PEAKCLOUD: BENCHMARKING AND LEARNING ON PEAK-LEVEL MASS SPECTRA FOR METABOLOMIC PATTERN RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Mass spectra acquired from non-invasive biospecimens such as urine, breath, and feces enable untargeted metabolomics to capture comprehensive molecular profiles for disease characterization. While machine learning can extract informative patterns from these spectra, current studies are largely confined to single cohorts and suffer from disparate data processing pipelines and evaluation protocols, limiting systematic assessments across multiple cohorts and data formats. To address these challenges, we establish **PeakCloud**, a benchmark covering nine classification tasks across three public datasets and an clinical cohort of 700 subjects. Crucially, this clinical cohort contains paired measurements across two instruments, enabling controlled evaluation of hardware shifts alongside cross-batch settings under unified protocols. Within this benchmark, although CNN-based deep learning models are competitive on several tasks, their reliance on dense spectral grids limits biological explainability. We therefore introduce **PeakCloudNet**, a set-based neural network that processes raw spectra while preserving individual peak identities for attribution. PeakCloudNet achieves competitive performance against grid-based models, accurately pinpoints disease-associated peaks, and retrieves candidate formulas aligned with known metabolic pathways. We hope that this benchmark and architecture will facilitate generalizable and interpretable pattern recognition in metabolomics. Upon publication, we will release the processed peak-level data for all datasets in PeakCloud, together with the PeakCloudNet implementation, to support reproducible research in this area.

1 INTRODUCTION

Untargeted mass spectrometry (MS) profiles thousands of small molecules from non-invasive biospecimens such as breath, urine, and feces, holding immense promise for early disease detection and phenotype screening (Alseekh et al., 2021). However, conventional clinical metabolomics workflows typically restrict analysis to a narrow subset of well-annotated metabolites, leaving the vast majority of spectral signals unexamined. Because experimentally annotating every detected metabolic peak prior to analysis is impractical, data-driven machine learning has emerged as an indispensable approach to discover diagnostic patterns directly from comprehensive, uncurated MS1 profiles (full-scan spectra of intact precursor ions, prior to any fragmentation). For machine learning, an MS1 profile is a variable-sized, unordered set of peaks, and these peaks are chemically related to one another through isotopes, adducts, and fragments (Quinn & Erb, 2020; Mahieu & Patti, 2017).

Despite growing interest, algorithmic progress in learning from MS1 profiles remains hard to assess reliably. Most existing studies optimize on single clinical cohorts using study specific preprocessing and random sample splits, reporting accuracies that are neither cross study comparable nor indicative of operational robustness (Mendez et al., 2019). Random splitting also overlooks batch and instrument drift, which are common in MS platforms (Kirwan et al., 2022; Han & Li, 2022), and clinical case and control samples are frequently collected across different sites or time periods, which a model can exploit instead of learning true pathology. Existing metabolomics benchmarks leave these issues open, focusing mainly on feature table resampling (Mendez et al., 2019) or on molecule identification from tandem spectra rather than metabolomic pattern recognition from MS1 profiles (Bushuiev et al., 2024).

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.

To establish rigorous grounds for comparison, we introduce **PeakCloud**, a standardized benchmark for sample-level metabolomic pattern recognition from MS1 profiles. The benchmark integrates three public liquid chromatography–mass spectrometry (LC–MS) cohorts (serum, urine, and feces) with an clinical proton-transfer-reaction time-of-flight mass spectrometry (PTR-TOF-MS) breath cohort of approximately 700 subjects, collected by collaborating clinical sites and provided in de-identified form. All datasets are unified into a shared peak-level format supporting both two-channel (m/z, intensity) and three-channel (m/z, retention time, intensity) spectra under fixed, subject-disjoint splits. We evaluate candidate methods under two complementary regimes: in-distribution (ID) discrimination and out-of-distribution (OOD) generalization across cross-instrument, temporal-batch, and external-cohort shifts, and we further examine every model with an attribution analysis for explainability. For the breath cohort, each subject was measured concurrently on two instruments, which lets us isolate hardware drift on matched biology, and every sample has a paired ambient air reference that we use during preprocessing to remove the influence of the collection environment from the breath spectrum before training. To our knowledge, MSP001 is the first public breath PTR-TOF-MS cohort to offer paired cross-instrument measurements over longitudinal batches.

Benchmarking existing methods also exposes a trade-off between alignment and discretization. Feature-table workflows depend on multi-sample alignment, where heuristic parameters decide which features are kept (Myers et al., 2017), while convolutional neural networks avoid alignment by binning continuous spectra into dense grids (Deng et al., 2024; Menacher et al., 2026), but this binning blurs spectral resolution and obscures individual peak identities. To resolve this trade-off, we propose **PeakCloudNet**. Across our experiments, high in-distribution accuracy does not always carry over to real-world robustness: classical tree ensembles perform well within their training domain but degrade under cross-instrument transfer, whereas PeakCloudNet transfers best in both directions, and its attribution analysis shows that its predictions rely on sparse, genuine chemical signals rather than intensity artifacts, and it recovers metabolic markers consistent with those reported in prior clinical studies (Mathé et al., 2014).

In summary, our primary contributions are:

1. An open-source benchmark featuring a unified peak-level format, fixed subject-disjoint splits, and evaluation protocols spanning in-distribution performance and out-of-distribution robustness, together with an attribution analysis for explainability.
2. The first public dataset of its kind, comprising 700 subjects measured across two instruments with longitudinal acquisition batches and paired environmental references, which together enable auditing hardware transfer and removing collection-site confounding. Upon publication, and subject to authorization from the data-providing institution, we will release the environment-filtered peak-level data for MSP001, together with our processed peak-level versions of the public datasets and the PeakCloudNet codebase.
3. A set-based framework that eliminates alignment and binning via resolution-aware Fourier and compositional transformations, reconciling competitive predictive accuracy with peak-level verifiable evidence.

2 Related Work

Mass-spectrometry metabolomics and public data resources. Mass spectrometry has become a central analytical platform for untargeted metabolomics (Patti et al., 2012; Alseekh et al., 2021), and repositories such as MetaboLights, Metabolomics Workbench, and GNPS/MassIVE have accumulated large collections of MS-based data (Yurekten et al., 2024). However, repository scale does not imply datasets suitable for machine learning (El Abiead et al., 2025), public multi-batch benchmarks are uncommon (Pelletier et al., 2024), cross-instrument datasets are rarer still (Martin et al., 2015), and the gap is especially pronounced for PTR-TOF-MS volatolomics (Roquencourt et al., 2022). Resources that simultaneously provide large cohorts, raw MS signals, explicit batch structure, and matched cross-instrument measurements thus remain scarce.

Learning from MS metabolomic profiles. Untargeted metabolomics commonly detects and aligns peaks into a feature table, then applies methods such as PLS-DA, random forests, and SVMs to identify discriminative biomarkers (Smith et al., 2006; Lange et al., 2008; Alseekh et al., 2021).

This pipeline has two limitations: alignment tools such as XCMS and MZmine require many parameter choices that affect which features are retained (Myers et al., 2017; Alseekh et al., 2021), and the resulting high-dimensional tables are hard to learn from reliably (Guan et al., 2009; Gaul et al., 2015; Mendez et al., 2019). Deep-learning methods on less processed signals including DeepPseudoMSI on LC–MS pseudo-images (Shen et al., 2022), and DeepMSProfiler and LCMS-Net on raw or near-raw signals (Deng et al., 2024; Menacher et al., 2026), reduce reliance on feature tables, but dense spectral binning obscures individual peak identities and cross-batch and cross-instrument generalization remains only partly addressed (Deng et al., 2024; Menacher et al., 2026). Our framework instead learns directly from individual peaks without alignment, preserving peak-level identity for the evaluations that follow.

Learning on Unordered Sets and Point Clouds. Deep Sets (Zaheer et al., 2017) and PointNet (Qi et al., 2017) handle permutation-invariant inputs; Set Transformer extends this with attention across elements (Lee et al., 2019), better capturing inter-peak relationships such as isotopes and adducts (Kuhl et al., 2012). In mass spectrometry, set-based methods have been applied at the spectrum level for peptide sequencing (Qiao et al., 2021; Yilmaz et al., 2024) and representation learning (Bushuiev et al., 2026), and at the profile level for feature detection (Wang et al., 2023) but not yet for phenotype prediction from whole untargeted profiles, which is the gap this work addresses.

3 PeakCloud Benchmark

PeakCloud standardizes disease classification from untargeted metabolomics MS1 profiles across datasets, input representations, and acquisition domains. It jointly evaluates classification performance and robustness to distribution shifts.

Tasks and datasets. A centroid MS1 profile represents measured signals as discrete ions with centroid m/z and intensity, plus retention time for LC–MS. Each benchmark task is formulated as $\mathcal{D} = \{(X_i, y_i, d_i)\}_{i=1}^{N}$, where X_i is this two- or three-channel peak profile, y_i is the disease label, and d_i denotes the acquisition domain, aiming to learn $f_\theta(X_i)$ under subject-disjoint in-distribution (ID) and out-of-distribution (OOD) protocols.

However, public mass spectrometry repositories (e.g., MetaboLights, Metabolomics Workbench, GNPS/MassIVE) exhibit substantial heterogeneity in data representations: many cohorts do not distribute raw peak-picked centroid profiles, but rather heavily filtered, pre-annotated metabolite tables or raw continuous intensities, precluding direct, peak-level learning. To establish a standardized benchmark, we systematically screened these repositories and unified qualifying raw centroid MS1 spectra with adequate sample sizes into a consistent peak format (Appendix A.1), yielding three public LC–MS cohorts: MTBLS17 (Ressom et al., 2012), MTBLS28 (Mathé et al., 2014), and ST001000 (Franzosa et al., 2019). Since public cohorts lack multi-instrument and longitudinal cross-batch settings, we additionally curate MSP001, a large-scale breath cohort of 700 subjects (tuberculosis cases and matched healthy controls) collected by collaborating clinical sites and provided in de-identified form. Subjects were measured concurrently across two separate PTR-TOF-MS instruments (designated as MSP001-1 and MSP001-2) with paired ambient-air references (Appendix A.4). Table 1 summarizes all benchmark tasks and dataset specifications.

Compared representations and models. With the datasets fixed, PeakCloud compares three representation families under identical subject partitions. For LC–MS, traditional methods operate on feature tables generated by the XCMS workflow, which detects chromatographic peaks, aligns them across samples, groups corresponding signals, and quantifies the resulting features (Smith et al., 2006; Tautenhahn et al., 2012); the PTR-TOF-MS counterpart uses an aligned peak-intensity table. Grid-based end-to-end models, DeepMSProfiler and LCMS-Net, discretize spectra for convolutional learning (Deng et al., 2024; Menacher et al., 2026). PeakCloudNet instead processes unordered centroid peaks directly, preserving their measured coordinates for attribution.

