no target validation labels are used, and updates are accepted by an entropy-based guard. Measured test-query spectra are reserved for evaluation only.

3.1 FRAGMENT-SPACE PRESERVING ADAPTATION

SPARC instantiates F_ω with Iceberg, which separates discrete fragment generation from continuous intensity prediction (Goldman et al., 2024). Writing $\omega = (\psi, \theta)$ for the parameters of these two stages respectively, we have

$$\mathcal{H}_{\mathbf{x}} = G_\psi(\mathbf{x}), \qquad \hat{\mathbf{y}} = f_\theta(\mathbf{x}, \mathcal{H}_{\mathbf{x}}). \tag{1}$$

Here, G_ψ autoregressively constructs a directed acyclic graph fragmentation $\mathcal{H}_{\mathbf{x}}$ of candidate fragments, and f_θ predicts their contributions to the spectrum. The source checkpoint provides the initial parameters (ψ_0, θ_0). Iceberg-Generate supplies a ranked set of candidate fragments to Iceberg-Score, allowing us to reuse the pretrained generator as a fixed fragmentation prior. Accordingly, SPARC

fixes G_{ψ_0} and adapts only θ. Conditioned on molecular and fragment representations and acquisition metadata, the intensity model uses chemically related reference spectra to learn target spectral patterns, emphasizing characteristic fragments and suppressing less relevant candidates. Adaptation therefore changes intensity allocation within the existing candidate space.

The adaptation pipeline uses retrieved spectra in two complementary ways. Their measured intensities directly supervise the student, while the reconstruction errors of an EMA teacher determine the bin-wise weights for consistency on the queries. Student and teacher states persist across successive queries, with stochastic restoration and rollback governing continual updates.

3.2 Condition-Aware Chemical Retrieval

SPARC constructs the chemical-local support neighborhood for each query, so the supervision follows the chemical region being processed. For a molecule $\mathbf{x}$, let $\phi(\mathbf{x})$ be the ℓ_2-normalized concatenation of its Morgan and hashed AtomPair bit fingerprints (Carhart et al., 1985; Rogers & Hahn, 2010). Each reference spectrum is scored by its fingerprint cosine similarity to the target molecule.

$$r_{t,i} = \frac{\phi(x_t)^\top \phi(x_i^s)}{\|\phi(x_t)\|_2 \|\phi(x_i^s)\|_2} \tag{2}$$

Acquisition compatibility takes precedence over chemical ranking. With the default adduct-based retrieval, references matching the query adduct form the candidate pool. Within this pool, references also matching the query instrument are ranked first, followed by the remaining adduct-compatible references. Each group is sorted by $r_{t,i}$, and the first K entries form S_t. The implementation falls back to global retrieval only when the adduct-compatible pool is empty. Collision energy and precursor m/z remain model inputs but are not hard retrieval filters.

Retrieved spectra contribute according to their chemical relevance. Writing $K_t = |\mathcal{S}_t|$ and reindexing locally reindexing the entries of $\mathcal{S}_t$, we define

$$a_{t,i} = \frac{\exp(r_{t,i}/\tau)}{\sum_{j=1}^{K_t} \exp(r_{t,j}/\tau)}, \qquad \tau > 0, \tag{3}$$

and minimize the similarity-weighted support reconstruction loss

$$\mathcal{L}_{\text{sup}}(\theta; \mathcal{S}_t) = \frac{1}{B} \sum_{i=1}^{K_t} a_{t,i} \, \| f_\theta(\mathbf{x}_i^s) - \mathbf{y}_i^s \|_2^2 \, . \tag{4}$$

This term anchors adaptation in measured spectra rather than relying exclusively on the model's own predictions.

3.3 Support-Calibrated Teacher Consistency

The teacher supplies a target prediction, but its reliability need not be uniform across the spectrum (Tarvainen & Valpola, 2017). We initialize the student θ and teacher $\bar{\theta}$ from θ_0. Before each update, we evaluate the teacher on the retrieved labeled spectra and estimate an error profile across m/z bins:

$$e_{t,b} = \frac{1}{K_t} \sum_{i=1}^{K_t} [f_{\bar{\theta}}(\mathbf{x}_i^s)_b - y_{i,b}^s]^2, \qquad c_{t,b} = \exp(-\gamma e_{t,b}), \gamma \geq 0. \tag{5}$$

Support errors are averaged uniformly over the retrieved set, whereas retrieval-similarity weights modulate the supervised reconstruction loss. Chemical locality enters the error estimate through support selection. The resulting $c_{t,b}$ is computed without gradients and controls the strength of teacher consistency on the shared m/z grid: larger support reconstruction errors yield weaker constraints. We use these weights as a support-derived regularization signal, rather than as calibrated probabilities of correctness for individual query fragments.

We regularize the student toward the unperturbed teacher prediction, reducing pressure in bins that the teacher reconstructs poorly:

$$\mathcal{L}_{\text{con}} = \frac{1}{B} \sum_{b=1}^{B} c_{t,b} \, [f_\theta(x_t)_b - \text{sg}(f_{\bar{\theta}}(x_t)_b)]^2 \tag{6}$$

where sg denotes stop-gradient. The core adaptation objective is

$$\mathcal{L}_{\text{adapt}} = \mathcal{L}_{\text{con}} + \lambda_{\text{sup}}\mathcal{L}_{\text{sup}} \tag{7}$$

where λ_{sup} controls the weight of direct support supervision. The measured support spectra drive student updates, while teacher consistency limits changes on the current query according to reliability. These weights determine how student model is constrained by the teacher at each bin.

3.4 Continual Updates and Drift Control

Mechanisms for stabilizing continual updates. SPARC uses two CoTTA-inspired mechanisms to stabilize continual updates, an EMA teacher and stochastic restoration (Wang et al., 2022). For each query and its retrieved support set, the student takes gradient steps on Eq. (7), and the accepted state is carried to the next query. At inner step u, the optimizer first produces a trial student state θ_u^+; the teacher then tracks this state according to

$$\bar{\theta}_{u+1} = \alpha\bar{\theta}_u + (1-\alpha)\theta_u^+. \tag{8}$$

The EMA decay coefficient α is fixed in our implementation. After the teacher update, each trainable student weight or bias entry is independently restored to its source value with probability ρ:

$$m_{u,j} \sim \text{Bernoulli}(\rho), \qquad \theta_{u+1,j} = m_{u,j}\theta_{0,j} + (1-m_{u,j})\theta_{u,j}^+. \tag{9}$$

Thus, the teacher averages the optimized student before restoration, whereas restoration anchors the student to the pretrained parameters. This ordering smooths the teacher trajectory while limiting cumulative drift in the model that is updated online.

Supervised and unsupervised rollback. Besides stabilizer, SPARC controls the remaining drift with a rollback rule whose evidence depends on whether labeled validation spectra are available. Before each trial update, we snapshot the whole model and optimizer states.

When validation spectra are available, a trial update is accepted only if its post-update validation CosSim remains at least η times the pre-update value. This comparison uses validation spectra for the rollback decision, while the adaptation loss remains unchanged.

When validation spectra are unavailable, the same rollback operation uses a label-free entropy guard instead. For a predicted intensity vector $\mathbf{z}$, define normalized spectral entropy as

$$\begin{aligned} p_b(\mathbf{z}) &= \frac{\max(z_b, 0)}{\sum_{j=1}^{B}\max(z_j,0)+\epsilon}, \\ h(\mathbf{z}) &= -\frac{1}{\log B}\sum_{b=1}^{B} p_b(\mathbf{z})\log[p_b(\mathbf{z})+\epsilon]. \end{aligned} \tag{10}$$

Let $\bar{y}_u^-$ and $\bar{y}_u^+$ denote the unperturbed teacher predictions for the current query before and after the trial update, respectively. We define

$$H_u^- = h(\bar{y}_u^-), \qquad H_u^+ = h(\bar{y}_u^+). \tag{11}$$

We reject the update when $H_u^+ - H_u^- > \delta$. This label-free guard rejects updates that abruptly broaden the predicted intensity distribution, providing a lightweight stability criterion.

4 Experiments

Datasets and benchmarks Following prior work, we first evaluate SPARC on MassSpecGym (MSG) (Bushuiev et al., 2024) and NPLIB1 (Dührkop et al., 2021). We consider adduct-focused specialization from the full MSG dataset to MSG-MNa subset and cross-library transfer from MSG to NPLIB1. MSG→MSG-MNa is an adduct-defined subset specialization, as the full-MSG source checkpoint includes M+Na examples. We further use five experimental libraries on GNPS: Drugs of Abuse, 3-Hydroxy Acyl Amides, ECG Acyl Amides C4–C24, Alkylamines–Bile Acids, and SelleckChem FDA. These collections span forensic toxicology, acyl-amide and lipid chemistry, bile-acid conjugates, and approved pharmaceuticals. For all five GNPS target libraries, the MSG training split serves as the labeled reference support library, while target spectra are reserved only for evaluation. To prevent information leakage, Murcko scaffolds were computed for molecules, and support molecules sharing a scaffold with any validation or test molecule were excluded before retrieval (Bemis & Murcko, 1996). Complete preprocessing setting are provided in Supplementary A.

