
Separability Is Not Alignability: Representations for Spectrum-to-Molecule Retrieval

Karol Musielinski^{1,2} Fanurs Chi-En Teh¹ Takahiko Koyama³ Artur Dubrawski¹

¹Auton Lab, Carnegie Mellon University

²University of Warsaw

³Keio University

{kmusieli,cteh,awd}@cs.cmu.edu

Abstract

Identifying a small molecule from its tandem mass spectrum is limited by paired data. Foundation models offer a way around this scarcity: pretrain a spectrum encoder on unlabeled spectra and a molecule encoder on a large corpus of structures, then align the two frozen representations, as in vision-language and as the Platonic Representation Hypothesis suggests. We audit this route on MassSpecGym with spectrum encoders retrained free of test-fold contamination. Alignment needs a metric relating distinct molecules, but the objectives that produce spectrum encoders never see molecular structure: we prove that some predictive objectives fix molecular identity without fixing that metric. Moreover, we find that a candidate-based objective rewards a spectrum-independent prior over the benchmark’s decoys, which we isolate. What ultimately addresses data scarcity is not architectural: simulated spectra surpass frozen geometry.

 [autonlab/msms-retrieval](https://github.com/autonlab/msms-retrieval)

1 Introduction

Tandem mass spectrometry (MS/MS) identifies a small molecule by fragmenting it and measuring the resulting fragment ions. *Retrieval* casts identification as ranking candidate structures against a query spectrum [5]. To overcome the small size of spectral libraries, Bushuiev et al. [6] developed a spectrum foundation model, DreaMS, pretrained on millions of unlabeled spectra, which Krzakala et al. [30] subsequently proposed to align with a molecule foundation model by learning a projection into a shared embedding space.

This rationale is borrowed from vision-language, where aligning frozen unimodal encoders rivals joint pretraining [67, 52]. More broadly, it echoes the Platonic Representation Hypothesis [24]. Kalia et al. [25] state outright that molecules and their spectra are “views of the same data.” These parallels motivate a closer examination of the assumptions behind transferring frozen-encoder alignment to molecule-spectrum retrieval, and of where those assumptions may break down.

Contributions. We audit this pipeline on MassSpecGym with spectrum encoders retrained free of test-fold contamination. **(i)** Satisfying the common assumptions is difficult in practice; frozen alignment instead relies on separability and metric compatibility (§2). **(ii)** We introduce MS-JEPA, a cross-condition predictive spectrum encoder, and prove that such objectives fix molecular identity without fixing the metric (§3). **(iii)** Measured, the two properties come apart: MS-JEPA separates molecules most strongly, yet retrieves worst when frozen (§4). **(iv)** The candidate objective rewards a spectrum-independent prior over the benchmark’s decoys, which a spectrum-blind control isolates (§5). **(v)** End-to-end training makes spectrum pretraining nearly irrelevant, while simulated spectra of 5.9M molecules double a fresh encoder’s MRR (§6).

2 The alignment premise

Let $\mathcal{M}$ be molecular structures, $\mathcal{A}$ acquisition conditions (adduct, instrument, and collision energy), and $\mathcal{S}$ spectra. For $m \in \mathcal{M}$ and $a \in \mathcal{A}$, a measurement is $s \sim p(\cdot \mid m, a)$. Retrieval ranks a candidate set $\mathcal{K}(s) \subset \mathcal{M}$ containing the true molecule m^* . Frozen alignment fixes $f : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^{d_s}$ and $g : \mathcal{M} \rightarrow \mathbb{R}^{d_m}$, learns adapters u_θ, v_θ into $\mathbb{R}^d$, and assigns the score

$$\rho_\theta(s, m) = \tau^{-1} \langle \bar{u}_\theta(f(s, a)), \bar{v}_\theta(g(m)) \rangle, \quad \bar{x} \equiv x / \|x\|_2. \quad (1)$$

What is borrowed. The Platonic Representation Hypothesis offers one explanation for why the vision-language recipe works [24]. It treats a representation as the kernel $K_f(x, x') = \langle f(x), f(x') \rangle$ it induces and claims that the kernels of different modalities converge as models scale, with convergence measured by neighborhood agreement (mutual- k NN), because coordinates need not coincide. If the kernels already agree, a *low-complexity adapter* should suffice.