Evaluation under distribution shifts. We establish a standardized evaluation protocol to assess the robustness and generalizability of each representation under controlled distribution shifts (Table 1). The protocol covers three complementary forms of shift: cross-dataset transfer between MSP001-1 and MSP001-2, temporal shift across acquisition batches within MSP001, and cross-venue shift in ST001000. Together, these settings probe whether representation quality is preserved

Table 1: Benchmark datasets, prediction tasks, and evaluation settings. Each peak-level data contains multiple peaks associate intensities with m/z coordinates (PTR-TOF-MS) or $(m/z, \mathrm{RT})$ coordinates (LC–MS). ID denotes in-distribution; OOD denotes out-of-distribution; TB denotes tuberculosis; HCC denotes hepatocellular carcinoma; IBD denotes inflammatory bowel disease.

Dataset	Reference	Task	Specimen	Platform (mode)	ID/OOD	OOD setting
MSP001 ($n = 700$)	This work	TB vs. control	Breath	PTR-TOF-MS (+)	ID & OOD	cross-instrument; temporal batch
MTBLS17 ($n = 262$)	Ressom et al. (2012)	HCC vs. cirrhosis	Serum	LC–MS (+/−)	ID	–
MTBLS28 ($n = 1004$)	Mathé et al. (2014)	Lung cancer vs. control	Urine	LC–MS (+/−)	ID	–
ST001000 ($n = 220$)	Franzosa et al. (2019)	IBD vs. control	Feces	LC–MS (−)	OOD	cross-venue

when the data distribution changes across cohorts, collection time, and acquisition environments. We further include a paired instrument-shift setting on MSP001 dataset, where source- and target-domain measurements are obtained from the same subjects, allowing us to isolate the effect of hardware variation from differences in cohort composition. Across all settings, target-domain data are held out from model fitting and model selection, providing a consistent protocol for evaluating robustness to distribution shift.

4 PeakCloudNet: Physically informed peak encoder

We propose PeakCloudNet, which directly processes unordered, variable-sized peak set; two-channel peaks contain m/z and intensity for PTR-TOF-MS, while LC–MS peaks additionally contain retention time (RT). For i-th peak, a three-channel-encoder is applied to encode it into embedding ϕ_i, where $\phi_i = [\phi_i^{\mathrm{mz}}, \phi_i^I, \phi_i^r]$ is the fusion of separate embeddings of peak m/z, intensity and RT. Specifically, a simple within-profile normalization is applied to RT denoted by r_i, i.e., $\phi_i^r = r_i / r_i^{\max}$, whereas m/z and intensity require specialized embeddings that account for measurement resolution and relative signal scale. The resulting peak representations are processed and pooled using a Set Transformer for sample-level classification, as summarized in Figure 1. During training, we use low-intensity deletion, resolution-bounded m/z shifts, intensity and baseline noise, and RT jitter when applicable; complete definitions of each are given in Appendix Table 5. We next describe the two physically informed encoders for m/z and intensity.

Resolution-aware m/z embedding. For measured mass m_i, PeakCloudNet uses the resolution-aware Fourier representation

$$\boldsymbol{\phi}_i^{mz} = [d_{ik}\cos(\omega_k m_i)\,;\, d_{ik}\sin(\omega_k m_i)]_{k=1}^{K}, \qquad d_{ik} = \exp\left(-\frac{1}{2}\omega_k^2\delta_i^2\right), \tag{1}$$

where ω_k represents a mass scale with wavelength $2\pi/\omega_k$, and δ_i is the effective mass measurement uncertainty at m_i (set by the instrument's resolustion). Gaussian damping then suppresses frequencies whose wavelength falls below this resolution limit, so features finer than δ_i do not propagate into ϕ_i^{mz}. The Fourier frequencies ω_k are fixed throughout training, while the resulting features ϕ_i^{mz} are mapped to m/z embeddings through a trainable projection.

Logarithm-rate-based intensity embedding. Because relative peak intensities in MS are more informative than absolute scale (Quinn & Erb, 2020), we apply the centered log-ratio (CLR) transform for a P-peaks MS as

$$\phi_i^I = clr(I_i) = \log I_i - \frac{1}{P}\sum_{i=1}^{P}\log I_i = \log\frac{I_i}{g(\mathbf{I})}, \qquad g(\mathbf{I}) = \left(\prod_{l=1}^{P} I_l\right)^{1/P}, \tag{2}$$

and map the centered values through a shared linear projection and LeakyReLU. We use $\log(1+I)$ to accommodate zeros; masked centering over valid peaks preserves permutation equivariance.

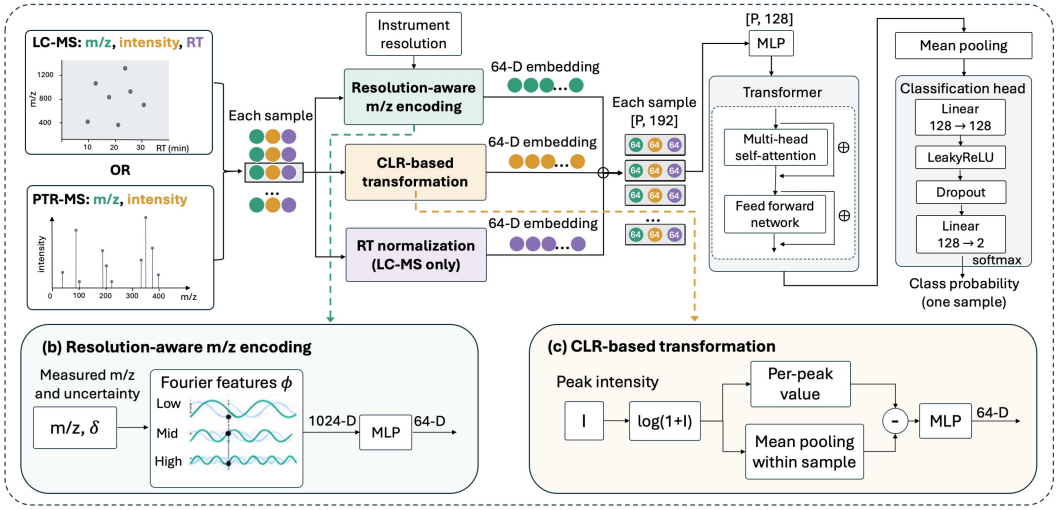

Figure 1: Overview of PeakCloudNet. (a) Each mass spectrometry sample is an unordered set of P peaks described by m/z, intensity, and, for LC–MS, retention time (RT). Channel-specific 64-dimensional embeddings are fused by an MLP, processed with Transformer self-attention, and mean-pooled for classification. (b) Resolution-aware m/z encoding damps Fourier features according to instrument-dependent uncertainty δ, suppressing frequencies finer than the measurement resolution. (c) CLR intensity encoding centers log-intensities by their within-sample mean, removing sample-level scale while remaining permutation-equivariant.

Peak-level attribution. To illustrate the explainability of the proposed model, we compute Peak-CAM attributions from the final per-peak representations before pooling. For spectrum, the target $s_i = z_i^+ - z_i^-$ is the logit margin between the task's positive and negative classes. Let ϕ_i^k denote channel k of the representation of valid peak p, and let P_i be the number of valid peaks. We compute

$$\alpha_{ik} = \frac{1}{P}\sum_{p=1}^{P} \frac{\partial s_i}{\partial \phi_i^k}, \qquad a_i = \sum_k \alpha_{ik}\phi_i^k. \tag{3}$$

without a ReLU, so positive and negative values support the positive and negative classes, respectively. Scores are normalized within each spectrum by their absolute sum. Cohort-level region construction, faithfulness tests, and candidate molecular-formula retrieval are detailed in the Appendix (Sections A.7 and A.9).

5 Experiments

Experiments instantiate the PeakCloud protocols defined in Section 3; ablations and attribution analyses further examine PeakCloudNet and its spectral evidence.

5.1 Experimental setup

Dataset setting. Dataset characteristics are summarized in Table 1. MTBLS17 and MTBLS28 are used for ID evaluation, whereas ST001000 supports OOD evaluation. The MSP001 dataset supports ID, cross-instrument, and temporal batch evaluations. Cross-instrument transfer is evaluated in both directions between MSP001-1 and MSP001-2 using paired measurements from 140 held-out subjects, comprising 70 TB subjects and 70 healthy controls. For each direction, measurements from the source and target instruments form the ID and OOD test sets, respectively. Temporal evaluation uses one acquisition batch as the source domain and another as the OOD test domain. All evaluations follow the subject-disjoint splitting protocol defined in the benchmark. Model selection uses training and validation data only. MSP001-1 and MSP001-2 each comprise a class-balanced development set of 560 participants and a class-balanced test set of 140 participants, with participant assignments shared across instruments. Cross-instrument evaluation uses measurements from the corresponding 140 test participants acquired on the other instrument. Cross-batch evaluation reserves 11 complete acquisition dates for testing, with the remaining dates forming the development set. For MTBLS17

Table 2: In-distribution performance on four mass-spectrometry datasets.