Evaluation We evaluate spectral fidelity using cosine similarity (CosSim), Jensen–Shannon entropy similarity (EntSim), and mean squared error (MSE), with entropy as an auxiliary diagnostic. Before scoring, we retain the 100 highest-intensity predicted bins. Comparisons use the successful intersection within methods, and coverage is reported in Supplementary A.5. We additionally evaluate peak-presence and ion-current quality using precision, recall, and F1 after exact-bin matching, with peak- and intensity-level metrics computed from matched prediction and ground-truth.

Implementation details and baselines SPARC retrieves 64 support spectra for each query. During adaptation, SPARC optimizes its objective with Adam using a learning rate of 1×10^{-4}. Each molecular representation concatenates 2,048 radius-2 Morgan fingerprint bits with 3,072 hashed atom-pair bits. We set the support-loss coefficient to $\lambda_{\text{sup}} = 1$ and use $\tau = 0.1$, $\gamma = 10$, $\alpha = 0.99$, $\rho = 0.02$, and $\delta = 0.1$ for the remaining adaptation and regularization terms. Finally, SPARC use student model for prediction.. We implement all models on NVIDIA A800 (80GB) GPUs.

We compare against NEIMS, GrAFF-MS, FIORA and pretrained Iceberg models as predictor baselines, and TENT and CoTTA on Iceberg as source-free adaptation baselines. Architectural details are reported in Supplementary B.

Adaptation protocols. For the MSG and NPLIB1 benchmarks, labeled target-training spectra provide support supervision. SPARC(val) specializes the source checkpoint using target-validation spectra for rollback and checkpoint selection, and is then evaluated without test-time updates. SPARC–TTT starts from the same selected checkpoint and continues adapting on test-query structures and acquisition metadata using entropy-based rollback. The GNPS experiments instead omit validation specialization and use MSG-training spectra as the reference library, with entropy-based rollback throughout online adaptation. Measured test-query spectra are never used in adaptation losses, rollback decisions, or checkpoint selection.

5 RESULTS

5.1 PREDICTION HETEROGENEITY AND SPECTRAL DOMAIN SHIFTS.

Cross-adduct shift changes both chemistry and fragmentation. In MassSpecGym, only 19.0% of molecules with M+H spectra also have M+Na spectra. These molecules are enriched for N-free, polyol/carbohydrate-like, and O-rich structures, but depleted for basic amines, aromatic nitrogen, and sulfone/sulfonamides, consistent with known adduct preferences (Kruve et al., 2013). Even among matched molecules, over 50% share no peaks and few neutral losses, reflecting strong adduct-dependent fragmentation (de Jonge et al., 2026; Liu et al., 2025). M+Na spectra also show lower entropy. Detailed analysis are provided in Supplementary C.

Locality motivates retrieval. We next ask whether the prediction error is uniform. We visualize this by projecting DreaMS embeddings and molecule fingerprints of dataset samples into a two-dimensional UMAP and coloring each sample by its prediction performance (Bushuiev et al., 2026). Regions of high and low accuracy occur within every split and in both molecule and spectra views (Fig. 2). This qualitative pattern is supported in Supplementary C.3. Because errors cluster locally, fine-tuning on the whole support set can improve performance across the entire local neighborhood, turning test-time training into genuine domain specialization (Huber et al., 2021b).

5.2 SPARC SPECIALIZES ACROSS AND WITHIN DOMAINS

We first evaluate SPARC under explicit domain shifts. On MSG→MSG-MNa, SPARC raises EntSim and CosSim from 0.180/0.280 for the pretrained MSG checkpoint to 0.274/0.370 after validation-stage specialization, with SPARC–TTT further reaching a CosSim of 0.376 (Table 1). The adapted model also exceeds the Iceberg model trained directly on M+Na, which obtains 0.262/0.348. SPARC therefore recovers a substantial part of the adduct-specific prediction gap through reference-guided specialization of the pretrained model.

Starting from the MSG checkpoint, specialization using NPLIB1 training spectra as support and validation-guided rollback and checkpoint selection improves EntSim and CosSim from 0.470/0.517 to 0.559/0.630 (Table 1). Across both transfer settings, validation-guided specialization provides the

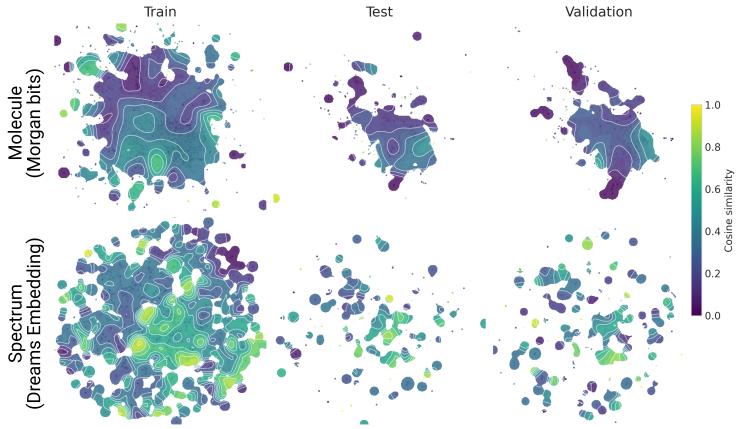


Figure 2: **Prediction performance is locally structured in spectral and molecule spaces.**

Table 1: Test performance for MSG→MSG-MNa and MSG→NPLIB1 transfer.

Method	Entropy	MSE	EntSim	CosSim	Method	Entropy	MSE	EntSim	CosSim
MSG→MSG-MNa (MSG initialized)					*MSG→NPLIB1 (MSG initialized)*				
NEIMS	4.150	14.461	0.072	0.072	NEIMS	3.568	20.246	0.392	0.363
GrAFF-MS	4.134	15.822	0.075	0.071	GrAFF-MS	3.396	21.170	0.330	0.280
FIORA	3.015	1.020	0.063	0.071	FIORA	1.900	1.317	0.369	0.412
Iceberg–MSG	3.461	1.242	0.180	0.280	Iceberg-MSG	3.586	1.269	0.470	0.517
Iceberg–MNa	2.485	1.006	0.262	0.348	Iceberg-NPLIB1	3.516	1.217	0.521	0.586
SPARC(val)	2.715	0.895	0.274	0.370	SPARC(val)	3.112	1.063	0.559	0.630
SPARC-TTT	2.923	**0.882**	**0.284**	**0.376**	SPARC–TTT	3.128	**1.041**	**0.575**	**0.650**
Iceberg(tent)	2.077	1.183	0.229	0.290	Iceberg(tent)	3.052	1.252	0.490	0.517
Iceberg(cotta)	3.419	1.189	0.191	0.288	Iceberg(cotta)	3.560	1.245	0.472	0.518

main performance gain, with test-stream continuation yielding additional refinement as target queries arrive. Thus, the benefit is not confined to the adduct subset but persists when transferring to a distinct library. A separate controlled comparison matches source initialization, target-training support access, validation protocol, and held-out test evaluation to isolate the benefit beyond retrieval-only fine-tuning (Supplementary D). On NPLIB1, similarity-weighted retrieved-support fine-tuning achieves 0.5873 CosSim, compared with 0.5545 for global fine-tuning, supporting the value of chemically conditioned training beyond access to labeled target-domain spectra alone.

Table 2: Test performance for MSG M+H and NPLIB1.

Method	Entropy	MSE	EntSim	CosSim	Method	Entropy	MSE	EntSim	CosSim
MSG M+H (MSG initialized)					*NPLIB1 (NPLIB1 initialized)*				
NEIMS	3.982	16.180	0.184	0.162	NEIMS	3.349	20.652	0.324	0.280
GrAFF-MS	3.756	16.713	0.169	0.137	GrAFF-MS	3.521	19.689	0.384	0.362
FIORA	1.742	0.917	0.434	0.460	FIORA	1.596	1.237	0.419	0.484
Iceberg–MSG	3.477	**0.853**	0.462	0.529	Iceberg-NPLIB1	3.516	1.217	0.521	0.586
SPARC(val)	2.866	0.907	0.494	0.532	SPARC(val)	2.925	1.107	0.559	0.616
SPARC–TTT	2.794	0.880	**0.505**	**0.547**	SPARC–TTT	2.944	**1.092**	**0.565**	**0.628**

SPARC can also refine predictors that are already matched. We apply the same adaptation procedure within MSG M+H and within NPLIB1, where the backbone was trained on the target domain itself (Table 2). On the dominant MSG M+H subset, SPARC increases EntSim and CosSim from 0.462/0.529 to 0.505/0.547, and increases EntSim and CosSim from 0.521/0.586 to 0.565/0.628 on NPLIB1. These within-domain gains show that retrieved reference spectra provide a useful local refinement signal even after the predictor has already learned the target-domain distribution, making SPARC a test-time specialization stage rather than only a mechanism for correcting domain shifts.