What it assumes. The hypothesis argues from a latent world state z observed through modality-specific maps $x = \text{obs}_x(z)$ and $y = \text{obs}_y(z)$, taken to be deterministic and bijective. Multitask pressure, capacity, and simplicity bias then push every modality’s kernel toward the kernel of z . However, the authors state that *lossy* or *stochastic* observations lie outside the argument, and mutual- k NN itself becomes ambiguous under non-bijective correspondence [48, 27]. If we assume the world state is the molecule itself, then the spectrum side departs from a bijection in two ways. First, one molecule can yield many spectra, as the measurement itself is a condition-dependent stochastic draw. Second, some isomers give near-identical spectra [9, 58].

What kernel agreement means here. To compare the kernels, we lift the molecule kernel to spectra, $K_g(s, s') = \langle g(m(s)), g(m(s')) \rangle$ with $m(s)$ the molecule that produced s . Write $\Sigma_m \subset \mathbb{R}^{d_s}$ for the support of normalized spectrum embeddings $f(s, a)$ over $s \sim p(\cdot \mid m, a)$. For unit-normalized features, kernel error at most ε forces $\text{diam}(\Sigma_m) \leq \sqrt{2\varepsilon}$, and exact agreement collapses every Σ_m to a point (Prop. 2; App. C). This is a stringent requirement for a spectrum encoder trained on variable, condition-dependent measurements. Indeed, across audited encoders, the mean squared distance between same-molecule embeddings remains 0.14–0.51 (Table 3; App. E).

Two ingredients. Such collapse is not necessary for retrieval: the correct molecule need only score above the other candidates after adaptation. Successful frozen alignment with limited paired data depends on two properties: *separability*, meaning that spectrum embeddings distinguish molecular identities, and *metric compatibility*, meaning that their between-molecule geometry is already similar to that of the molecule tower. Clustering spectra by molecular identity may facilitate alignment: in vision-language, Zhang et al. [67] find that clustering quality strongly correlates with retrieval performance after lightweight alignment. Yet separability alone does not constrain the arrangement of different molecules and therefore cannot guarantee metric compatibility.

3 MS-JEPA: learning separable spectrum representations

To investigate whether stronger separability supports better retrieval performance, we train an encoder to capture molecular information shared across measurements under different acquisition conditions.

Objective. Molecular identity is invariant across acquisition conditions, and under fixed conditions a molecule’s fragmentation pattern is highly reproducible [32, 55, 3]. We therefore predict a spectrum’s representation from replicate spectra of the same molecule and the target condition [1]. With online encoder f_θ , a target encoder f_ξ whose weights are an exponential moving average of θ [19], a stop-gradient sg, a context aggregator Agg, and a condition-aware predictor h_ψ ,

$$\mathcal{L}_{\text{JEPA}} = \mathbb{E}_m \mathbb{E}_{(C, t)} \left[1 - \cos(h_\psi(\text{Agg}_{i \in C}(f_\theta(s_i, a_i), a_i), \text{sg}[f_\xi(s_t, a_t)]) \right]. \quad (2)$$

Ghaemi et al. [13] showed that this family of models learns invariant and equivariant representations from sequences of views and actions. Here the action is the change of acquisition condition.

Architecture. f_θ is a transformer over peak tokens with FiLM conditioning [39]: with $e(a)$ an embedding of the ion condition (adduct, charge, polarity), block ℓ scales its attention and MLP residual updates by gains $1 + \gamma_\ell(e(a))$. The representation is a learned [CLS] token. The predictor h_ψ is a shallow MLP whose hidden activations are FiLM-modulated by the full target condition. Each training episode samples one molecule, its target spectrum, and up to three context spectra.

The EMA teacher provides slowly evolving prediction targets to stabilize training. In principle, the aggregator combines the representations of multiple context spectra into a molecule-level representation, then h_ψ predicts the condition-dependent features. We use a cosine loss because it cannot be reduced by simply shrinking embedding norms. Training details are in App. B.3.