	MSP001-1 *Internal*		MSP001-2 *Internal*		MTBLS17 *POS+NEG*		MTBLS28 *POS+NEG*	
Method	**AUC**	**BA**	**AUC**	**BA**	**AUC**	**BA**	**AUC**	**BA**
Traditional machine-learning baselines								
RF	0.94±0.04	0.89±0.06	**0.97±0.02**	0.92±0.05	0.68±0.11	0.60±0.08	0.71±0.06	0.66±0.06
SVM	0.89±0.05	0.82±0.06	0.79±0.10	0.78±0.08	0.62±0.11	0.60±0.06	0.66±0.07	0.62±0.06
PLS-DA	0.86±0.09	0.83±0.06	0.93±0.03	0.87±0.04	0.66±0.12	0.57±0.05	0.72±0.07	0.65±0.06
PCA + LR	0.79±0.07	0.77±0.08	0.79±0.09	0.76±0.07	0.52±0.13	0.49±0.10	0.56±0.07	0.57±0.06
MLP	0.92±0.04	0.84±0.06	0.93±0.04	0.87±0.06	0.48±0.12	0.50±0.06	0.53±0.07	0.54±0.06
Deep-learning baselines								
DeepMSProfiler	0.87±0.06	0.86±0.04	0.91±0.05	0.86±0.06	0.68±0.11	**0.65±0.08**	0.60±0.07	0.58±0.06
LCMS-Net	0.93±0.04	0.88±0.06	0.93±0.04	0.86±0.05	0.65±0.12	0.63±0.09	**0.80±0.06**	**0.74±0.05**
Proposed method								
PeakCloudNet (Ours)	**0.97±0.02**	**0.93±0.03**	**0.97±0.02**	**0.94±0.03**	**0.69±0.09**	**0.65±0.09**	0.75±0.06	0.70±0.06

and MTBLS28, we use the first fold of stratified three-fold and five-fold partitions, respectively, as the test set and the remaining folds as the development set. Participant assignments are consistent across positive and negative ionization modes, and technical replicates in MTBLS17 are grouped together. ST001000 comprises 155 development participants; traditional methods and neural networks are evaluated on 65 external test participants. Uncertainty is estimated using 2,000 bootstrap resamples. PTR-TOF-MS data are resampled by acquisition date, retaining all observations within each sampled date, whereas non-PTR data are resampled at the participant level within each class, keeping technical replicates in MTBLS17 together. Predictions and classification thresholds remain fixed during resampling. Results are reported as point estimates ±1.96 bootstrap standard errors, yielding normal-approximation 95% confidence intervals.

Evaluation protocol. Performance is evaluated using the area under the receiver operating characteristic curve (AUC) and balanced accuracy (BA), where $\mathrm{BA} = (\mathrm{sensitivity} + \mathrm{specificity})/2$. Reported results are point estimates ± 1.96 bootstrap standard errors based on 2,000 resamples. Bootstrap resampling is performed at the collection-date level for PTR-TOF-MS data and at the subject level for LC–MS data, with technical replicates grouped within subjects. Predicted probabilities are averaged over three random seeds before computing the evaluation metrics.

5.2 Main results

5.2.1 In-distribution results

As shown in Table 2, PeakCloudNet achieves the best overall performance across both PTR-TOF-MS breath datasets, reaching an AUC of 0.97 on each instrument and tying or outperforming Random Forest. Grid-based CNNs perform worse than tree models on these cohorts. This drop occurs because PTR-TOF-MS spectra are sparse and lack retention time, which makes dense grid representations inefficient. On the more difficult LC–MS datasets with high class overlap, PeakCloudNet remains highly competitive. Overall, PeakCloudNet matches or exceeds grid-based models without peak alignment or binning, while keeping individual peak identities intact for attribution.

5.2.2 Out-of-distribution results

Cross-instrument transfer. Under instrument shift, PeakCloudNet achieves the best transfer in both directions, whereas high ID accuracy does not guarantee robustness for either feature-table or grid-based baselines (Table 3(a)). Each model is trained on one instrument and tested on paired measurements from the other, so performance changes reflect hardware differences, and the ID and OOD results share the same training samples. RF, the strongest baseline in-distribution, loses 0.21 and 0.14 AUC, respectively. Grid-based CNNs are direction-sensitive: DeepMSProfiler retains an AUC of 0.80 in the first direction but falls to chance level (0.50) in the second. PeakCloudNet, by comparison, loses only 0.09 and 0.02 AUC and attains the highest AUC and BA in both directions. Robustness to instrument shift thus likely depends on building in the relevant measurement invariances rather than on ID accuracy or end-to-end training alone: PeakCloudNet's resolution-aware mass encoding damps m/z differences below the instrument's mass accuracy, and its log-ratio intensities remove global scale differences between devices.

Table 3: Out-of-distribution performance across instruments, batches, and cohorts.

(a) Cross-instrument evaluation

Method	MSP001-1 → MSP001-2 AUC	MSP001-1 → MSP001-2 BA	MSP001-2 → MSP001-1 AUC	MSP001-2 → MSP001-1 BA
Traditional				
RF	0.73±0.09	0.67±0.08	0.83±0.08	0.79±0.07
SVM	0.84±0.07	0.77±0.08	0.77±0.11	0.75±0.08
PLS-DA	0.73±0.10	0.69±0.07	0.91±0.05	0.85±0.06
PCA + LR	0.53±0.13	0.58±0.07	0.59±0.13	0.61±0.10
MLP	0.67±0.10	0.64±0.07	0.79±0.09	0.75±0.09
Deep-learning				
DeepMSProfiler	0.80±0.09	0.77±0.09	0.50±0.10	0.55±0.04
LCMS-Net	0.82±0.07	0.80±0.07	0.73±0.10	0.69±0.09
Proposed				
PeakCloudNet	**0.88±0.05**	**0.82±0.06**	**0.95±0.03**	**0.88±0.06**

(b) Cross-batch and external-cohort evaluation

Method	MSP001-1 Cross-batch AUC	MSP001-1 Cross-batch BA	MSP001-2 Cross-batch AUC	MSP001-2 Cross-batch BA	ST001000 External cohort AUC	ST001000 External cohort BA
Traditional						
RF	0.92±0.05	0.85±0.06	0.92±0.04	0.87±0.04	0.92±0.07	0.74±0.08
SVM	0.68±0.14	0.71±0.11	0.65±0.10	0.63±0.09	**0.94±0.06**	**0.86±0.07**
PLS-DA	0.69±0.08	0.67±0.06	0.79±0.09	0.74±0.10	0.93±0.06	0.79±0.08
PCA + LR	0.57±0.12	0.60±0.07	0.65±0.07	0.62±0.07	0.88±0.08	0.79±0.09
MLP	0.71±0.12	0.69±0.09	0.76±0.09	0.70±0.09	0.72±0.12	0.72±0.08
Deep-learning						
DeepMSProfiler	0.81±0.05	0.80±0.04	0.86±0.08	0.79±0.08	0.81±0.10	0.72±0.11
LCMS-Net	0.88±0.06	0.82±0.07	0.75±0.09	0.72±0.07	0.89±0.07	0.74±0.11
Proposed						
PeakCloudNet	**0.94±0.06**	**0.89±0.06**	**0.94±0.04**	**0.88±0.06**	0.90±0.07	**0.86±0.08**

Temporal batch and external-cohort shifts. Under batch and cohort shifts, PeakCloudNet remains among the strongest methods, but RF baselines also stay competitive, in contrast to their degradation across instruments (Table 3(b)). Cross-batch evaluation trains on one acquisition batch and tests on another from the same instrument, whereas ST001000 tests transfer to an independent LC–MS cohort. Across batches, PeakCloudNet ranks first on both instruments, and RF ranks second, ahead of both grid-based CNNs. Robustness therefore cannot be assessed from a single type of shift: judged by batch or cohort transfer alone, RF would appear nearly as robust as PeakCloudNet, and only the paired cross-instrument evaluation exposes the gap between them.

5.3 Ablation study

We assess m/z encoding, intensity encoding and the self-attention block through targeted replacements with raw m/z, raw intensity, and MLP, keeping the original splits, augmentation, training budget, and backbone fixed. On MSP001-1 and MSP001-2, we report same-instrument AUC.

Contribution of m/z encoding. To assess the m/z representation, we keep the original intensity branch and self-attention block, replacing only the resolution-aware mass embedding with a single-layer MLP operating on raw m/z. This replacement reduces same-instrument AUC dramatically, which reveals the necessity of m/z encoding. (Table 4).

Contribution of intensity encoding. With the resolution-aware mass embedding and self-attention block retained, we replace the CLR-based intensity embedding with a single-layer MLP on raw intensity. This replacement omits both the logarithmic transformation and spectrum-level centering. The resulting AUC is 0.89 on MSP001-1 and 0.69 on MSP001-2. Replacing both encoders while retaining self-attention yields 0.77 and 0.65, respectively. The joint replacement is therefore not uniformly worse than either individual replacement across datasets.

Contribution of the self-attention block. To assess self-attention, we feed the concatenated branch embeddings directly into a three-layer MLP, retaining the original masked mean pooling and classifier. With both encoders retained, this variant achieves AUC of 0.92 on MSP001-1 and 0.96 on MSP001-2, compared with 0.97 for the full model on each dataset. Replacing both encoders and the self-attention block yields 0.70 and 0.63, respectively. The full model has the highest AUC point estimate among the eight configurations on both datasets.

Table 4: Ablation study. ✓/✗: component enabled/removed.

m/z encoder	✓	✗	✓	✗	✓	✗	✓	✗
Intensity encoder	✓	✓	✗	✗	✓	✓	✗	✗
Self-attention	✓	✓	✓	✓	✗	✗	✗	✗
MSP001-1	**0.97**	0.62	0.89	0.77	0.92	0.65	0.60	0.70
MSP001-2	**0.97**	0.72	0.69	0.65	0.96	0.54	0.58	0.63

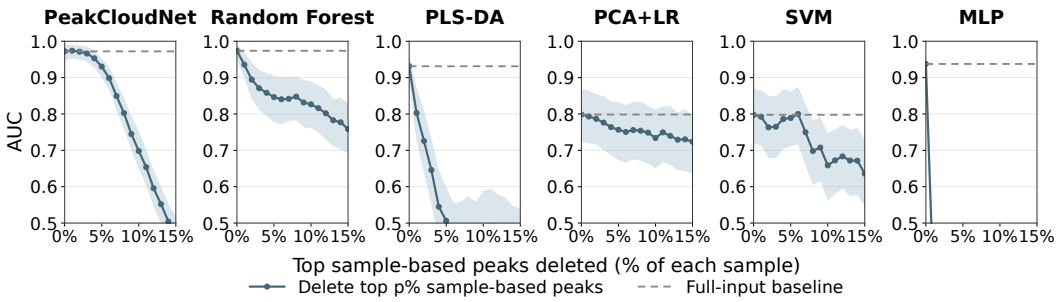

Figure 2: Sample-based peak perturbation on MSP001. For each test spectrum, peaks were ranked by absolute Peak-CAM attribution for PeakCloudNet or model-specific attribution for the baseline classifiers, and the top 0–15% were removed. Solid curves show mean AUC across three seeds; dashed lines indicate each model's full-input baseline. Shaded bands represent pointwise 95% confidence intervals. Attribution methods are described in the Appendix.

