Which Distance Metric Predicts Cross-Instrument MS/MS Degradation?

A Systematic Comparison on MassBank Fingerprint Embeddings

Elliot Tower

July 2026

1 Question

When deploying a spectral library search model (such as LSM-MS2) to a new mass spectrometer, performance degrades. Can we predict *how much* degradation to expect from the embeddings alone, without running the model on the target instrument?

We compare six distance metrics between instrument-specific embedding distributions and measure which one best correlates with actual cross-instrument degradation on a compound identification task.

2 Setup

Data. 2,140 MS/MS spectra from MassBank, spanning 10 instrument types across three ionization families: ESI (7 instruments), EI (2), and MALDI (1). This yields 45 pairwise instrument comparisons.

Embedding. Binary fingerprints: m/z values binned into 2048 bins across 0–2000 Da. Each spectrum becomes a 2048-dimensional binary vector. This is a deliberately simple baseline; a learned embedding (e.g. LSM-MS2) should improve all metrics.

Degradation metrics (what we want to predict).

- **Classifier degradation**: Train a logistic regression to distinguish instrument A from instrument B using one half of each, test on the held-out half. Degradation = internal CV AUC − external AUC.
- **Spectral matching degradation**: For compounds appearing in both instruments, compute within-instrument hit@1 (nearest-neighbor has same compound label) minus cross-instrument hit@1. Higher = worse cross-instrument retrieval.

Distance metrics (what we compare).

1. **Geodesic distance** on the Grassmannian $\mathrm{Gr}(k, 2048)$: $\|\boldsymbol{\theta}\|_2$ where $\theta_i = \arccos(\sigma_i)$ are principal angles between top-k PCA subspaces.
2. **Centroid distance**: $\|\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B\|_2$.

3. **Domain classifier AUC**: 5-fold CV AUC of a logistic regression trained to distinguish the two instruments from their embeddings.

4. **MMD** (Maximum Mean Discrepancy): $\hat{D}_k^2$ with RBF kernel, $\gamma = 1/d$. The standard two-sample test from domain adaptation (Gretton et al., 2012).

5. **Sliced Wasserstein distance**: Average 1D Wasserstein distance over 100 random projections.

6. **Proxy $\mathcal{A}$-distance**: $d_{\mathcal{A}} = 2(1 - 2\epsilon)$ where ϵ is the error rate of a linear SVM distinguishing the two domains (Ben-David et al., 2007). The theoretical bound on domain adaptation loss.

3 Results

Table 1: Spearman ρ between each distance metric and cross-instrument degradation. All 45 instrument pairs (10 instruments, binary fingerprint embeddings). Starred entries reach $p < 0.05$; bootstrap 95% CIs in brackets.

	Clf degradation		Match degradation	
Metric	ρ [95% CI]	p	ρ [95% CI]	p
Geodesic (k=10)	-0.149 [−.44, +.18]	.329	-0.040 [−.35, +.25]	.793
Centroid distance	-0.217 [−.54, +.11]	.153	$+0.417$ [+.10, +.66]	.004**
Domain classifier AUC	-0.302 [−.62, +.04]	.044*	$+0.508$ [+.20, +.72]	<.001***
MMD (RBF)	-0.217 [−.54, +.11]	.153	$+0.417$ [+.10, +.66]	.004**
Sliced Wasserstein	-0.247 [−.56, +.09]	.101	$+0.422$ [+.09, +.66]	.004**
Proxy $\mathcal{A}$-distance	-0.303 [−.61, +.02]	.043*	$+0.317$ [+.01, +.57]	.034*

At first glance, five of six metrics significantly predict match degradation ($\rho = 0.32$–0.51, $p < 0.05$). Domain classifier AUC is the strongest predictor ($\rho = 0.51$, $p < 0.001$). Geodesic distance on the Grassmannian predicts neither outcome.

3.1 Confound: ionization family

Twelve of the 45 pairs involve EI-B instruments, which share zero compounds with ESI instruments. These pairs have match degradation = 1.0 by construction (no shared compounds means hit@1 drops to zero). Removing these 12 fallback pairs collapses all correlations (Table **??**).

Within the ESI family alone (21 pairs), no metric reaches even $p < 0.25$. The apparent signal in Table **??** is driven by the ionization-family boundary, not by gradual distributional differences that a metric could usefully rank.

3.2 Permutation null

All 45 geodesic distances exceed the permutation null ($p = 0.000$, 50 permutations per pair). The PCA subspace differences are real, even though they fail to predict degradation.

Table 2: Same analysis excluding 12 pairs with match degradation = 1.0 (EI-B instruments sharing no compounds with the partner). $n = 33$.

	Clf degradation		Match degradation	
Metric	ρ	p	ρ	p
Geodesic (k=10)	−0.154	.391	−0.297	.094
Centroid distance	+0.028	.875	−0.162	.367
Domain classifier AUC	−0.075	.680	−0.038	.832
MMD (RBF)	+0.028	.875	−0.162	.367
Sliced Wasserstein	+0.001	.994	−0.142	.432
Proxy $\mathcal{A}$-distance	−0.090	.619	−0.093	.606

4 Interpretation

Binary fingerprints map spectra into a high-dimensional space where all instruments land at near-maximal Grassmannian distance (range 4.22–4.80 out of a theoretical max ∼4.97). The geodesic has almost no dynamic range, so it cannot rank instrument pairs by similarity.

The discriminative metrics (domain classifier AUC, proxy $\mathcal{A}$-distance) reach marginal significance against classifier degradation ($p \approx 0.04$) and apparent strong significance against match degradation. The controlled analysis reveals this is an artifact: EI-B instruments share no compounds with ESI instruments, producing match degradation of exactly 1.0 by construction. These 12 pairs create a clean separation between "high distance, high degradation" and "low distance, low degradation" that any metric can fit. Once removed, nothing predicts degradation.

The failure mode is clear. Binary fingerprints do not encode enough spectral structure for any distance metric—distributional, geometric, or discriminative—to track gradual cross-instrument performance loss. The metrics can detect whether two instruments use the same ionization method (a coarse binary distinction), but cannot rank instruments within a family.

5 The pitch

This analysis establishes a benchmark for the question: "given a new instrument, how much should I expect my spectral search model to degrade?"

We compared six standard domain-adaptation distance metrics on binary fingerprint embeddings across 45 instrument pairs from MassBank. None predicts within-family degradation. The bottleneck is the embedding, not the metric.

The hypothesis: LSM-MS2 embeddings, trained on 1.8M spectra with learned spectral similarity, should produce distributions whose inter-instrument distances track real performance loss. The metric comparison framework is ready; what it needs is a better embedding.

Testing this requires running the same 2,140 MassBank spectra through LSM-MS2 inference and recomputing Tables **??**–**??**.