5.3 SPARC Improves Spectral Purity by Suppressing Spurious Peaks

We next examine how SPARC redistributes intensity within the frozen fragmentation space. For peak-quality evaluation on M+H and M+Na, we follow the Iceberg evaluation setting: retain the 100 highest-intensity predicted bins, remove peaks below 1% of the predicted base-peak intensity, and match the remaining bins to the experimental spectrum. Both SPARC and SPARC–TTT improve intensity-weighted precision and F1 (Table 3), concentrating more predicted intensity on experimentally supported peaks. We further rank the saved M+Na top-100 predictions by intensity and assess agreement with experimental peak support and intensities using average precision (AP), NDCG, and Spearman correlation on matched bins. Improvements across these measures indicate better prioritization of supported peaks and more faithful relative intensity ordering.

Besides, we compare Iceberg and the validation-selected SPARC checkpoint over the complete candidate space, progressively adding peaks in descending predicted-intensity order until matched experimental peaks account for 25%, 50%, 75%, or 100% of the experimental ion current recoverable within the frozen candidate space. We compare purity at each coverage level and find that purity increases at full recoverable coverage, showing that less unsupported predicted intensity accompanies comparable experimental signal recovery (Supplementary. D.2).

Table 3: Peak quality, candidate ranking, and spectral purity of Iceberg and SPARC.

Method	MSG M+H Peak			MSG M+H Inten			MSG M+Na Peak			MSG M+Na Inten		
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Iceberg	0.203	**0.550**	0.297	0.465	**0.756**	0.576	0.039	**0.159**	0.062	0.153	**0.314**	0.206
SPARC	0.264	0.449	0.333	0.581	0.691	0.632	**0.061**	0.097	**0.075**	**0.361**	0.297	0.326
SPARC–TTT	**0.271**	0.438	**0.335**	**0.593**	0.687	**0.637**	0.045	0.112	0.074	0.340	0.312	**0.327**

Method	MSG M+Na Ranking				MSG M+Na Purity				MSG M+H Purity			
	AP	N@10	N@50	Spear	P@25	P@50	P@75	P@100	P@25	P@50	P@75	P@100
Iceberg	0.402	0.508	0.567	0.268	0.459	0.444	0.414	0.383	0.804	0.765	0.722	0.664
SPARC	0.474	0.612	0.648	0.409	**0.561**	**0.573**	**0.569**	**0.567**	**0.809**	**0.771**	**0.740**	**0.699**
SPARC–TTT	**0.485**	**0.624**	**0.664**	**0.410**	–	–	–	–	–	–	–	–

Fig. 3 illustrates this effect: the source prediction assigns substantial intensity to low-m/z peaks, whereas the adapted predictions suppress many of these peaks and concentrate relative intensity on the dominant peak. Together, these results show that useful target-domain corrections can be made within the existing candidate space by changing which fragments receive substantial intensity.

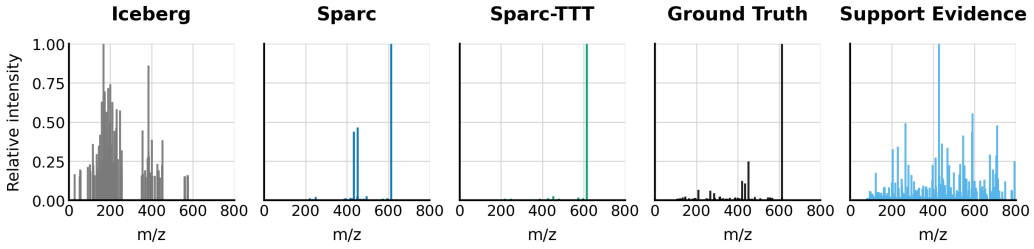


Figure 3: **Representative example of prediction from SPARC.** Each displayed spectrum is normalized, and peaks with intensities below 1% are removed.

5.4 SPARC Transfers to Application-Specific Target Domains

We next evaluate target-only specialization on five application-specific GNPS libraries using MSG-training spectra as references, without a validation-specialization stage. Retrieval follows the acquisition-aware policy in Sec. 3.2, and online updates use entropy-based rollback.

Measured target-library spectra are reserved for evaluation. SPARC achieves the best EntSim among the compared predictors on every library and the best scores across all three reported metrics on GNPS-A/B and SC-FDA. These results demonstrate specialization to application-defined query streams using an external reference library.

Table 4: Performance on Drugs of Abuse, 3HAA, ECG, GNPS-A/B, and SC-FDA. All models are initially trained on the MSG dataset.

Method	DoA			3HAA			ECG			GNPS-A/B			SC-FDA		
	MSE	EntSim	Cos	MSE	EntSim	Cos	MSE	EntSim	Cos	MSE	EntSim	Cos	MSE	EntSim	Cos
NEIMS	19.90	0.198	0.127	12.84	0.386	0.382	34.33	0.184	0.145	18.05	0.409	0.242	14.75	0.148	0.110
GrAFF-MS	19.98	0.149	0.084	15.23	0.238	0.173	34.49	0.145	0.117	19.84	0.324	0.179	15.57	0.157	0.107
FIORA	1.078	0.427	0.458	0.935	0.395	0.397	2.836	0.140	0.117	1.009	0.393	0.488	0.893	0.362	0.405
Iceberg	**1.057**	0.415	0.483	0.904	0.490	0.509	**2.621**	0.178	0.223	0.960	0.504	0.511	0.925	0.322	0.379
SPARC	1.058	**0.482**	**0.509**	**0.901**	**0.508**	**0.514**	2.654	**0.205**	0.246	**0.919**	**0.534**	**0.569**	**0.839**	**0.389**	**0.435**
w/o rollback	1.068	0.451	0.488	0.940	0.493	0.497	2.631	0.202	**0.252**	1.109	0.516	0.536	0.920	0.367	0.391

5.5 Component Ablation

Table 5 compares component ablations under adaptation on the MSG → MSG-MNa transfer. All tested ablations reduce CosSim relative to the full method. Removing similarity-based support retrieval produces the largest decrease, whereas removing stochastic restoration has the smallest effect. These comparisons identify similarity-based support use as the most influential component among the tested ablations.

Table 5: **Ablation Study of SPARC Components.**

Method	MSE	EntSim	CosSim
-similarity-based support retrieval	0.9859	0.2595	0.3441
-support-calibrated bin confidence	0.8889	0.2637	0.3514
-stochastic restoration	0.8840	0.2720	0.3692
-per-query rollback	0.8977	0.2656	0.3631
SPARC	**0.8804**	**0.2742**	**0.3755**

These results support the roles of chemically relevant references, direct support supervision, and stochastic restoration in continual adaptation. Removing similarity-based support produces the clearest degradation, indicating that adaptation depends largely on local chemical relevance. Removing stochastic restoration leads to the smallest decrease in CosSim, suggesting that this mechanism is not uniformly necessary and that allowing the adapted model to move farther from the source parameters may be beneficial for mass spectrum prediction.

6 Limitations

SPARC has two limitations. First, it is bounded by the fragmentation space inherited from the pretrained predictor. Freezing the fragment generator stabilizes adaptation but restricts SPARC to reweighting candidate fragments and it can't recover an omitted diagnostic fragment. Jointly expanding the candidate space without destabilizing online adaptation can be an important direction for domains with unseen fragmentation pathways.

Second, SPARC assumes that chemically similar, acquisition-compatible support spectra provide transferable supervision. Poor support coverage or a fingerprint-caused mismatch can bias both the supervised update and the support-derived reliability estimate. Continual updates further introduce dependence on target-stream order, while entropy-based rollback limits abrupt drift without guaranteeing correct peak locations. Query-specific uncertainty estimates and order-robust safeguards aligned more directly with spectral fidelity could mitigate these limitations.

7 Conclusion

We introduced SPARC, a retrieval-guided test-time specialization framework for pretrained MS/MS predictors. SPARC preserves the fragmentation space and adapts fragment intensities using chemically related reference spectra, with support reconstruction errors weighting teacher consistency. Validation-assisted specialization improves benchmark spectral similarity, with further gains from online continuation. Without validation specialization, entropy-guarded adaptation using MSG references improves EntSim across all GNPS application libraries. Peak-level analyses associate these gains with higher spectral purity and intensity redistribution within the existing candidate space. Together, these results support reference-guided specialization as an extension of pretrained predictors when compatible labeled reference spectra are available.