Limitations. The objective compares spectra of the same molecule only, without explicitly constraining the arrangement of different molecules. In particular, it provides no direct incentive for their geometry to agree with that of the molecule tower.

Proposition 1 (Identity without metric). *Let $L(f) = \inf_h L(f, h)$ over measurable heads h , where $L(f, h)$ is the expectation of a function of the cosine between the head’s prediction from a same-molecule context $\{(f(s_i, a_i), a_i)\}_{i \in C}$ with target condition a_t and the target $f(s_t, a_t)$.*

For a finite or countable molecule family, if the supports $\{\Sigma_m\}$ are pairwise disjoint and R acts on each Σ_m as an orthogonal map Q_m with the images again pairwise disjoint, then $L(R \circ f) = L(f)$.

Moreover, if molecular identity is determined by (s, a) , then for any molecule tower with distinct unit-normalized vectors in dimension $d \geq 2$ and candidate lists determined by the true molecule, each containing at least one distractor, some such pair f , R ranks the true molecule first for every query under direct cosine retrieval with f and never first with $R \circ f$.

Proof in App. C. The proposition idealizes the objective (2) with $f = f_\theta = f_\xi$. It shows that, even under perfect separability, any preference for a geometry compatible with g must arise from additional inductive biases. DreaMS’s contrastive loss falls outside this class, but likewise receives no direct supervision from g .

4 Against separability as a sufficient condition

To test the expectations outlined in §2, we measure both separability and cross-modal agreement on the challenging MCES split of MassSpecGym [5], which tests generalization to structurally distinct molecules. We compare three frozen encoders: DreaMS after self-supervised pretraining (masked-peak prediction), its version after contrastive fine-tuning, and MS-JEPA, whose [CLS] token and aggregated representation we probe separately. We reproduce both DreaMS stages to avoid test-spectra contamination from the released model (App. B.6).

Evaluation framework. Within-modality separability is scored by the silhouette [44] and by k NN purity, normalized by its ceiling. Cross-modality agreement is measured by mutual- k NN between molecule-level spectrum centroids and a frozen molecule tower, following Huh et al. [24] (App. E). Agreement among molecule encoders ranges from 0.11 to 0.29, providing a same-modality reference (App. G). Frozen retrieval uses the adapters of Krzakala et al. [30], trained with in-batch negatives under one recipe shared by all models, and is evaluated by Recall@1 and mean reciprocal rank (MRR) of the true molecule (App. D).

DreaMS. The self-supervised stage separates weakly, with silhouette +0.14 (Fig. 1) and purity 0.68, and its agreement with MoLFormer is 0.09: above chance, but only a third of what the best-agreeing molecule encoders share. Contrastive fine-tuning doubles the silhouette to +0.28 and increases purity to 0.78, but leaves cross-modal agreement and retrieval MRR (0.23) essentially unchanged.

MS-JEPA. Despite using only molecule-grouped spectra from the MassSpecGym training fold, without large-scale pretraining, MS-JEPA achieves substantially stronger separability: silhouette +0.49 and normalized purity 0.91 for the post-aggregation embedding. Thus, cross-condition prediction is promising for learning a representation that distinguishes molecular identities, even with limited training data. Yet MS-JEPA has the weakest agreement with MoLFormer (mutual- k NN 0.03) and the worst frozen retrieval performance: 0.12 MRR and 0.05 Recall@1. The strongest clustering therefore produces the weakest retrieval of the three encoders, contrary to the expectation that separability alone should facilitate frozen alignment.

Supervising cross-modal geometry. We added VA-VAE’s alignment loss [60] to MS-JEPA’s pretraining and DreaMS’s contrastive stage, directly biasing the encoders towards MoLFormer’s geometry. Notably, both interventions have a positive effect on retrieval while reducing same-molecule silhouette. MS-JEPA’s frozen MRR rises to 0.19, while mutual- k NN increases from 0.03 to 0.06. For DreaMS, linear CKA with MoLFormer rises from 0.20 to 0.46, with little change in mutual- k NN and frozen MRR (App. H).