5.4 Explainability analysis

Sparse spectral evidence within individual samples. To examine whether Peak-CAM identifies peaks that are functionally relevant to the classifier, we removed the highest-attribution peaks independently from each test spectrum. Removing the top 5% of Peak-CAM-ranked peaks reduced AUC from 0.9722 to 0.9304, whereas intensity-matched random removal produced little change (0.9671). Removing the top 10% further reduced AUC to 0.6981, compared with 0.9619 under random removal. These interventions indicate that PeakCloudNet's predictions depend on a relatively sparse subset of peaks within each spectrum. We also performed sample-based peak perturbation using model-specific attributions for five baseline classifiers (Figure 2). The rate of degradation reflects how each model handles deleted peaks. PLS-DA and MLP collapse within the first few percent, as zeroed bins become out-of-distribution values in their standardized feature space. Random Forest plateaus after an initial drop because correlated features compensate, while PCA+LR and SVM decline slowly as their decisions are spread across many peaks. In contrast, PeakCloudNet removes peaks as set elements and is trained with deletion augmentation, so it remains stable under small deletions and degrades sharply only once its attributed evidence is exhausted.

Heterogeneous evidence across subjects. To assess whether this sparse predictive evidence is shared across subjects, we identified cohort-level global m/z regions using the training and validation subjects and froze them before held-out test evaluation. Retaining only peaks within globally recurrent regions did not recover full-input discrimination (Figure 3). Global recurrence also showed limited correspondence with local attribution rankings. Among the ten highest-ranked global regions, only 14.7% of matched occurrences belonged to the corresponding sample's top 5% attribution set, and 31.2% belonged to its top 10%. Thus, regions recurring across the cohort were not consistently among the most influential peaks within individual spectra. Together, these findings suggest that PeakCloudNet relies on sparse but heterogeneous spectral evidence across subjects, with a fixed set of recurrent m/z regions capturing only part of the sample-specific evidence.

Biological explainability of peak-level features. We use pulmonary tuberculosis (TB) breath mass spectra as a case study to investigate the potential biological relevance of peak-level features identified by PeakCloudNet. Using the candidate molecular-formula retrieval procedure described in Appendix A.9, we annotate the top 200 global peaks with candidate formulas and interpret their associated chemical classes and possible connections to TB in the context of proton-transfer-reaction mass spectrometry (PTR-TOF-MS) and prior studies. Previous work suggests that TB-associated breath profiles may reflect contributions from both microbial activity and host oxidative stress, providing a biological context for interpreting these features (Phillips et al., 2007).

For the peak at m/z 115.11, molecular-formula retrieval returns $C_7H_{14}O$ as a neutral-formula candidate, corresponding to the protonated ion $C_7H_{15}O^+$, with a mass error of -51.76 ppm. This formula is compatible with 3-heptanone and other isomers. As an aliphatic ketone, 3-heptanone belongs to the class of carbonyl-containing volatile organic compounds. Previous work detected

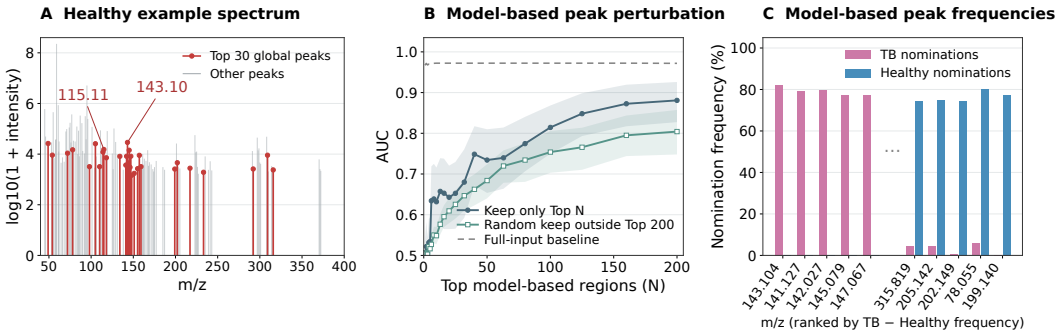

Figure 3: Model-based peak analysis on MSP001. (A) A healthy test spectrum with detected peaks from the model-based Top 30 highlighted in red. Peaks matching m/z 115.11 and 143.10 are labeled by their nominal target values. (B) AUC when retaining the top N model-based regions, compared with retaining the same number of peaks sampled without replacement from outside the Top 200. The dashed line indicates the full-input baseline. (C) Nomination frequencies of the five regions with the largest and five with the smallest TB–healthy frequency differences.

3-heptanone among the more abundant volatiles in culture systems involving *Mycobacterium tuberculosis*, suggesting a potential connection to microbial-associated volatile profiles. This peak therefore provides a ketone-related candidate for further investigation.

For the peak at m/z 143.10, retrieval returns $C_{11}H_{10}$ as a neutral-formula candidate, corresponding to the protonated ion $C_{11}H_{11}^{+}$, with a mass error of $+75.92$ ppm. Its candidates include 1-methylnaphthalene, 2-methylnaphthalene, and other isomers. An aromatic hydrocarbon, 1-methylnaphthalene has been detected in culture systems involving *M. tuberculosis* and reported as a TB-associated volatile in human breath studies. These findings identify methylnaphthalenes as a candidate chemical class linking the current peak-level feature to previously reported TB-associated breath volatiles.

Biological explainability on MTBLS28. Additionally, we extend the explainability analysis to the MTBLS28 dataset. The global peaks identified by PeakCloudNet include features annotated as creatine riboside, N-acetylneuraminic acid (NANA), and the 561^{+} feature, consistent with the urinary biomarkers reported by Mathé et al. (2014). That study found all three features to be elevated in lung cancer patients relative to population controls, with creatine riboside and NANA also enriched in tumors compared with paired adjacent tissues. As discussed in the study, creatine riboside may be associated with altered creatine-related energy metabolism, whereas NANA is linked to sialic-acid biology and the roles of sialylated glycoconjugates in cell recognition and tumor–immune interactions. The 561^{+} feature was tentatively characterized as a glucuronidated compound, although its exact structure and biological role remained unresolved. These findings show that PeakCloudNet highlights peaks consistent with previously reported lung cancer-associated metabolic signals, providing biological context for its learned representations. Results on MTBLS17 and ST001000 are reported in Appendix A.8.

6 Conclusion

This work establishes a standardized benchmark for sample-level prediction from MS1 profiles, bringing feature-table methods and end-to-end models into a shared evaluation framework. Unified data partitions and evaluation protocols enable comparisons of in-distribution performance and generalization across instruments, temporal batches, and external cohorts. PeakCloudNet shows strong predictive performance and generalization on the PTR-TOF-MS evaluations, whereas grid-based CNNs are less consistent across shift directions, while the varying method rankings on LC–MS datasets underscore the need for evaluation across complementary settings.

Within this benchmark, PeakCloudNet offers a competitive balance between classification performance and peak-level explainability through its set-based representation. Attribution and peak-removal analyses indicate that its predictions rely on sparse, sample-specific spectral evidence that varies across subjects. Together, PeakCloud and PeakCloudNet provide a basis for further research on generalizable and interpretable learning from mass spectrometry profiles.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing the writing and with literature discovery, including identifying related work for citation. We did not use generative AI tools for research ideation, experimental design, code development, data analysis, or generation of results. All AI-assisted text was reviewed and revised by the authors, and all literature identified through generative AI tools was verified against the original sources before being cited or discussed. The authors take full responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

The MSP001 breath data were collected by collaborating clinical sites (a hospital outpatient clinic and a affiliated health examination center) as part of a study approved by the site's institutional ethics committee. Written informed consent was obtained from all participants by the collaborating clinical sites.

All personally identifiable information and direct patient identifiers were removed by the collaborating clinical sites prior to the data being provided to the authors for research use. The dataset available to the authors contains only the PTR-TOF-MS measurements and the non-identifying information used in this study, and the authors had no access to information that would enable re-identification of individual participants.

The present study involved only secondary analysis of the de-identified data and did not involve any direct interaction with human participants. The authors did not attempt to re-identify any individual represented in the dataset.

Upon publication, and as authorized by the data-providing institution, we will release the environment-filtered, de-identified peak-level MSP001 data, in a form that retains no patient identifiers, replaces exact collection dates with anonymized batch indices, and complies with the data-sharing terms approved by the originating institution's ethics committee.

REPRODUCIBILITY STATEMENT

The PeakCloud benchmark and the PeakCloudNet implementation, including data preprocessing scripts, training and evaluation code will be made publicly available upon publication. Architectural and training details for PeakCloudNet are described in Section 4. All reported metrics use fixed, subject-disjoint splits (Section 5.1) and bootstrap resampling procedures described therein, which are deterministic given the released splits and random seeds.

REFERENCES

Saleh Alseekh, Asaph Aharoni, Yariv Brotman, Kévin Contrepois, John D'Auria, Jan Ewald, Jennifer C. Ewald, Paul D. Fraser, Patrick Giavalisco, Robert D. Hall, Matthias Heinemann, Hannes Link, Jie Luo, Steffen Neumann, Jens Nielsen, Leonardo Perez de Souza, Kazuki Saito, Uwe Sauer, Frank C. Schroeder, Stefan Schuster, Gary Siuzdak, Aleksandra Skirycz, Lloyd W. Sumner, Michael P. Snyder, Huiru Tang, Takayuki Tohge, Yulan Wang, Weiwei Wen, Si Wu, Guowang Xu, Nicola Zamboni, and Alisdair R. Fernie. Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nature Methods*, 18:747–756, 2021. doi: 10.1038/s41592-021-01197-1.

Roman Bushuiev, Anton Bushuiev, Niek F de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, Marcus Ludwig, Nils A. Haupt, Apurva Kalia, Corinna Brungs, Robin Schmid, Russell Greiner, Bo Wang, David S. Wishart, Li-Ping Liu, Juho Rousu, Wout Bittremieux, Hannes Rost, Tytus D. Mak, Soha Hassoun, Florian Huber, Justin J. J. van der Hooft, Michael A. Stravs, Sebastian Böcker, Josef Sivic, and Tomáš Pluskal. MassSpecGym: A benchmark for the discovery and identification of molecules. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110010–110027. Curran Associates, Inc., 2024.

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*, 44:630–640, 2026. doi: 10.1038/s41587-025-02663-3.

Yongjie Deng, Yao Yao, Yanni Wang, Tiantian Yu, Wenhao Cai, Dingli Zhou, Feng Yin, Wanli Liu, Yuying Liu, Chuanbo Xie, Jian Guan, Yumin Hu, Peng Huang, and Weizhong Li. An end-to-end deep learning method for mass spectrometry data analysis to reveal disease-specific metabolic profiles. *Nature Communications*, 15:7136, 2024. doi: 10.1038/s41467-024-51433-3.