REFERENCES

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

Parker Ladd Bremer, Arpana Vaniya, Tobias Kind, Shunyang Wang, and Oliver Fiehn. How well can we predict mass spectra from structures? benchmarking competitive fragmentation modeling for metabolite identification on untrained tandem mass spectra. *Journal of Chemical Information and Modeling*, 62(17):4049–4056, 2022. doi: 10.1021/acs.jcim.2c00936.

Corinna Brungs, Robin Schmid, Steffen Heuckeroth, et al. MS^nLib: Efficient generation of open multi-stage fragmentation mass spectral libraries. *Nature Methods*, 22:2028–2031, 2025. doi: 10.1038/s41592-025-02813-0.

Roman Bushuiev, Anton Bushuiev, Niek F de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, et al. Massspecgym: A benchmark for the discovery and identification of molecules. *Advances in Neural Information Processing Systems*, 37:110010–110027, 2024.

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.

Raymond E. Carhart, Dennis H. Smith, and R. Venkataraghavan. Atom pairs as molecular features in structure-activity studies: Definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73, 1985. doi: 10.1021/ci00046a002.

Taoyong Cui, Chenyu Tang, Dongzhan Zhou, Yuqiang Li, Xingao Gong, Wanli Ouyang, Mao Su, and Shufei Zhang. Online test-time adaptation for better generalization of interatomic potentials to out-of-distribution data. *Nature Communications*, 16:1891, 2025. doi: 10.1038/s41467-025-57101-4.

Niek F. de Jonge, Joris J. R. Louwen, Elena Chekmeneva, Stephane Camuzeaux, Femke J. Vermeir, Robert S. Jansen, Florian Huber, and Justin J. J. van der Hooft. MS2Query: Reliable and scalable MS2 mass spectra-based analogue search. *Nature Communications*, 14:1752, 2023. doi: 10.1038/s41467-023-37446-4.

Niek F. de Jonge, Elena Chekmeneva, Robin Schmid, David Joas, Lem-Joe Truong, Justin J. J. van der Hooft, and Florian Huber. Cross ionization mode chemical similarity prediction between tandem mass spectra in metabolomics. *Nature Communications*, 17:2483, 2026.

Victoria Deleray, Vincent Charron-Lamoureux, Kyle Vittali, Helena Mannochio-Russo, Crystal X Wang, Corinn Walker, Karsten Zengler, Dionicio Siegel, and Pieter C Dorrestein. Synthesis-driven reverse metabolomics reveals 3-hydroxy n-acyl amides as gut microbial molecules. *bioRxiv*, 2026. doi: 10.64898/2026.01.07.698252.

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

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

Kai Dührkop, Louis-Félix Nothias, Markus Fleischauer, et al. 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.

Anis Elloumi, Lindsay Mas-Normand, Jamie Bride, et al. From MS/MS library implementation to molecular networks: Exploring oxylipin diversity with NEO-MSMS. *Scientific Data*, 11:193, 2024. doi: 10.1038/s41597-024-03034-4.

Emily C Gentry, Stephanie L Collins, Morgan Panitchpakdi, et al. Reverse metabolomics for the discovery of chemical structures from humans. *Nature*, 626(7998):419–426, 2024. doi: 10.1038/s41586-023-06906-8.

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.

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.

Corey Hoang, Winnie Uritboonthai, Linh Hoang, Elizabeth M. Billings, Aries Aisporna, Farshad A. Nia, Rico J. E. Derks, James R. Williamson, Martin Giera, and Gary Siuzdak. Tandem mass spectrometry across platforms. *Analytical Chemistry*, 96(14):5478–5488, 2024. doi: 10.1021/acs.analchem.3c05576.

Yuhui Hong, Sujun Li, Christopher J Welch, Shane Tichy, Yuzhen Ye, and Haixu Tang. 3dmolms: prediction of tandem mass spectra from 3d molecular conformations. *Bioinformatics*, 39(6): btad354, 2023. doi: 10.1093/bioinformatics/btad354.

Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, et al. MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714, 2010. doi: 10.1002/jms.1777.

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

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

Anneli Kruve, Karl Kaupmees, Jaanus Liigand, Merit Oss, and Ivo Leito. Sodium adduct formation efficiency in ESI source. *Journal of Mass Spectrometry*, 48(6):695–702, 2013. doi: 10.1002/jms.3218.

Yuanyue Li, Tobias Kind, Jacob Folz, Arpana Vaniya, Sajjan Singh Mehta, and Oliver Fiehn. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nature Methods*, 18(12):1524–1531, 2021. doi: 10.1038/s41592-021-01331-z.

Botao Liu, Zhifeng Tang, and Tao Huan. Adduct-induced variability in tandem mass spectrometry. *Analytical Chemistry*, 97(31):17058–17066, 2025. doi: 10.1021/acs.analchem.5c02792.

Laura Mismetti, Marvin Alberts, Andreas Krause, and Mara Graziani. Test-time tuned language models enable end-to-end de novo molecular structure generation from MS/MS spectra. *arXiv preprint arXiv:2510.23746*, 2026. URL `https://arxiv.org/abs/2510.23746`.

Michael Murphy, Stefanie Jegelka, Ernest Fraenkel, Tobias Kind, David Healey, and Thomas Butler. Efficiently predicting high resolution mass spectra with graph neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25549–25562. PMLR, 2023.

Yannek Nowatzky, Francesco Friedrich Russo, Jan Lisec, Alexander Kister, Knut Reinert, Thilo Muth, and Philipp Benner. Fiora: Local neighborhood-based prediction of compound mass spectra from single fragmentation events. *Nature Communications*, 16:2298, 2025. doi: 10.1038/s41467-025-57422-4.

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

Robin Schmid, Daniel Petras, Louis-Felix Nothias, Mingxun Wang, Allegra T. Aron, et al. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nature Communications*, 12:3832, 2021. doi: 10.1038/s41467-021-23953-9.

Emma L. Schymanski, Junho Jeon, Rebekka Gulde, Kathrin Fenner, Matthias Ruff, Heinz P. Singer, and Juliane Hollender. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science & Technology*, 48(4):2097–2098, 2014. doi: 10.1021/es5002105.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

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.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7191–7201. IEEE, 2022.

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

Jianbai Ye, Xiangnan He, Shujuan Wang, Meng-Qiu Dong, Feng Wu, Shan Lu, and Fuli Feng. Test-time training for deep MS/MS spectrum prediction improves peptide identification. *Journal of Proteome Research*, 23(2):550–559, 2024. doi: 10.1021/acs.jproteome.3c00229.

Adamo Young, Hannes Röst, and Bo Wang. Tandem mass spectrum prediction for small molecules using graph transformers. *Nature Machine Intelligence*, 6(4):404–416, 2024a. 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. *arXiv preprint arXiv:2404.02360*, 2024b.

Ozgur Yurekten, Thomas Payne, Noemi Tejera, Felix Xavier Amaladoss, Callum Martin, Mark Williams, and Claire O'Donovan. Metabolights: open data repository for metabolomics. *Nucleic acids research*, 52(D1):D640–D646, 2024.

AI USE STATEMENT

We used generative AI tools to assist with parts of method implementation, experimental workflow development, and manuscript writing and editing, and to obtain feedback on research methodology and experimental design. AI tools were used as supporting tools rather than autonomous decision makers: all proposed methods, experimental designs, code, analyses, and scientific interpretations were reviewed by the authors. AI-assisted code was inspected and tested against the intended implementation, and AI-assisted scientific claims and manuscript text were checked against the corresponding experimental results and relevant literature.

Generative AI was not used to make final decisions regarding methodology, evaluation protocols, interpretation of results, or scientific conclusions. We take responsibility for the final content of this work, including all text, claims, analyses, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops and evaluates computational methods for small-molecule tandem mass-spectrum prediction using existing spectral benchmarks and publicly available experimental spectral libraries. The study does not involve collecting new human-subject data or personally identifiable information, or conducting interventions involving human participants. All datasets are used for scientific benchmarking and method evaluation in accordance with their intended research purposes. The proposed method is intended to support molecular annotation and mass-spectrometry analysis; as with other computational prediction tools, its outputs should be experimentally validated before being used in safety-critical, clinical, or regulatory settings.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details required to reproduce the reported results throughout the paper and appendices. Section 3 specifies the complete SPARC adaptation procedure, including condition-aware retrieval, support-weighted supervision, reliability-weighted teacher consistency, continual updates, stochastic restoration, and rollback. Section 4 describes the datasets, evaluation metrics, adaptation protocols, hyperparameters, and baselines.

Supplementary A documents dataset construction, support-set eligibility, spectral preprocessing, evaluation-cohort construction, and prediction coverage, while Appendix B provides the backbone architecture, online state-management procedure, optimization details, rollback rules, baseline configurations, and computational budget. Appendices C and D describe the additional domain-shift analyses, locality experiments, retrieval controls, order-sensitivity analysis, matched-concentration controls, and support counterfactual experiments. Together, these sections specify the data processing, model configuration, evaluation procedure, and statistical analyses used for the reported experiments.