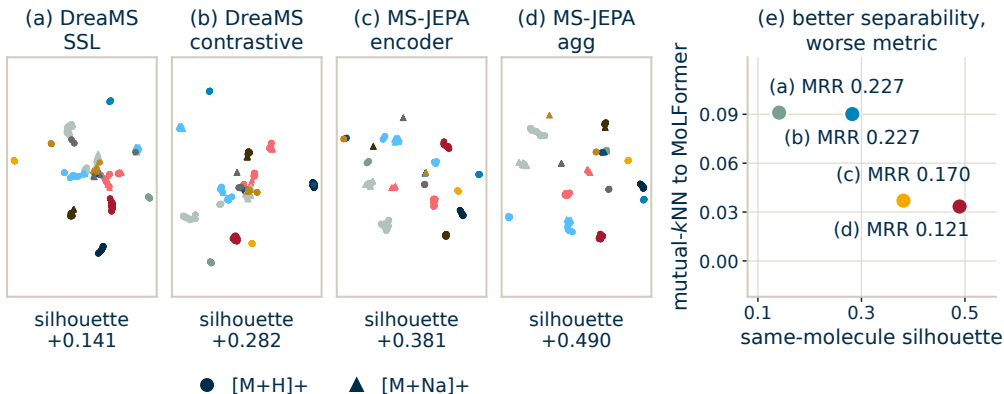


Figure 1: **Separability is not alignability.** (a–d) UMAP of the same test spectra (12 molecules with ≥ 12 spectra) using four different representations. Color is the molecule. (e) Separability against cross-modal agreement with MoLFormer, annotated with frozen MRR (Table 3).

5 What the adapter learns

If the frozen towers do not agree, the adapters carry the performance, as the molecule foundation model makes it easy to exploit the benchmark’s decoy distribution.

The candidate objective. MSAAlign optimizes InfoNCE [51] over the candidate set $\mathcal{K}(s)$, drawn from PubChem using a mass or formula filter (App. I, Eq. (4)). Writing q for the induced proposal and p for the library-structure marginal, the population minimizer is, up to a function of s alone [51, 41],

$$\rho^*(s, m) = \log \frac{p(m | s)}{q(m)} = \underbrace{\log \frac{p(m)}{q(m)}}_{(A) \text{ prior}} + \underbrace{\log \frac{p(s | m)}{p(s)}}_{(B) \text{ PMI}}. \quad (3)$$

(B) is the spectrum–structure pointwise mutual information. (A) does not depend on s : it is the log-odds that a structure is a library compound rather than a decoy. With in-batch negatives $p = q$, so (A) vanishes. However, for conventional benchmark candidate pools p and q are substantially different [18], raising a potential concern.

Experiments. We rank every candidate by its expected score over the training spectra, $\mathbb{E}_s[\rho(s, m)]$, which sees neither the spectrum nor its acquisition conditions (App. I), and measure the remaining ranking accuracy. With in-batch negatives, performance falls to chance ($\Delta_0 \leq 0$, Table 1, left), consistent with (3). With benchmark candidates as negatives, Δ_0 reaches 0.12–0.14, accounting for at least half the full model’s gain over random. Candidate training increases blind MRR more than full MRR, reducing the spectral margin Δ_s . Similar but weaker effects occur in other candidate-trained models, particularly without pretrained molecule representations (App. I.1). This is particularly significant because the decoys do not follow the same stringent train–test splitting protocol used to support generalization claims [36].

6 Improved baselines

Frozen alignment offers an efficient way to exploit foundation models trained on unpaired data, while preserving pretrained features that fine-tuning can distort under distribution shift [31]. Yet it depends on the geometry of independently trained encoders. One alternative is to use forward spectrum predictors to generate paired data for millions of molecular structures. Unlike spectrum-only pretraining, this directly supervises spectrum–molecule alignment and can cover molecules absent from experimental spectral libraries.