Yasin El Abiead, Michael Strobel, Thomas Payne, Eoin Fahy, Claire O'Donovan, Shankar Subramamiam, Juan Antonio Vizcaíno, Ozgur Yurekten, Victoria Deleray, Simone Zuffa, Shipei Xing, Helena Mannochio-Russo, Ipsita Mohanty, Haoqi Nina Zhao, Andres M. Caraballo-Rodriguez, Paulo Wender P. Gomes, Nicole E. Avalon, Trent R. Northen, Benjamin P. Bowen, Katherine B. Louie, Pieter C. Dorrestein, and Mingxun Wang. Enabling pan-repository reanalysis for big data science of public metabolomics data. *Nature Communications*, 16(1):4838, 2025. doi: 10.1038/s41467-025-60067-y.

Eric A. Franzosa, Alexandra Sirota-Madi, Julian Avila-Pacheco, Nadine Fornelos, Henry J. Haiser, Stefan Reinker, Tommi Vatanen, A. Brantley Hall, Himel Mallick, Lauren J. McIver, Jenny S. Sauk, Robin G. Wilson, Betsy W. Stevens, Justin M. Scott, Kerry Pierce, Amy A. Deik, Kevin Bullock, Floris Imhann, Jeffrey A. Porter, Alexandra Zhernakova, Jingyuan Fu, Rinse K. Weersma, Cisca Wijmenga, Clary B. Clish, Hera Vlamakis, Curtis Huttenhower, and Ramnik J. Xavier. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology*, 4:293–305, 2019. doi: 10.1038/s41564-018-0306-4.

David A. Gaul, Roman Mezencev, Tran Q. Long, Christina M. Jones, Benedict B. Benigno, Alexander Gray, Facundo M. Fernández, and John F. McDonald. Highly-accurate metabolomic detection of early-stage ovarian cancer. *Scientific Reports*, 5:16351, 2015. doi: 10.1038/srep16351.

Wei Guan, Manshui Zhou, Christina Y. Hampton, Benedict B. Benigno, L. DeEtte Walker, Alexander Gray, John F. McDonald, and Facundo M. Fernández. Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. *BMC Bioinformatics*, 10:259, 2009. doi: 10.1186/1471-2105-10-259.

Wei Han and Liang Li. Evaluating and minimizing batch effects in metabolomics. *Mass Spectrometry Reviews*, 41(3):421–442, 2022. doi: 10.1002/mas.21672.

Jennifer A. Kirwan, Helen Gika, Richard D. Beger, Dan Bearden, Warwick B. Dunn, Royston Goodacre, Georgios Theodoridis, Michael Witting, Li-Rong Yu, and Ian D. Wilson. Quality assurance and quality control reporting in untargeted metabolic phenotyping: mQACC recommendations for analytical quality management. *Metabolomics*, 18(9):70, 2022. doi: 10.1007/s11306-022-01926-3.

Carsten Kuhl, Ralf Tautenhahn, Christoph Böttcher, Tony R. Larson, and Steffen Neumann. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry*, 84(1):283–289, 2012. doi: 10.1021/ac202450g.

Eva Lange, Ralf Tautenhahn, Steffen Neumann, and Clemens Gröpl. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics*, 9:375, 2008. doi: 10.1186/1471-2105-9-375.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019.

Nathaniel G. Mahieu and Gary J. Patti. Systems-level annotation of a metabolomics data set reduces 25,000 features to fewer than 1,000 unique metabolites. *Analytical Chemistry*, 89(19):10397–10406, 2017. doi: 10.1021/acs.analchem.7b02380.

Jean-Charles Martin, Matthieu Maillot, Gerard Mazerolles, Alexandre Verdu, Bernard Lyan, Carole Migné, Catherine Defoort, Cécile Canlet, Christophe Junot, Claude Guillou, Claudine Manach, Daniel Jabob, Delphine Jouan-Rimbaud Bouveresse, Estelle Paris, Estelle Pujos-Guillot, Fabien Jourdan, Franck Giacomoni, Frédérique Courant, Gaëlle Favé, Gwenaëlle Le Gall, Hubert Chassaigne, Jean-Claude Tabet, Jean-Francois Martin, Jean-Philippe Antignac, Laetitia Shintu, Marianne Defernez, Mark Philo, Marie-Cécile Alexandre-Gouaubau, Marie-Josephe Amiot-Carlin, Mathilde Bossis, Mohamed N. Triba, Natali Stojilkovic, Nathalie Banzet, Roland Molinié, Romain Bott, Sophie Goulitquer, Stefano Caldarelli, and Douglas N. Rutledge. Can we trust untargeted metabolomics? Results of the metabo-ring initiative, a large-scale, multi-instrument interlaboratory study. *Metabolomics*, 11(4):807–821, 2015. doi: 10.1007/s11306-014-0740-0.

Ewy A. Mathé, Andrew D. Patterson, Majda Haznadar, Soumen K. Manna, Kristopher W. Krausz, Elise D. Bowman, Peter G. Shields, Jeffrey R. Idle, Philip B. Smith, Katsuhiro Anami, Dickran G. Kazandjian, Emmanuel Hatzakis, Frank J. Gonzalez, and Curtis C. Harris. Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. *Cancer Research*, 74(12):3259–3270, 2014. doi: 10.1158/0008-5472.CAN-14-0109.

Lisa M. Menacher, Liam J. Ward, Fredrik Heintz, Henrik Green, and Oleg Sysoev. LCMS-Net: Deep learning for raw high resolution mass spectrometry data applied to forensic cause-of-death screening. *Analytical Chemistry*, 98(9):6589–6597, 2026. doi: 10.1021/acs.analchem.5c05404.

Kevin M. Mendez, Stacey N. Reinke, and David I. Broadhurst. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics*, 15:150, 2019. doi: 10.1007/s11306-019-1612-4.

Owen D. Myers, Susan J. Sumner, Shuzhao Li, Stephen Barnes, and Xiuxia Du. Detailed investigation and comparison of the XCMS and MZmine 2 chromatogram construction and chromatographic peak detection methods for preprocessing mass spectrometry metabolomics data. *Analytical Chemistry*, 89(17):8689–8695, 2017. doi: 10.1021/acs.analchem.7b01069.

Gary J Patti, Oscar Yanes, and Gary Siuzdak. Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, 13(4):263–269, 2012.

Simon J Pelletier, Mickaël Leclercq, Florence Roux-Dalvai, Matthijs B de Geus, Shannon Leslie, Weiwei Wang, TuKiet T Lam, Angus C Nairn, Steven E Arnold, Becky C Carlyle, Frédéric Precioso, and Arnaud Droit. BERNN: Enhancing classification of liquid chromatography mass spectrometry data with batch effect removal neural networks. *Nature Communications*, 15(1):3777, 2024. doi: 10.1038/s41467-024-48177-5.

Michael Phillips, Renee N. Cataneo, Rany Condos, Gerald A. Ring Erickson, Joel Greenberg, Vincent La Bombardi, Muhammad I. Munawar, and Olaf Tietje. Volatile biomarkers of pulmonary tuberculosis in the breath. *Tuberculosis*, 87(1):44–52, 2007. doi: 10.1016/j.tube.2006.03.004.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660, 2017.

Rui Qiao, Ngoc Hieu Tran, Lei Xin, Xin Chen, Ming Li, Baozhen Shan, and Ali Ghodsi. Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nature Machine Intelligence*, 3:420–425, 2021. doi: 10.1038/s42256-021-00304-3.

Thomas P. Quinn and Ionas Erb. Interpretable log contrasts for the classification of health biomarkers: A new approach to balance selection. *mSystems*, 5(2):e00230–19, 2020. doi: 10.1128/mSystems.00230-19.

Habtom W. Ressom, Jun Feng Xiao, Leepika Tuli, Rency S. Varghese, Bin Zhou, Tsung-Heng Tsai, Mohammad R. Nezami Ranjbar, Yi Zhao, Jinlian Wang, Cristina Di Poto, Amrita K. Cheema, Mahlet G. Tadesse, Radoslav Goldman, and Kirti Shetty. Utilization of metabolomics to identify serum biomarkers for hepatocellular carcinoma in patients with liver cirrhosis. *Analytica Chimica Acta*, 743:90–100, 2012. doi: 10.1016/j.aca.2012.07.013.

Camille Roquencourt, Stanislas Grassin-Delyle, and Etienne A Thévenot. ptairMS: real-time processing and analysis of PTR-TOF-MS data for biomarker discovery in exhaled breath. *Bioinformatics*, 38(7):1930–1937, 2022.

Xiaotao Shen, Wei Shao, Chuchu Wang, Liang Liang, Songjie Chen, Sai Zhang, Mirabela Rusu, and Michael P. Snyder. Deep learning-based pseudo-mass spectrometry imaging analysis for precision medicine. *Briefings in Bioinformatics*, 23(5):bbac331, 2022. doi: 10.1093/bib/bbac331.

Colin A. Smith, Elizabeth J. Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787, 2006. doi: 10.1021/ac051437y.

Ralf Tautenhahn, Gary J. Patti, Duane Rinehart, and Gary Siuzdak. XCMS Online: a web-based platform to process untargeted metabolomic data. *Analytical Chemistry*, 84(11):5035–5039, 2012. doi: 10.1021/ac300698c.

Ruimin Wang, Miaoshan Lu, Shaowei An, Jinyin Wang, and Changbin Yu. 3D-MSNet: A point cloud-based deep learning model for untargeted feature detection and quantification in profile LC-HRMS data. *Bioinformatics*, 39(5):btad195, 2023. doi: 10.1093/bioinformatics/btad195.

Melih Yilmaz, William E. Fondrie, Wout Bittremieux, Carlo F. Melendez, Rowan Nelson, Varun Ananth, Sewoong Oh, and William Stafford Noble. Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nature Communications*, 15:6427, 2024. doi: 10.1038/s41467-024-49731-x.

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.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3391–3401. Curran Associates, Inc., 2017.

A Appendix

A.1 Dataset screening details

Candidate datasets required centroid MS1 measurements, reliable diagnostic labels, and acquisition metadata describing available collection times, sites, or instruments. Public datasets were additionally required to have appeared in a published machine-learning study. We used approximately 20 independent test subjects per class as a practical screening target.

A.2 Dataset distribution and class structure

Figure 4 visualizes the spectral profiles using t-SNE after total-intensity normalization and log1p transformation. The OOD datasets exhibit acquisition-domain differences across instruments, temporal batches, and cohorts, consistent with distribution shifts in the cross-batch evaluation settings (Figure 4(A)). For the ID datasets, the visualized class structure broadly aligns with classification performance across the evaluated methods (Figure 4(B)). Datasets on which all methods perform well show clearer separation between the two classes, whereas those with consistently lower performance exhibit extensive class overlap. These observations suggest two broad dataset groups: those with relatively distinct class structure and those with strongly interspersed classes. We hypothesize that differences in intrinsic data quality, including the strength of class-discriminative signals, may contribute to this distinction and help explain the performance differences across datasets.