An anonymized implementation will be made available to reviewers during the review process.

A Datasets and Evaluation

A.1 Benchmark Datasets

MassSpecGym (MSG) is a large curated MS/MS benchmark assembled from GNPS, MoNA, MassBank, and in-house measurements. It harmonizes molecular structures and acquisition metadata, retaining positive-mode electrospray spectra with $[\mathrm{M}+\mathrm{H}]^+$ or $[\mathrm{M}+\mathrm{Na}]^+$ precursors acquired on Orbitrap- or QTOF-type instruments. MSG also provides standardized, generalization-demanding splits for molecular discovery and spectrum-simulation tasks, making it suitable for evaluating adduct-aware spectrum prediction across diverse chemical and acquisition conditions (Bushuiev et al., 2024). MSG→MSG-MNa specializes a full-MSG model that has already seen M+Na examples, rather than transferring from an exclusively M+H-trained model.

NPLIB1 is a public natural-product MS/MS collection originally prepared for CANOPUS; it was subsequently renamed to distinguish the library from the CANOPUS classifier. Its standard processed release contains 10,709 spectra from 8,553 unique molecular structures and uses structure-disjoint training, validation, and test partitions. NPLIB1 therefore complements the broader MSG benchmark with a smaller, chemically focused library of natural-product-like molecules (Dührkop et al., 2021; Goldman et al., 2024).

A.2 Downstream Spectral Libraries

We evaluate the models on five experimental spectral libraries obtained from GNPS. These libraries were selected to cover complementary downstream domains, including forensic toxicology, pharmaceutical compounds, microbiome-associated lipids and synthetically generated metabolite families. Collectively, they introduce locality in molecular structure, chemical class, instrument platform and precursor-adduct distribution relative to the large curated training dataset. All evaluated spectra are experimentally acquired rather than generated in silico (Wang et al., 2016).

Drugs of Abuse. This library is a community reference provided by Prof. Heather Barkholtz. It contains spectra of known drugs of abuse and related psychoactive compounds. It was developed primarily for forensic and clinical toxicology applications. The library mainly consists of positive-mode protonated precursors acquired on Q-TOF instruments. It therefore provides a focused drug-oriented domain with both forensic relevance and an instrument distribution distinct from general-purpose metabolomics collections.

3-Hydroxy Acyl Amides. This library is a synthesis-derived reference developed by Prof. Victoria Deleray and colleagues in the Dorrestein laboratory using a synthesis metabolomics strategy. It contains experimentally measured spectra of 3-hydroxy N-acyl amides constructed from fatty-acyl chains of varying lengths and diverse amine headgroups such as amino acids, polyamines and neuroactive amines. These molecules represent a recently characterized class of microbiome-associated bioactive lipids. The spectra were acquired primarily in positive mode on Orbitrap instruments and include protonated and dehydrated precursor ions. It therefore provides a structured lipid-oriented domain with a distinct chemical space and fragmentation behavior compared with general metabolomics collections (Deleray et al., 2026).

ECG Acyl Amides C4–C24. This library is a community reference introduced by Prof. Emily C. Gentry and colleagues for synthesis-based reverse metabolomics. The library was generated through combinatorial coupling of fatty-acyl chains with diverse amine headgroups, producing a systematically structured family of N-acyl amides spanning C4–C24 chains. Spectra were acquired predominantly from protonated precursors in positive mode using a Bruker MaXis HD Q-TOF instrument. It therefore provides a controlled chemical-series domain where closely related molecules differ in fatty-chain length or amine identity, enabling evaluation of generalization within a densely sampled metabolite family (Gentry et al., 2024).

GNPS Alkylamines–Bile Acids. The library is a community collection from the Dorrestein laboratory. Its principal constituents are bile-acid amidates formed by conjugating bile-acid cores with structurally diverse amines and amino-acid-derived headgroups. These compounds combine a conserved steroidal scaffold with substantial variation in the conjugated headgroup. The spectra were

acquired primarily by positive-mode LC–ESI Orbitrap MS/MS and contain protonated precursors together with alternative positive-ion adducts such as $[\mathrm{M} + \mathrm{Na}]^+$. It therefore provides a lipid-like metabolite domain with complex fragmentation patterns influenced by both conserved scaffolds and variable conjugated groups (Gentry et al., 2024).

GNPS SelleckChem FDA. This library is a community reference generated by Sirenas MD from SelleckChem approved-drug collection. It contains experimentally measured spectra of approved drugs and active pharmaceutical ingredients spanning diverse therapeutic and chemical classes. The library therefore provides a broad pharmaceutical domain with heterogeneous drug-like scaffolds, complementing the chemically focused acyl-amide and bile-acid libraries and evaluating model generalization across diverse pharmaceutical structures (Dührkop et al., 2015; Wang et al., 2016).

A.3 Support Set Eligibility

For the MSG→MSG-MNa and MSG→NPLIB1 transfer experiments, the M+Na subset of the MSG training split and the NPLIB1 training split serve as the respective support libraries. For the MSG-MH and NPLIB1 self-transfer experiments, the M+H subset of the MSG training split and the NPLIB1 training split serve as the respective support libraries. Spectra from the validation split determine update acceptance and checkpoint selection, yielding SPARC(val), and SPARC–TTT continues from that selected checkpoint on the test stream.

The GNPS application experiments omit validation specialization and only use the label-free rollback rule. In all settings, a test query supplies only its molecular structure and acquisition metadata, while its measured spectrum is reserved for evaluation. The method therefore uses existing labeled references without accessing test-query spectra.

A.4 Spectral Representation and Evaluation Details

Preprocessing and filtering pipeline Molecular inputs use the processed structures supplied by the corresponding benchmark or library. The Iceberg backbone represents spectra using 15,000 bins spanning 0–1,500 m/z. Support reconstruction, query prediction, and evaluation use the same binning grid. In GNPS application dataset, we filter candidates who occur in support set.

Evaluation-cohort construction. In evaluation setting, we restrict each comparison to spectra with valid predictions from every included method. We use cosine similarity, entropy similarity, and MSE as evaluation metrics. Reported MSE values are multiplied by 10^4. For the separate peak-quality analysis, we additionally remove predicted peaks below 1% of the predicted base-peak intensity. Peak-presence precision, recall, and F1 use exact-bin matching, whereas ion-current metrics account for the intensities of matched and unmatched bins. Predicted spectral entropy is an auxiliary measure of intensity concentration, not a fidelity metric with a preferred direction.

A.5 Prediction Coverage

Table 6 separates individual-method coverage from the size of such an intersection: the MSG-MNa input pool contains 3,490 spectra, while coverage varies across predictors. For each of the five application libraries, all methods are evaluated on the same cohort (Table 7); these counts should not be interpreted as the sizes of the original source collections.

Table 6: Prediction coverage on the MSG-MNa and MSG M+H test cohorts. The eligible query pools contain 3490 and 9954 spectra, respectively; counts below are measured before taking the intersection across compared methods.

MSG-MNa (n=3490)		*MSG M+H (n=9954)*	
Predictor	Evaluable spectra	Predictor	Evaluable spectra
NEIMS / GrAFF-MS / FIORA	3490	NEIMS / GrAFF-MS / FIORA	9954
Iceberg–M+Na/MSG	3331	Iceberg–MSG	9954
SPARC(val) / SPARC–TTT	3331	SPARC(val) / SPARC–TTT	9954

Table 7: Evaluated spectrum counts for the application-specific libraries.

Library	Evaluated spectra
Drugs of Abuse	242
3-Hydroxy Acyl Amides	528
ECG Acyl Amides C4–C24	169
GNPS Alkylamines–Bile Acids	2,103
SelleckChem FDA	127

B Implementation Details

B.1 Backbone Architecture

SPARC uses the pretrained Iceberg fragment generator and intensity model. The generator constructs a directed acyclic graph of candidate fragments for each molecule and remains frozen throughout adaptation. The resulting candidate graphs are reused, while only the intensity model is optimized. Table 8 summarizes the backbone configuration.

Table 8: Iceberg backbone architecture used by SPARC. The fragment generator is fixed, and the intensity model is adapted.

Component	Configuration
Generator model	6-layer GGNN; hidden size 512, dropout 0.2, positional-encoding size 14
Intensity model	4-layer GGNN; hidden size 256; dropout 0.1, 3 fragment-set 1 MLP layer
Candidate space	At most 100 candidate nodes per molecule
Conditioning	Adduct, collision-energy, instrument, and element-group embeddings
Additional inputs	Formula encoding, explicit hydrogens, and unshifted m/z

B.2 Online Execution and State Management

The retrieval hierarchy, adaptation objective, reliability weights, EMA update, stochastic restoration, and entropy guard are defined in Sections 3.2–3.4. Here we specify the execution details needed to reproduce the reported runs. A fresh support set is retrieved for every query, and each support spectrum is evaluated with its own acquisition metadata. Queries follow their order in the input file without shuffling; additional passes, when used, repeat the same order.