End-to-end adaptation. We first unfreeze the spectrum encoder while keeping MoLFormer fixed (App. B.4, J). DreaMS improves from 0.227 to 0.244–0.255 MRR across three learning rates, and MS-JEPA from 0.121 to 0.243 using its aggregated representation. Remarkably, a fresh encoder matching MS-JEPA’s architecture reaches 0.232 MRR. Thus, despite the motivation for freezing

Table 1: **Spectrum-blind ranking (left) and improved baselines (right)**. Left: frozen DreaMS contrastive; $\Delta_0 = \text{MRR}_{\text{blind}} - \text{MRR}_{\text{rand}}$, $\Delta_s = \text{MRR}_{\text{full}} - \text{MRR}_{\text{blind}}$. Right: Ablation of simulated pretraining, fine-grained scoring, and candidate-aware training for fresh and MS-JEPA-pretrained encoders (App. J).

Tower	Negatives	MRR	Δ_0	Δ_s	Training	Score	fresh		pretrained	
							MRR	R@1	MRR	R@1
ChemBERTa	in-batch	0.147	-0.004	0.127	real	global	0.232	0.124	0.243	0.128
ChemBERTa	candidate	0.211	0.122	0.064	real	MaxSim	0.241	0.130	0.234	0.129
MolFormer	in-batch	0.227	-0.006	0.209	sim→real	global	0.369	0.229	0.374	0.238
MolFormer	candidate	0.303	0.144	0.135	sim→real	MaxSim	0.463	0.334	0.476	0.341
					sim→real	MaxSim + cand.	0.566	0.425	0.542	0.400

under distribution shift, spectrum pretraining provides little advantage once paired supervision is available.

Simulated paired data. We pretrain on spectra of 5.9M leakage-audited molecules, simulated by GLACIER and FraGNNNet [54, 63] (App. K), before fine-tuning on the 25k measured training molecules. Table 1 compares global cosine scoring with fine-grained MaxSim [7] (App. B.4) under both initializations. Simulated pretraining substantially improves global retrieval (+0.13–0.14 MRR), but its greater benefit is enabling fine-grained scoring: MaxSim adds virtually nothing on real data alone, yet gains another 0.09–0.10 MRR after simulated pretraining. The fresh encoder reaches 0.463 MRR, within 0.013 of pretrained MS-JEPA (0.476).

Candidate-aware training. Adding the candidate objective raises the fresh model to 0.566 MRR and 0.425 Recall@1, ahead of the pretrained model (0.542 MRR). However, as in §5, it raises spectrum-blind MRR by 0.070 and 0.085. Meanwhile, longer training on the real corpus saturates and eventually hurts validation performance (App. J). Expanding molecular coverage, rather than refining the frozen representation or repeatedly fitting the same molecules, delivers the largest gains. Fine-grained scoring benefits most once that broader supervision is available.

7 Conclusion

Frozen spectrum–molecule alignment borrows the premise that independently pretrained encoders share enough geometry for lightweight alignment [24, 67]. Our results expose the limits of this premise for MS/MS. Molecular separability does not determine between-molecule geometry: MS-JEPA achieves the strongest same-molecule clustering yet the weakest frozen retrieval. Across spectrum and molecule encoders, cross-modal agreement remains low. Concurrent work likewise reports near-chance label-free alignment between DreaMS and molecular fingerprints on MassSpecGym [45]. More generally, spectral similarity is an imperfect proxy for molecular similarity: the cosine–Tanimoto correlation [6, 7] of self-supervised DreaMS is 0.63 under similarity-balanced sampling but only 0.23 under uniform pair sampling, with a substantial correlation also obtained from precursor mass alone (App. F). Together, these findings reveal a broader difficulty with frozen alignment in this modality.

Retrieval performance can also obscure what the spectrum representation contributes. Candidate-trained adapters retain substantial accuracy without query-specific spectral input, which reinforces the need for spectrum-blind controls.

Our improved baselines point to one possible alternative: rather than relying on compatible frozen representations, forward spectrum models provide paired supervision across millions of molecular structures. End-to-end training largely eliminates the advantage of spectrum pretraining, while simulated spectra from 5.9M molecules unlock fine-grained scoring and raise a fresh encoder from 0.232 to 0.463 MRR without candidate negatives. These results identify molecular coverage as the central constraint in our setting. Simulated paired data supplies this missing supervision at a scale that measured spectral libraries cannot currently provide.