A.3 Training-Time Data Augmentation

Table 5 provides the detailed formulations of the training-time augmentations used for PeakCloudNet.

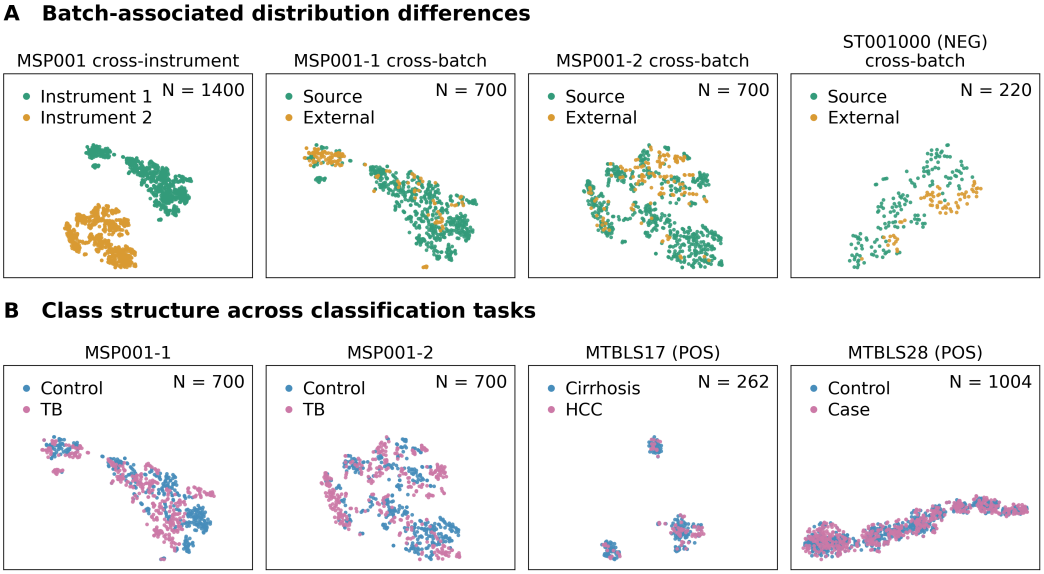

Figure 4: Acquisition-domain variation in OOD datasets and class structure in ID datasets, visualized using t-SNE after total-intensity normalization and log1p transformation. (A) OOD dataset distributions across instruments, temporal batches, and cohorts. (B) ID dataset class structure, showing clearer class separation in datasets with consistently high classification performance and extensive class overlap in those with consistently lower performance. Positive-ion data are shown for MTBLS17 and MTBLS28.

Table 5: Training-time augmentations for PeakCloudNet. Sample indices are omitted for readability.

Augmentation	**Formulation**
Low-intensity deletion	$d_j \sim \text{Bernoulli}(0.6)$, delete peak j if $I_j \le Q_{0.2}(\mathbf{I})$ and $d_j = 1$
m/z shift	$m'_j = m_j + s_j \Delta m_j$, $s_j \sim \mathcal{U}\{-k, \dots, k\}$, $\lvert s_j \Delta m_j \rvert \le 0.5\,\text{FWHM}_j$
Multiplicative intensity noise	$I'_j = I_j\, 2^{\epsilon_j}$, $\epsilon_j \sim \mathcal{N}(0,1)$
Baseline augmentation	$I'_j = I_j + a$, $a \sim \mathcal{U}(0, \ell)$
RT jitter	$r'_j = \text{clamp}\big(r_j + 0.002\, r^{\max} \eta_j, 0, r^{\max}\big)$, $\eta_j \sim \mathcal{N}(0,1)$

Notation. $Q_{0.2}(\mathbf{I})$ is the 20th percentile of valid peak intensities. Δm_j is the m/z perturbation step, k controls its discrete range, and FWHM_j is the full width at half maximum. ℓ is the sample-specific baseline bound, and $r^{\max}$ is the maximum retention time. RT jitter applies only to measurements containing RT.

A.4 Dataset Details and Environmental Confound Removal

MSP001 cohort and sample collection. According to the collection protocol documented by the collaborating clinical sites, the MSP001 cohort comprises 700 participants, including microbiologically confirmed pulmonary tuberculosis (TB) patients recruited from an outpatient clinic and healthy controls recruited over an overlapping period from an affiliated health examination center. All participants fasted for at least 10 hours and refrained from odoriferous foods, alcohol, and tobacco for 24 hours. After resting quietly in a controlled environment for at least 10 minutes to stabilize res-

piratory patterns, participants provided breath samples via standardized sampling bags using sterile, single-use mouthpieces. Triplicate breath samples were collected per participant alongside a concurrent ambient-air reference sample. Every sample was concurrently measured on two independent PTR-TOF-MS instruments under identical operational protocols, yielding paired cross-instrument profiles across longitudinal acquisition batches.

Environmental confound removal. To eliminate site-specific volatile contaminants between the clinical outpatient clinic and the health examination center, paired environmental ambient-air spectra were used to construct an instrument-specific peak exclusion mask prior to model training:

1. **Differential peak filtering:** Two-sided Mann–Whitney U tests with Benjamini–Hochberg correction were conducted on ambient-air peak intensities across collection venues. Peaks exhibiting significant site differences (adjusted $q < 0.1$) were flagged as differential environmental contaminants.
2. **Correlation proxy filtering:** Within each collection site, Pearson correlation coefficients were computed across environmental peaks. Any peak exhibiting $|r| > 0.75$ with a differential environmental peak or its neighboring spectral region was identified as an environmental proxy and added to the exclusion mask.
3. **Spectrum filtering:** For each breath spectrum, peaks within ± 200 ppm (matching instrument resolution) of an exclusion mask center were discarded.

Cross-instrument evaluations strictly applied the environmental exclusion set derived from the training source instrument, ensuring zero leakage of target hardware or test-site background profiles.

A.5 Sample-Level Peak-CAM Visualizations

Figures 5–11 present sample-level Peak-CAM visualizations for four TB samples and three healthy samples. Each spectrum plots $\log_{10}(1 + I)$ against the mass-to-charge ratio (m/z), where I denotes peak intensity. Dark gray bars highlight the top 10% of peaks ranked by absolute Peak-CAM score, while light gray bars show the remaining peaks. All panels use the same axis limits. The examples include a TB sample predicted as healthy (Figure 8), alongside correctly classified TB and healthy samples.

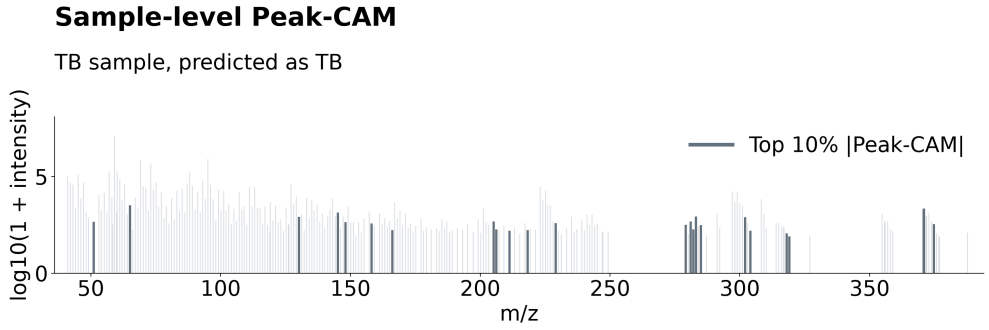

Figure 5: Sample-level Peak-CAM visualization for TB sample `1_1207`, correctly predicted as TB. Dark gray bars indicate the top 10% of peaks ranked by absolute Peak-CAM score.

A.6 Local peak perturbation for baseline classifiers

We used the original trained models, test splits, preprocessing, and proxy-peak filtering, without retraining or out-of-fold analysis. Spectra were represented by feature table.

For RF, we computed interventional TreeSHAP contributions to TB probability using the preprocessed training set as background. For PLS-DA, PCA+LR, contributions were $\phi_j = \beta_j(z_j - \mu_j)$, where z_j is the preprocessed feature, μ_j its training mean, and β_j the effective coefficient. These

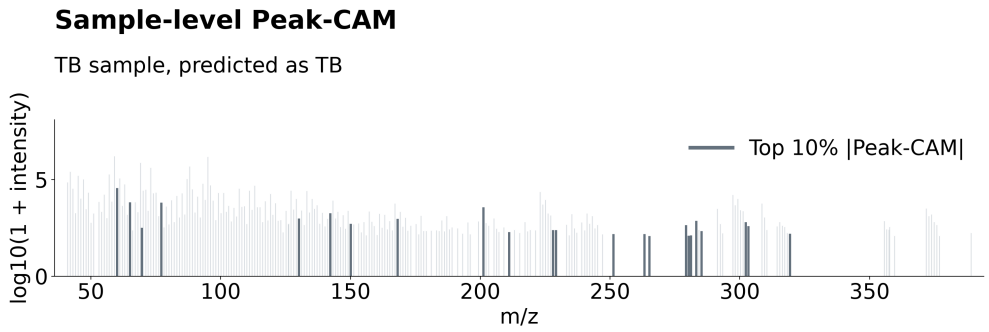

Figure 6: Sample-level Peak-CAM visualization for TB sample `1_1459`, correctly predicted as TB. The plotting conventions follow Figure 5.

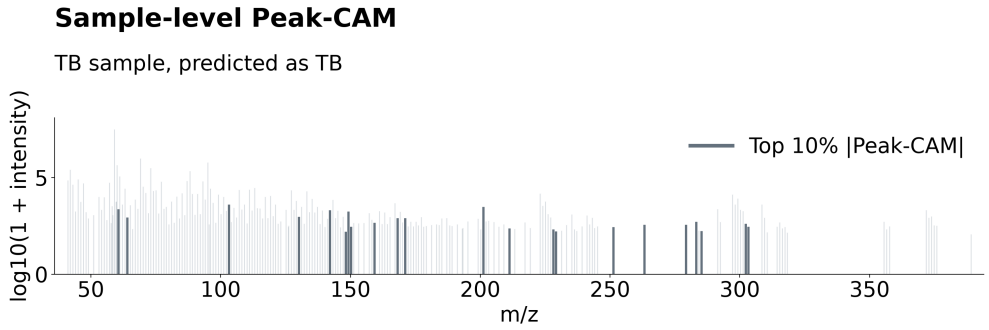

Figure 7: Sample-level Peak-CAM visualization for TB sample `1_1560`, correctly predicted as TB. The plotting conventions follow Figure 5.