For each query, the implementation first records the pre-update prediction, executes U inner Adam trials, and then records the prediction from the final accepted state. Before every trial, it snapshots the student, EMA teacher, and optimizer together; the rollback decision therefore either retains or restores all three states. An accepted post-query state becomes the initialization for the next query. The validation-specialization stage follows the same stateful procedure and ranks checkpoints only after a complete pass through the validation stream. The reported SPARC(val) model is the selected student evaluated without test-time updates; SPARC–TTT is initialized from the same checkpoint and continues the procedure on the ordered test stream. In neither case is a measured test spectrum available to the adaptation loss or rollback rule.

B.3 Optimization, Rollback, and Spectrum Construction

All reported runs use 4 inner updates per query, $K = 64$ retrieved support spectra, and a constant learning rate of 10^{-4}. Table 9 lists the rollback rule used in each phase, and the rollback rule is the only task-specific protocol choice listed in the table.

Adam uses the PyTorch defaults $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ with zero weight decay. We use neither a learning-rate scheduler nor gradient clipping. The fragment generator runs in inference mode and is absent from the differentiated computation graph. Consequently, it receives no gradient or parameter update even though the optimizer is constructed from the joint model parameter list; all weights and biases of the intensity predictor remain trainable.

For phases using the supervised guard, query-level CosSim is computed after retaining the 100 highest-intensity predicted bins, without an additional intensity cutoff. A trial is rejected if its post-update CosSim is less than $\eta = 0.8$ times the pre-update value. At pass boundaries, validation

Table 9: Rollback rules for validation specialization and online continuation. For the GNPS application libraries, no validation-specialization phase is used.

Target setting	Validation specialization	Test TTT continuation
MSG→MSG-MNa	CosSim, $\eta = 0.8$	Entropy, $\delta = 0.1$
MSG→NPLIB1	CosSim, $\eta = 0.8$	Entropy, $\delta = 0.1$
NPLIB1	CosSim, $\eta = 0.8$	Entropy, $\delta = 0.1$
MSG-MH	CosSim, $\eta = 0.8$	Entropy, $\delta = 0.1$
GNPS application libraries	—	Entropy, $\delta = 0.1$

checkpoints are ordered lexicographically by higher CosSim, lower predicted entropy, and lower MSE. Validation labels are used only for rollback and checkpoint selection, not as terms in the adaptation objective. Online continuation instead uses the entropy rule in Eq. (10) with $\delta = 0.1$.

The adaptation layer does not renormalize measured support intensities before Eqs. (5)–(7). Peaks are placed directly on the 15,000-bin grid, and when multiple peaks map to the same bin, the largest supplied intensity is retained. This dense vector is used for both support reconstruction and the teacher's bin-wise error estimate. Metric-specific normalization is applied only when computing CosSim, EntSim, or spectral entropy. Hyperparameters were selected on the corresponding target validation split before final testing; test labels were not used for tuning. The five GNPS target-only runs share one configuration and use no supervised rollback.

B.4 Baselines and Optimization

We compare against NEIMS, GrAFF-MS, FIORA, and pretrained Iceberg models. Each baseline are evaluated using their provided implementations and checkpoints, with standardized data loading and evaluation. Predictions from all models are mapped to the common grid before retaining the 100 highest-intensity bins for scoring. Source-trained, target-trained, and target-adduct-trained Iceberg checkpoints are evaluated separately, without support retrieval or adaptation.

We train NEIMS and GrAFF-MS on the MSG and NPLIB1 datasets for 200 epochs with early stopping, using a hidden size of 128, a dropout rate of 0.1, cosine loss, the Adam optimizer, learning rate 10^{-4}, and a batch size of 8. GrAFF-MS additionally uses adduct and collision-energy conditioning. We train FIORA on the MSG or NPLIB1 dataset for 300 epochs using a batch size of 32, a learning rate 2×10^{-4}, weight decay of 10^{-5}, and graphwise KL loss. For Iceberg, we use the architecture described above with the official weights and retrain the model on the MSG-MNa subset.

TENT and CoTTA adapt the Iceberg intensity model without support spectra or target labels. Both use Adam with learning rate 10^{-4} and zero weight decay; CoTTA uses an EMA coefficient of 0.99 and a restoration probability of 0.1.

For SPARC, we use Adam optimizer with a learning rate of 10^{-4}, a support-set size of 64, $\lambda_{\text{sup}} = 1$, $\tau = 0.1$, $\gamma = 10$, $\alpha = 0.99$, and $\rho = 0.02$. The task-specific rollback rules are listed in Table 9. All models are implemented in PyTorch and evaluated on NVIDIA A800 (80GB) GPUs.

B.5 Computational Budget

In the final protocol, for each query, $K = 64$ labeled support spectra are retrieved and four Adam updates are applied to the intensity model. Table 10 reports wall-clock measurements from SPARC runs on one NVIDIA A800 GPU. Peak GPU memory was approximately 12,000 MB (about 12 GB). Across the representative cohorts in Table 10, adaptation required 2.74–7.47 s per query.

Table 10: Measured computational budget for representative target cohorts. Per-query times are normalized by the number of valid queries and measured on one NVIDIA A800 GPU.

Cohort	Queries	Support/query	Adapt. time/query (s)
MSG-MNa	3331	64	2.74
MSG-MH	9954	64	7.47
NPLIB1	902	64	5.93
GNPS-3HAA	528	64	4.87

C Additional Analyses of Adduct and Spectral Domain Shifts

C.1 Molecular Coverage and Chemical Composition

We examine all M+H and M+Na records across the MSG training, validation, and test splits, identifying molecules by InChIKey connectivity and representing each molecule by its most frequent SMILES–formula pair. The collection contains 27,987 M+H-observed and 6,272 M+Na-observed molecules, of which 5,330 occur in both groups. Thus, only 19.0% of M+H-observed molecules also have an M+Na record. These library-wide counts establish that the two observed-adduct domains differ in molecular coverage.

Table 11: Functional-motif associations with concurrent M+Na observation among M+H-observed molecules. Percentages give motif prevalence in each group. Crude odds ratios (ORs) are unstratified; CMH ORs are calculated with stratification by dataset split, 50-Da neutral exact-mass bin, and M+H instrument set. Brackets denote 95% confidence intervals.

Feature	M+H/M+Na (%)	M+H only (%)	Crude OR [95% CI]	CMH OR [95% CI]
Sulfone/sulfonamide	1.7	6.0	0.26 [0.21, 0.33]	0.26 [0.21, 0.32]
N-free	41.2	18.8	3.01 [2.83, 3.21]	3.11 [2.88, 3.36]
Polyol/carbohydrate-like	36.1	12.1	4.09 [3.82, 4.38]	3.06 [2.81, 3.34]
O-rich ($\geq$ 4 O atoms)	73.1	43.1	3.59 [3.36, 3.83]	2.93 [2.71, 3.17]
Basic amine proxy	19.7	38.4	0.39 [0.37, 0.42]	0.38 [0.35, 0.41]
Phosphate/phosphonate	1.6	2.9	0.55 [0.44, 0.70]	0.39 [0.31, 0.49]
Aromatic nitrogen	26.2	44.8	0.44 [0.41, 0.47]	0.49 [0.45, 0.53]
Alcohol/phenol	51.1	28.8	2.59 [2.44, 2.75]	2.04 [1.90, 2.19]
Ester/lactone	29.8	16.0	2.23 [2.08, 2.38]	1.83 [1.69, 1.98]
Ether	68.5	51.1	2.08 [1.95, 2.22]	1.76 [1.63, 1.89]
Aldehyde/ketone	18.8	11.1	1.86 [1.71, 2.01]	1.53 [1.40, 1.67]
Amide/urea	38.3	45.5	0.74 [0.70, 0.79]	0.69 [0.64, 0.73]
Multiple amides ($\geq$ 2)	13.0	14.3	0.89 [0.82, 0.98]	0.75 [0.68, 0.83]
Long aliphatic chain (C8+)	10.2	10.2	1.00 [0.91, 1.10]	0.80 [0.71, 0.89]
Fatty-acyl/lipid-like	6.1	6.5	0.94 [0.83, 1.06]	0.83 [0.72, 0.95]
Carboxylic acid/carboxylate	20.0	18.8	1.08 [1.00, 1.16]	1.03 [0.95, 1.12]

The overlap is also chemically non-uniform. Within the 27,987 M+H-observed molecules, we compare 16 structure-derived functional motifs between molecules with and without an M+Na record. Cochran–Mantel–Haenszel (CMH) estimates adjust jointly for dataset split, 50-Da neutral exact-mass bin, and the instrument classes associated with each molecule's spectra. The group with concurrent M+Na observations is enriched for nitrogen-free, polyol/carbohydrate-like, and oxygen-rich molecules, with odds ratios of 3.11, 3.06, and 2.93, respectively (Table 11, Figure 4). Basic-amine, aromatic-nitrogen, and sulfone/sulfonamide motifs are depleted, with odds ratios of 0.38, 0.49, and 0.26. The confidence interval for carboxylic acid/carboxylate includes one. These associations describe the recorded library rather than intrinsic adduct-formation probabilities, but they establish a relevant prediction challenge: specialization to M+Na also changes the chemical composition of the queried domain.