Acknowledgments and Disclosure of Funding

This work was partially supported by NSF award #2406231 and the Artur Dubrawski Alumni Research Professorship Chair funds. This work used Bridges-2 at Pittsburgh Supercomputing Center

[4] through allocation CHE260128 from the ACCESS program, which is supported by NSF grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

- [1] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, volume 162, pages 1298–1312. PMLR, 2022.
- [2] M. Bohde, M. Manjrekar, R. Wang, S. Ji, and C. W. Coley. DiffMS: Diffusion generation of molecules conditioned on mass spectra. In *International Conference on Machine Learning*, volume 267, pages 4737–4756. PMLR, 2025.
- [3] P. L. Bremer, A. Vaniya, T. Kind, S. Wang, and O. 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.
- [4] S. T. Brown, P. Buitrago, E. Hanna, S. Sanielevici, R. Scibek, and N. A. Nystrom. Bridges-2: A platform for rapidly-evolving and data intensive research. In *Practice and Experience in Advanced Research Computing*, pages 1–4. ACM, 2021.
- [5] R. Bushuiev, A. Bushuiev, N. F. de Jonge, A. Young, F. Kretschmer, R. Samusevich, J. Heirman, F. Wang, L. Zhang, K. Dührkop, M. Ludwig, N. A. Haupt, A. Kalia, C. Brungs, R. Schmid, R. Greiner, B. Wang, D. S. Wishart, L.-P. Liu, J. Rousu, W. Bittremieux, H. Röst, T. D. Mak, S. Hassoun, F. Huber, J. J. J. van der Hooft, M. A. Stravs, S. Böcker, J. Sivic, and T. Pluskal. MassSpecGym: A benchmark for the discovery and identification of molecules. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pages 110010–110027, 2024.
- [6] R. Bushuiev, A. Bushuiev, R. Samusevich, C. Brungs, J. Sivic, and T. Pluskal. Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. *Nature Biotechnology*, 44(4):630–640, 2026.
- [7] Y. Z. Chen, B. Rushing, and S. Hassoun. FLARE: Fine-grained Learning for Alignment of spectra-molecule REpresentation enhances metabolite annotation. bioRxiv preprint, 2026.
- [8] S. Chithrananda, G. Grand, and B. Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885, 2020.
- [9] S. Crotti, M. Menicatti, M. Pallecchi, and G. Bartolucci. Tandem mass spectrometry approaches for recognition of isomeric compounds mixtures. *Mass Spectrometry Reviews*, 42(4):1244–1260, 2023.
- [10] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024.
- [11] K. Dührkop, H. Shen, M. Meusel, J. Rousu, and S. Böcker. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences*, 112(41):12580–12585, 2015.
- [12] Q. Garrido, L. Najman, and Y. LeCun. Self-supervised learning of Split Invariant Equivariant representations. In *International Conference on Machine Learning*, volume 202, pages 10975–10996. PMLR, 2023.
- [13] H. Ghaemi, E. B. Muller, and S. Bakhtiari. seq-JEPA: Autoregressive predictive learning of invariant-equivariant world models. In *Advances in Neural Information Processing Systems*, volume 38, pages 37034–37064, 2025.
- [14] R. Giné, I. Pérez-López, J. M. Badia, J. Capellades, and O. Yanes. Benchmarking MS/MS featurization strategies for machine learning-driven metabolite structure annotation. *Journal of the American Society for Mass Spectrometry*, 37(7):1550–1561, 2026.