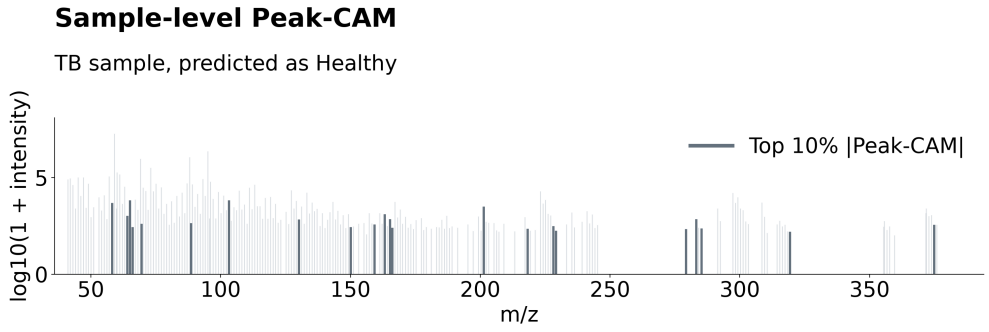

Figure 8: Sample-level Peak-CAM visualization for TB sample `1_1589`, incorrectly predicted as healthy. The plotting conventions follow Figure 5.

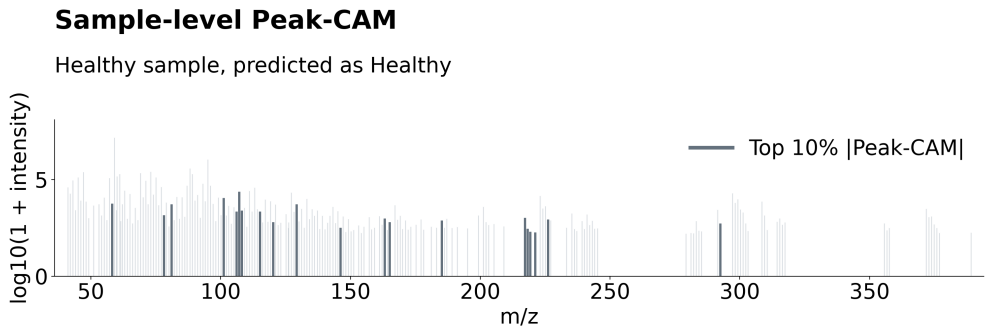

Figure 9: Sample-level Peak-CAM visualization for healthy sample `0_1155`, correctly predicted as healthy. The plotting conventions follow Figure 5.

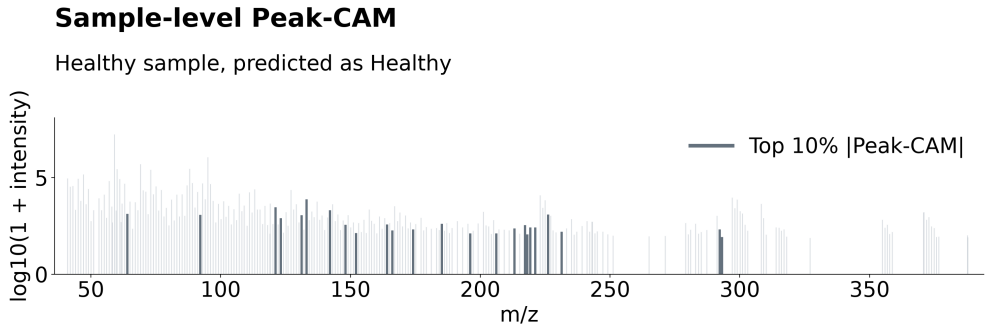

Figure 10: Sample-level Peak-CAM visualization for healthy sample `0_820`, correctly predicted as healthy. The plotting conventions follow Figure 5.

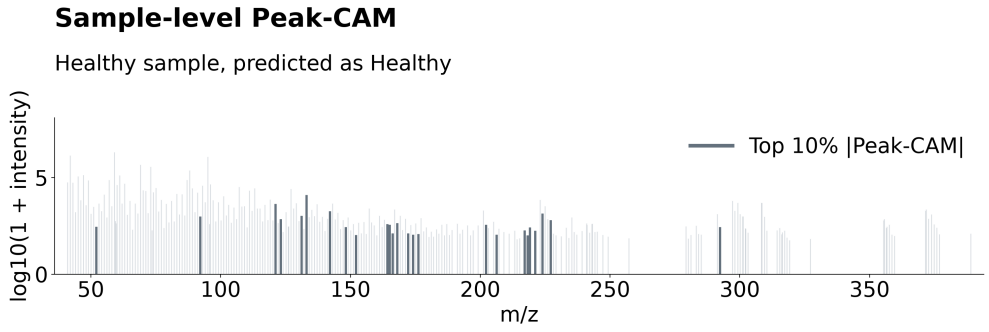

Figure 11: Sample-level Peak-CAM visualization for healthy sample `0_856`, correctly predicted as healthy. The plotting conventions follow Figure 5.

explain the pre-sigmoid PLS score, LR logit. PCA+LR coefficients were mapped back through the PCA components. For the SVM and MLP, attribution was the change in TB probability after setting one occupied bin to zero before preprocessing.

Each occupied bin corresponded to one physical peak in the test spectra. For each sample and seed, we ranked peaks by absolute attribution and removed the top $p\%$, for $p = 0, 1, \ldots, 15$, rounding the number removed upward. Rankings were fixed from the full input. After deletion, we rebuilt the binned input and applied the frozen preprocessing, with empty bins represented by zero before transformation.

Curves show the mean test AUROC across seeds 42, 43, and 44, with each model's full-input AUROC as its baseline. Shading represents pointwise 95% percentile intervals from 1,000 subject-bootstrap replicates stratified by disease class, using the same resampled subjects across seeds.

A.7 Cohort-level attribution regions

Global regions are constructed from five-fold out-of-fold attributions on the 560-subject MSP001-1 training pool. We average each peak's absolute normalized attribution across three seeds and nominate the top 20% within each subject, rounding upward. Nominated peaks are ordered by m/z and grouped while all members remain within 200 ppm of a shared center, defined as the median observed m/z when feasible and otherwise as the midpoint of the feasible interval. The resulting 832 regions are ranked by the fraction of subjects nominating them, counting each subject once per region and breaking ties by increasing center m/z; nomination frequencies are also reported by class.

A.8 Biological explainability on MTBLS17 and ST001000

Biological explainability on MTBLS17. We further extend the explainability analysis to MTBLS17, which compares serum samples from patients with hepatocellular carcinoma (HCC) and cirrhotic controls. The global peaks identified by PeakCloudNet include features whose m/z values match ions reported for glycocholic acid (GCA), a GCA fragment, and taurochenodeoxycholate (TCDCA) by Ressom et al. (2012). In particular, the GCA-matching feature ranks 19th among the negative-mode global peaks. That study reported lower serum levels of GCA and TCDCA in HCC than in cirrhosis and discussed their association with altered bile acid metabolism and transport. Additional matches include features tentatively assigned in the original study to lysophosphatidylcholine and lysophosphatidylethanolamine species, including LysoPC 15:0, LysoPC 22:6, LysoPE 20:4, and LysoPE 20:3. These lipid classes were discussed in relation to membrane phospholipid metabolism, cell proliferation, invasion, and inflammation. Together, these mass-based correspondences provide biological context for the model's selected features through bile acid and phospholipid metabolism, while remaining putative assignments pending structural confirmation.

Biological explainability on ST001000. We further extend the explainability analysis to the ST001000 dataset. The global peaks identified by PeakCloudNet include features whose m/z values match the $[\mathrm{M}-\mathrm{H}]^-$ ions of cholate, lactate, and pantothenate reported by Franzosa et al. (2019). Their study found elevated fecal cholate in Crohn's disease, consistent with altered microbial bile acid transformation, and reported increased lactate and decreased pantothenate in both Crohn's disease and ulcerative colitis. These changes suggest perturbed microbial metabolic activity, as lactate is produced by gut bacteria, while pantothenate is a precursor of coenzyme A involved in fatty acid metabolism and can also be supplied by the gut microbiota. These mass-based correspondences remain putative assignments pending structural confirmation, but indicate that PeakCloudNet highlights peaks consistent with previously reported IBD-associated metabolic alterations.

A.9 Candidate molecular-formula retrieval

A.9.1 Query and molecular-formula representations

We use a pretrained spectral matching model to associate attribution-selected peaks with candidate molecular formulas. A query is a set of neutral-mass coordinates $Q = \{(x_j, \delta_j, w_j)\}_{j=1}^{J}$, where δ_j is

an effective mass uncertainty and the nonnegative intensity weights satisfy $\sum_j w_j = 1$. Single-peak retrieval corresponds to $J = 1$ and $w_1 = 1$.

The model uses $F = 512$ fixed, logarithmically spaced angular frequencies ω_k, with wavelengths ranging from 10^{-3} to 3850 Da. Its positive spectral amplitudes are $a = \exp(W\theta)$, where $\theta \in \mathbb{R}^{16}$ contains learned interpolation knots and $W \in \mathbb{R}^{F\times 16}$ is a fixed linear-interpolation matrix. The complex query coordinates are

$$u_k(Q) = \sum_{j=1}^{J} w_j \exp\left(-\frac{\omega_k^2 \delta_j^2}{2}\right) \exp(\mathrm{i}\omega_k x_j), \qquad \psi(Q) = [a \odot \operatorname{Re} u(Q);\, a \odot \operatorname{Im} u(Q)]\,. \tag{4}$$

The Gaussian damping attenuates frequencies that resolve mass differences below the query uncertainty. Concatenating the real and imaginary coordinates gives a 1024-dimensional representation.