C.2 Same-Molecule Paired Spectral Comparisons

To distinguish differences in molecular coverage from differences in the observed spectra themselves, we pair M+H and M+Na measurements of the same molecule on the same instrument. Direct-m/z cosine, peak Jaccard, and shared-ion F1 assess overlap in observed peak locations; neutral-loss cosine compares precursor-relative mass losses. This paired design holds molecular identity and instrument fixed without assuming that all remaining acquisition conditions are matched.

Direct overlap is low (Figure 4). The medians of direct-m/z cosine, peak Jaccard, and shared-ion F1 are all zero (Table 12). Neutral-loss cosine has a broader distribution but remains low at the median (0.041). Together with the coverage analysis, these results show that the M+Na prediction challenge is not simply a matter of adding spectra for previously unobserved molecules. Even when molecular identity and instrument match, the observed spectral pattern can change substantially across adducts.

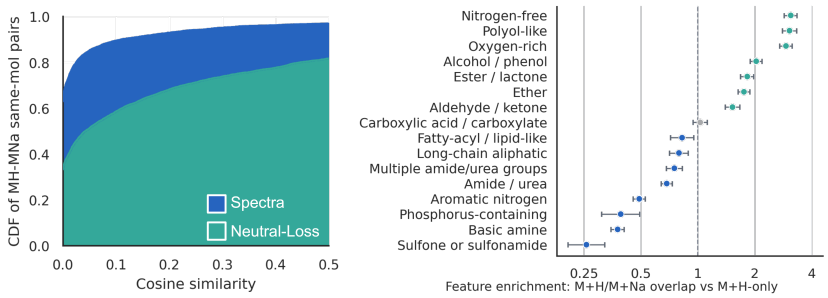


Figure 4: **Cross-adduct shifts affect both molecular composition and fragmentation patterns.** **Left:** Empirical CDFs of direct-spectrum and neutral-loss cosine similarity for same-molecule, same-instrument M+H/M+Na pairs, showing limited spectral overlap across adducts. **Right:** Functional-motif enrichment among M+H-observed molecules with versus without an M+Na record. Points show Cochran–Mantel–Haenszel odds ratios with 95% confidence intervals, adjusted for dataset split, neutral-mass bin, and instrument class; the dotted line denotes an odds ratio of one.

Table 12: Same-molecule, same-instrument M+H/M+Na spectral comparisons. Values are medians [interquartile ranges], accompanied by bootstrap 95% confidence intervals for the median. Sample sizes are reported separately for each metric.

	Metric	n	Median [IQR]	Bootstrap 95% CI
Same-Instrument Cross-Ion pairs	Direct m/z cosine	5,133	0.000 [0.000, 0.012]	[0.000, 0.000]
Same-Instrument Cross-Ion pairs	Neutral-loss cosine	5,000	0.041 [0.000, 0.326]	[0.035, 0.050]
Same-Instrument Cross-Ion pairs	Direct peak Jaccard	5,133	0.000 [0.000, 0.054]	[0.000, 0.000]
Same-Instrument Cross-Ion pairs	Shared-ion F1	5,133	0.000 [0.000, 0.039]	[0.000, 0.000]

Observed spectral entropy. We remove non-finite entries and non-positive intensities from measured peak lists, then merge peaks with identical m/z values by summing their intensities. After normalizing each spectrum to unit total intensity, we calculate Shannon entropy $H = -\sum_i p_i \log p_i$ using natural logarithms. All valid observed peaks are retained, and entropy is not divided by $\log B$. The histograms in Figure 5 use shared bins of width 0.1 nats and within-group normalization.

Figure 5 shows systematic entropy shifts across adduct and instrument groups, together with substantial within-group variation and overlap. Because these pooled comparisons are not matched for molecular identity, collision energy, or chemical composition, they characterize differences between the sampled domains rather than isolated acquisition effects.

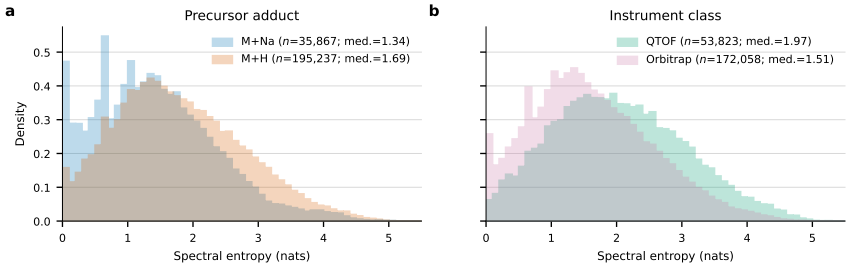


Figure 5: **Observed spectral entropy across adduct and instrument domains.** Histograms are normalized within each group using shared bins.

C.3 Neighborhood Performance Prediction Supports Locality

The regional patterns in Figure 2 motivate a quantitative question: can performance on nearby training samples predict whether an unseen query will be easy or difficult? To avoid relying on a two-dimensional projection, we define neighborhoods in the original DreaMS embedding and Morgan fingerprint spaces. Each query's predicted performance is the similarity-weighted mean CosSim of its K nearest training neighbors. Validation selects K; held-out test performance is then evaluated with that choice fixed. This analysis estimates prediction quality without updating the spectrum predictor.

Neighbor performance is informative in both representations (Figure 6). In DreaMS space, test-spectrum performance has Spearman $\rho = 0.667$ with the neighborhood estimate (95% CI $[0.620, 0.707]$; permutation $p = 9.99 \times 10^{-4}$), Pearson $r = 0.656$, and mean absolute error (MAE) 0.186, compared with 0.322 for the training-global-mean predictor. In Morgan fingerprint space, the molecule-level association is weaker but remains positive: Spearman $\rho = 0.355$ (95% CI $[0.276, 0.436]$; permutation $p = 9.99 \times 10^{-4}$), Pearson $r = 0.397$, and MAE 0.245, compared with 0.273 for the global mean.

This locality should not be confused with a monotonic relationship between training-set proximity and accuracy. The similarity to the nearest training reference alone correlates negatively with test performance in both DreaMS space ($\rho = -0.344$, 95% CI $[-0.430, -0.249]$) and Morgan space ($\rho = -0.172$, 95% CI $[-0.246, -0.093]$). What transfers across neighbors is information about local model performance, not a guarantee that a more similar query is easier.

DreaMS provides a retrospective spectral-space diagnostic, whereas molecular fingerprints are available from the query structure without its measured spectrum. The molecular result therefore supplies the more direct motivation for chemical retrieval. Neither association alone establishes an adaptation benefit; the controlled reference interventions in Section D test whether using this local information improves predictions.

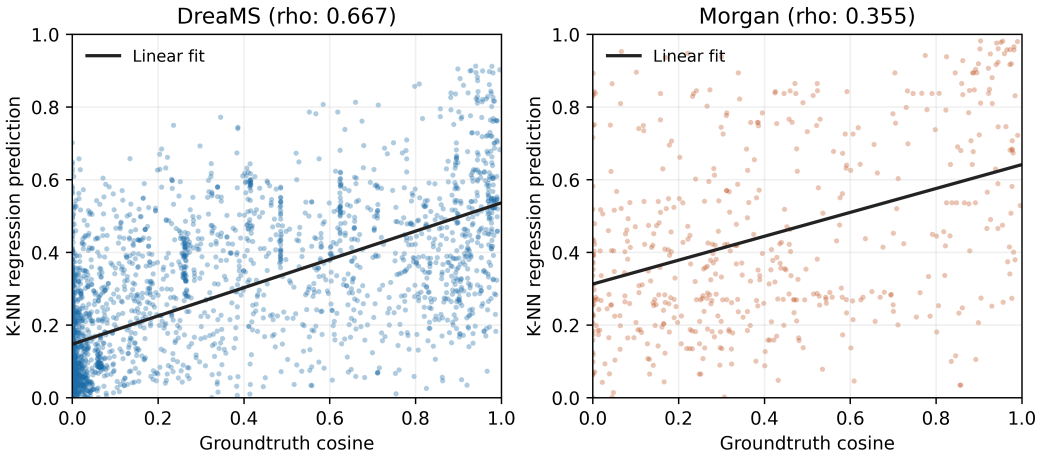


Figure 6: **Training-neighborhood performance predicts held-out query performance.** Test cosine scores are compared with similarity-weighted K-NN estimates in the original DreaMS space and Morgan fingerprint space. Neighborhood size is selected on validation data. Lines are ordinary least-squares fits; titles report Spearman correlations and molecule-cluster bootstrap 95% confidence intervals.

D Additional Analyses of Adaptation

D.1 Contribution Beyond Retrieval-Only Fine-Tuning

Chemically retrieved reference spectra already provide a strong supervision signal for target-domain specialization. We therefore ask how much of SPARC's gain can be attributed to this signal alone, and what is contributed by the remaining adaptation mechanisms. We compare SPARC(val) with simplified fine-tuning variants using the same source initialization, target-training support pool, validation protocol, and held-out test set. Retrieved-FT retains query-conditioned retrieval and similarity-weighted support supervision, while removing support-calibrated teacher consistency, the EMA teacher, stochastic restoration, and rollback. Global-FT instead fine-tunes on the same target-training support pool without query-conditioned retrieval.

As shown in Table 13, chemically localized supervision accounts for a substantial fraction of the specialization gain: Retrieved-FT consistently outperforms Global-FT, indicating that adaptation benefits from selecting and weighting references according to chemical relevance rather than from target-domain supervision alone. SPARC(val) further improves over retrieval-only fine-tuning, isolating the additional contribution of support-calibrated consistency and adaptation stabilization.

The M+Na controls further separate neighborhood selection from similarity weighting. Uniformly weighting either random or Top-K supports gives similar performance, whereas similarity-weighted Retrieved-FT performs substantially better. Thus, the retrieval benefit arises not only from restricting adaptation to a local chemical neighborhood, but also from emphasizing the most relevant references within that neighborhood.

Table 13: Comparison of SPARC with simplified target-domain specialization baselines.

Method	Support/update policy	CosSim	EntSim	MSE
	MSG M+Na			
Iceberg	No adaptation	0.2802	0.1798	1.242
Global-FT	Global random support; fixed model	0.2915	0.1932	1.053
Random-uniform	Random support; uniform weights	0.3063	0.2126	0.946
TopK-uniform	Top-K support; uniform weights	0.3170	0.2142	0.932
Retrieved-FT	Weighted Top-K; supervision only	<u>0.3268</u>	<u>0.2503</u>	<u>0.917</u>
SPARC(val)	Full SPARC	**0.3704**	**0.2739**	**0.895**
	NPLIB1			
Iceberg	No adaptation	0.5173	0.4698	1.269
Global-FT	Global random support; fixed model	0.5545	0.5103	1.277
Retrieved-FT	Weighted Top-K; supervision only	<u>0.5873</u>	<u>0.5346</u>	<u>1.108</u>
SPARC(val)	Full SPARC	**0.6298**	**0.5593**	**1.063**

D.1.1 Sensitivity to Query Order and Stochasticity

Because SPARC maintains model states across successive queries, its performance can depend on both stream order and stochastic restoration. We therefore repeat the complete specialization and online-adaptation procedure with five independent random seeds for MSG→MSG-MNa and MSG→NPLIB1. Each run independently performs validation-stage specialization, yielding a seed-specific validation-selected checkpoint. SPARC(val) denotes the corresponding checkpoint evaluated on the held-out test set without further updates, whereas SPARC–TTT continues adaptation from that checkpoint on an independently shuffled test stream. The random seed also controls stochastic restoration during adaptation. Table 14 reports the mean and standard deviation across the five runs.

Across both transfer settings, the validation-selected models remain stable across independent runs, while online continuation introduces additional variation from test-stream ordering and stochastic updates. The resulting performance remains close across runs, indicating that the observed gains are not driven by a particular query ordering.

Table 14: Sensitivity of online adaptation to query order and stochasticity. Results are mean ± standard deviation over 5 independent runs.

Method	EntSim	CosSim	Method	EntSim	CosSim
	MSG→MSG-MNa			*MSG→NPLIB1*	
SPARC(val)	0.273 ± 0.015	0.370 ± 0.019	SPARC(val)	0.561 ± 0.014	0.634 ± 0.014
SPARC–TTT	0.287 ± 0.013	0.379 ± 0.016	SPARC–TTT	0.573 ± 0.017	0.656 ± 0.017
Δ	0.014 ± 0.007	0.009 ± 0.005	Δ	0.012 ± 0.007	0.021 ± 0.010

D.2 Per-Spectrum Matched-Concentration Control

To test whether the M+Na improvement can be reproduced by simply making the source predictions sharper or sparser, we construct two post-processing controls from the Iceberg predictions. For each spectrum independently, we apply a positive power transformation $p'_i \propto p_i^{\beta}$ with $\beta \geq 1$. The entropy-matched control selects β so that the transformed Iceberg prediction matches the entropy of the corresponding frozen SPARC prediction. The count-matched control instead matches the number of peaks retained under the same 1% base-peak threshold used for peak-quality evaluation. In both controls, β is determined solely from the Iceberg and frozen SPARC predictions, before comparison with the experimental spectrum.

Table 15: **Per-spectrum matched-concentration controls on MSG M+Na.** Results are reported on the **entropy-feasible cohort subset**. Inten P and Inten F1 denote Intensity precision and F1 after the 1% base-peak filter.

Method	CosSim	EntSim	Inten Pre.	Inten. F1
Iceberg	0.2485	0.1692	0.1642	0.2177
Iceberg (adjusted entropy)	0.2539	0.2067	0.2535	0.2593
Iceberg (adjusted peaks)	0.2531	0.2137	0.2840	0.2604
SPARC	**0.3322**	**0.2878**	**0.3648**	**0.3364**

Both transformations make Iceberg substantially more concentrated and improve its peak-level intensity metrics (Table 15). Nevertheless, entropy-matched Iceberg reaches only 0.2539 CosSim and 0.2067 EntSim, compared with 0.3322 and 0.2878 for frozen SPARC. Count matching produces a similar result and also remains below frozen SPARC in ion-current precision and F1. Thus, the M+Na improvement reflects a learned redistribution of intensity across candidate fragments and cannot be explained by global sharpening or by simply retaining fewer peaks.

D.2.1 Query-conditioned chemical retrieval

SPARC retrieves a chemically local support set independently for each query. For a query x_t, each eligible reference spectrum is ranked by its fingerprint similarity,

$$r_{t,i} = \mathrm{sim}(\phi(x_t), \phi(x_i^s)) \,, \tag{12}$$

and the top-K references form the query-specific support set S_t. Acquisition compatibility is enforced before chemical ranking, prioritizing references that match the query adduct and instrument condition. The resulting support set and similarity weights are therefore conditioned on the individual query and provide local supervision for its adaptation step.

We additionally compare this query-conditioned retrieval against acquisition-matched random support and a spectrum-label-shuffled control under the same online adaptation protocol. Query-conditioned retrieval performs better than both controls, indicating that the correspondence between chemically related support molecules and their measured spectra provides useful adaptation information beyond acquisition matching alone.

D.2.2 Interpretation of Peak-Level Changes

The peak-presence and ion-current comparisons in the main text show that adaptation increases precision while reducing recall. Because the fragment generator is frozen, these changes arise from reallocating intensity among existing candidate fragments, not from generating new fragmentation pathways. Suppressing low-intensity candidates can remove spurious peaks while also discarding weak true peaks, accounting for the observed precision–recall trade-off.

This distinction also defines the method's scope. SPARC can improve the relative intensities and ranking of fragments already represented by the source model, but a diagnostic fragment absent from the candidate graph remains unavailable to the intensity model.

D.3 Support counterfactual Ablation

We additionally ran the same intervention on the complete M+Na test stream. The serial replay used the same archived validation-selected setting. The correct arm used condition-aware fingerprint Top-K retrieval and similarity weighting. The random arm instead drew a deterministic uniform sample from the same adduct- and instrument-matched source pool and used uniform support weighting. The shuffled arm retained the retrieved molecules and their similarity weights, but circularly permuted the associated measured support spectra within each query-specific support set.

Correct retrieval achieved the highest CosSim and EntSim and the lowest MSE on the full test stream (Table 16). Acquisition-matched random support reduced spectral fidelity, while shuffling the measured spectra among retrieved molecules caused a substantially larger degradation. Interestingly, random support yielded slightly lower predicted entropy than correct retrieval, reinforcing that entropy concentration alone is not a proxy for spectral fidelity.

Table 16: Serial full-test M+Na support counterfactual. All three arms used the same 3,490-row test stream; metrics are computed on all valid evaluated spectra. Higher CosSim and EntSim and lower loss and entropy are better.

Support condition	CosSim	EntSim	MSE	Entropy
Correct retrieval	**0.3755**	**0.2742**	**0.8804**	2.9230
Random retrieval	0.2994	0.2052	0.9242	2.9468
Spectrum-label shuffled	0.1854	0.0987	0.9457	4.3728