For a formula f, let n_e be the count of element e and M_f its theoretical monoisotopic mass. The characteristic function of its isotope distribution is

$$C_{f,k} = \exp(\mathrm{i}\omega_k M_f) \prod_e \left(\sum_r p_{e,r} \exp(\mathrm{i}\omega_k \Delta m_{e,r})\right)^{n_e}, \tag{5}$$

where $p_{e,r}$ and $\Delta m_{e,r}$ are fixed isotope probabilities and mass offsets. This factorization computes the distribution in the frequency domain without enumerating every isotopologue. A learned composition gate adjusts its frequency contributions:

$$\begin{aligned} g(f) &= \operatorname{softplus}(G(\log(1+n_f)) + 0.5413)\,, \\ \Phi(f) &= [a \odot g(f) \odot \operatorname{Re} C_f;\, a \odot g(f) \odot \operatorname{Im} C_f]\,, \\ s(Q,f) &= \frac{\psi(Q)^\top \Phi(f)}{\tau} + b(f), \qquad b(f) = B(\log(1+n_f)). \end{aligned} \tag{6}$$

Here G and B are multilayer perceptrons with dimensions $13 \to 128 \to 512$ and $13 \to 128 \to 1$, respectively, each using a GELU hidden activation. The 13 input coordinates represent element counts. The model has 69,777 parameters, with one additional learned temperature parameter stored separately. Frequencies and isotope constants are fixed; spectral amplitudes, composition gates, composition biases, and temperature are learned. The deployed temperature is $\tau = 0.152627$. Retrieval uses the unnormalized inner product in Equation 6.

A.9.2 Pretraining data and query construction

The released training pipeline constructs a formula-labeled corpus from MassBank records. It retains supported adducts, formulas with theoretical masses between 10 and 1500 Da, and precursor measurements within 200 ppm of their theoretical ion masses. The corpus builder also requires at least three eligible spectral peaks. A deterministic formula-based partition assigns approximately 10% of formulas to validation, keeping formulas disjoint between the training and validation partitions. The spectral retriever uses these formula labels and adduct-corrected precursor masses to synthesize MS1 isotope queries; the stored MS2 fragment intensities are not inputs to its training objective.

For each training record, the generator constructs a theoretical isotope envelope through $M + 8$ and places it relative to the adduct-corrected measured precursor mass. It samples an instrument resolution curve, merges adjacent isotope peaks according to the corresponding full width at half maximum, and adds a shared mass shift and independent centroid residuals. The effective uncertainty is

$$\delta = \sqrt{\sigma_{\mathrm{err}}^2 + \left(\frac{m}{2R(m)}\right)^2}, \tag{7}$$

where m is the query reference mass and σ_{err} combines the simulated error components. Instrument families are sampled with conditional probabilities of 0.42, 0.37, 0.16, and 0.05 for Orbitrap, time-of-flight, quadrupole, and FT-ICR settings, respectively; a separate 5% branch samples additional resolution settings. The calibration scale is restricted to 0.2–10 ppm, with a family-dependent factor before clipping. Twenty percent of queries receive an additional shared error with a scale between 1.5 and 6 mDa.

Isotope-intensity perturbations and contaminating peaks are sampled from the supplied empirical noise profile, which records 544 calibration observations. Relative detection thresholds range from 0.003 to 0.05, and an additional missing-$M+2$ event has probability 0.254. Two independent augmentation decisions, each with probability 0.15, retain only the monoisotopic component or randomly discard higher isotope components. Contaminating peaks can subsequently be added. Retained intensities are normalized to sum to one. These augmentations expose the retriever to incomplete isotope patterns, including single-peak inputs.

A.9.3 Candidate sampling and training objective

Training candidates comprise the correct formulas, mass-window candidates, and up to 2048 randomly selected training records contributing additional candidate formulas per batch. The window centres are $x_1 - k \times 1.00336\,\mathrm{Da}$ for $k \in \{0, \ldots, 4\}$. The unshifted half-width is $4\delta + 0.01\,\mathrm{Da}$; shifted windows add $0.01 + 0.005k\,\mathrm{Da}$. Their respective candidate budgets are 800, 200, 120, 60, and 60. Candidate formulas are deduplicated across the batch.

The objective identifies the correct formula within this shared candidate set $\mathcal{C}$:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\{s(Q_i, f_i) + c_{i,f_i}\}}{\sum_{f\in\mathcal{C}} \exp\{s(Q_i, f) + c_{i,f}\}}. \tag{8}$$

For sampled window candidates, the implementation adds a correction of the form $c = \log(\widehat{N}_{\mathrm{window}}/N_{\mathrm{sampled}})$ when the estimated number of legal candidates exceeds the sampled count; entries without such a correction use zero. This is a one-hot cross-entropy objective with a sampled-candidate correction. The released checkpoint is trained for 30 epochs with batch size 384, AdamW, initial learning rate 3×10^{-4}, cosine learning-rate decay, and gradient-norm clipping at 5. The optimizer uses its default weight decay of 0.01. We use the final checkpoint, indexed as epoch 29. The training script sets the PyTorch and Python random seeds to 0 and 1234, respectively.

A.9.4 Retrieval evaluation protocol

The released evaluation scripts operate on measured MassBank MS1 isotope patterns. Dataset construction locates the monoisotopic peak within $\max(10\,\mathrm{ppm}, 0.002\,\mathrm{Da})$ of the theoretical ion mass and requires an $M+1$ peak within $\max(15\,\mathrm{ppm}, 0.003\,\mathrm{Da})$. The evaluator samples up to 600 eligible records with seed 999. High-, medium-, and low-resolution groups use $(R, p) = (70000, 5), (30000, 10), (2000, 30)$, where p is the assumed accuracy in ppm, giving

$$\delta_{\mathrm{eval}} = \sqrt{(p\,10^{-6}x)^2 + (x/(2R))^2}. \tag{9}$$

Candidate masses lie within $x \pm (4\delta_{\mathrm{eval}} + 0.005\,\mathrm{Da})$. Up to 32,000 candidates are sampled before explicitly adding the reference formula. Records with fewer than two candidates are skipped. Thus, the reported ranking metric conditions on the reference formula being present in the candidate set.

The isotope-selected input includes the monoisotopic peak and peaks within 0.02 Da of offsets $k \times 1.00336\,\mathrm{Da}$ for $k = 1, \ldots, 8$. When a stored isotope ratio is available without a selected local peak, the evaluator uses its nominal isotope position. The input ablation compares the full selected pattern, M with $M+1$, M with $M+2$, and M alone. For each query, the reference rank is one plus the number of candidates with a strictly greater score; top-k accuracy is the fraction with rank at most k. The single-peak condition matches the input structure used in our attribution analysis. The scripts also stratify results by whether a formula occurs anywhere in the source corpus, including its validation partition; this stratification denotes corpus membership rather than training-set membership.

A.9.5 Reported evaluation results and model analysis

The release materials report retrieval accuracy for a single-peak reference version, a summary of the learned spectral allocation, and empirical isotope-intensity statistics. These results characterize ranking performance, representation properties, and the measurement variation used to construct training queries, respectively.

Table 6: Release-reported molecular-formula retrieval accuracy. The entry is the single-peak reference result documented in the evaluation code, not a measurement on the attribution-selected peaks.

Reported model	Top-1 (%)	Top-5 (%)
Single-peak reference	81.2	92.5

Table 7: Learned spectral allocation reported in the release design document. Wavelength is $2\pi/\omega$. The listed bands are a partial summary and their reported shares do not sum to 100%.

Wavelength range	Reported share (%)
Above 10 Da	41.5
0.1–1 Da	18.0
10–100 mDa	14.5
1–10 mDa	11.9
Below 1 mDa	0.0

Molecular-formula ranking. The release evaluation code reports top-1 and top-5 accuracies of 81.2% and 92.5%, respectively, for its single-peak reference version (Table 6). The top-five list therefore increases reference-formula recovery by 11.3 percentage points relative to the first-ranked candidate. These are historical results reported with the release. Their checkpoint identity and evaluated sample count are not specified in the available record, so they are attributed to the single-peak reference rather than to a new evaluation of the deployed cluster checkpoint.

Learned spectral allocation. Table 7 reproduces the allocation summary in the release design document. The reported share is largest at wavelengths above 10 Da, while the 1–10 mDa band retains an 11.9% share. The reported bands thus span coarse and fine mass scales. The document reports a zero share below 1 mDa. This is a partial band summary: the 1–10 Da interval and the exact normalization used to compute the shares are not specified, so the entries are retained as reported rather than renormalized.

Empirical isotope-intensity variation. The release design document summarizes deviations between measured and theoretical isotope intensities (Table 8). The reported typical deviation increases from $\pm 14\%$ for $M+1$ to $\pm 48\%$ for $M+4$, while the reported frequency of more-than-twofold deviations increases from 2.8% to 40.0%. These observations motivate the use of empirical intensity perturbations during training, particularly for higher isotope orders. We retain the source's term "typical deviation" because it does not specify whether this statistic is a quantile or a dispersion measure. These are measurement statistics underlying the training noise model, rather than formula-ranking accuracies.

Table 8: Empirical isotope-intensity statistics reported in the release design document. Deviations compare measured with theoretical isotope intensity; the column descriptions follow the source.

Isotope component	Typical deviation	More-than-twofold deviation (%)
$M+1$	$\pm 14\%$	2.8
$M+2$	$\pm 29\%$	14.2
$M+3$	$\pm 33\%$	18.8
$M+4$	$\pm 48\%$	40.0

A.9.6 Application to attribution-selected peaks

We apply the frozen retriever to peaks in the selected mass regions and selected peak modules of the PTR-TOF-MS attribution analysis. Region selection is fixed before annotating the test and transfer samples. Annotation uses the peak records from the classification run with random seed 42. For an

observed mass-to-charge value z, we assume a singly protonated ion and set

$$x = z - 1.007276\,\mathrm{Da}, \qquad \delta = \frac{z}{2 \times 2500}, \qquad \mathcal{C}(x, \delta) = \{f : |M_f - x| \leq 4\delta + 0.005\,\mathrm{Da}\}\,. \tag{10}$$

The uncertainty follows the $R = 2500$ resolution model used for these peaks. No measured mass-accuracy term is available for this analysis.

Candidates are generated by meet-in-the-middle enumeration over C, H, N, O, P, S, F, Cl, Br, and I, followed by necessary valence-feasibility checks. The respective count limits for C, N, O, P, S, F, Cl, Br, and I are 120, 20, 45, 8, 8, 24, 10, 6, and 6; hydrogen counts are bounded by the mass range. The enumerator enforces a nonnegative integral ring-and-double-bond equivalent and the total-valence connectivity condition. It enumerates all candidates when feasible and samples larger windows, with a maximum of 32,000 scored candidates and a deterministic query-specific seed. No reference formula is inserted during annotation.

Each selected peak is encoded as a single-peak query and scored with Equation 6. We retain the ten highest-scoring formulas together with their theoretical masses and signed mass errors, $x - M_f$. Identical (z, δ) queries share a cached ranking. The retriever receives the peak mass and uncertainty independently of the classification network; peak attribution scores, class labels, and classification-network hidden states do not enter its matching score. The resulting lists provide candidate molecular formulas under the stated ion assumption, rather than unique structural assignments.