- [15] S. Goldman, J. Wohlwend, M. Stražar, G. Haroush, R. J. Xavier, and C. W. Coley. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence*, 5(9):965–979, 2023.
- [16] S. Goldman, J. Li, and C. W. Coley. Generating molecular fragmentation graphs with autoregressive neural networks. *Analytical Chemistry*, 96(8):3419–3428, 2024.
- [17] C. M. Grulke, A. J. Williams, I. Thillanadarajah, and A. M. Richard. EPA’s DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research. *Computational Toxicology*, 12:100096, 2019.
- [18] V. Gupta, C. Xu, E. Herbst, F. Wang, D. S. Wishart, and M. A. Skinnider. Confronting spurious evaluations of computational methods in small molecule mass spectrometry. *bioRxiv preprint*, 2026.
- [19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.
- [20] J. Heirman and W. Bittremieux. Reusability report: Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence*, 6(11):1296–1302, 2024.
- [21] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger. Deep networks with stochastic depth. In *European Conference on Computer Vision*, pages 646–661. Springer, 2016.
- [22] F. Huber, L. Ridder, S. Verhoeven, J. H. Spaaks, F. Diblen, S. Rogers, and J. J. J. van der Hoof. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology*, 17(2):e1008724, 2021.
- [23] F. Huber, S. van der Burg, J. J. J. van der Hoof, and L. Ridder. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *Journal of Cheminformatics*, 13(1):84, 2021.
- [24] M. Huh, B. Cheung, T. Wang, and P. Isola. Position: The Platonic Representation Hypothesis. In *International Conference on Machine Learning*, volume 235, pages 20617–20642. PMLR, 2024.
- [25] A. Kalia, Y. Z. Chen, D. Krishnan, and S. Hassoun. JESTR: Joint Embedding Space Technique for Ranking candidate molecules for the annotation of untargeted metabolomics data. *Bioinformatics*, 41(7):btaf354, 2025.
- [26] K. Kläser, B. Banaszewski, S. Maddrell-Mander, C. McLean, L. Müller, A. Parviz, S. Huang, and A. Fitzgibbon. MiniMol: A parameter-efficient foundation model for molecular learning. *arXiv preprint arXiv:2404.14986*, 2024.
- [27] A. S. Koepke, D. Zverev, S. Ginosar, and A. A. Efros. Back into Plato’s Cave: Examining cross-modal representational convergence at scale. *arXiv preprint arXiv:2604.18572*, 2026.
- [28] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, volume 97, pages 3519–3529. PMLR, 2019.
- [29] F. Kretschmer, J. Seipp, M. Ludwig, G. W. Klau, and S. Böcker. Coverage bias in small molecule machine learning. *Nature Communications*, 16(1):554, 2025.
- [30] P. Krzakala, G. Melo, C. Lançon, C. Laclau, R. Flamary, E. Thévenot, and F. d’Alché-Buc. MSAlign: Aligning molecule and mass spectra foundation models for metabolite identification. *arXiv preprint arXiv:2605.19752v1*, 2026.
- [31] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

- [32] H. Lam. Building and searching tandem mass spectral libraries for peptide identification. *Molecular & Cellular Proteomics*, 10(12):R111.008565, 2011.
- [33] B. Li, Y. Zhang, X. Wang, W. Liang, L. Schmidt, and S. Yeung-Levy. Closing the modality gap for mixed modality search. *Transactions on Machine Learning Research*, 2026.
- [34] Y. Li, T. Kind, J. Folz, A. Vaniya, S. S. Mehta, and O. Fiehn. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nature Methods*, 18(12):1524–1531, 2021.
- [35] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 17612–17625, 2022.
- [36] H. Liu, R. Bushuiev, I. Lightheart, M. Manjrekar, A. Bushuiev, M. Lederbauer, F. Jozefov, Y. Wang, S. Hassoun, J. Sivic, J. Taylor, R. Wang, D. Healey, T. Pluskal, and C. W. Coley. MassSpecGym in the wild: Uncovering and correcting evaluation pitfalls in AI-driven molecule discovery. arXiv preprint arXiv:2606.19624, 2026.
- [37] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, volume 97, pages 4114–4124. PMLR, 2019.
- [38] M. Murphy, S. Jegelka, E. Fraenkel, T. Kind, D. Healey, and T. Butler. Efficiently predicting high resolution mass spectra with graph neural networks. In *International Conference on Machine Learning*, volume 202, pages 25549–25562. PMLR, 2023.
- [39] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, volume 32, pages 3942–3951, 2018.
- [40] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, S. Johansson, H. Chen, S. Nikolenko, A. Aspuru-Guzik, and A. Zhavoronkov. Molecular Sets (MOSES): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11:565644, 2020.
- [41] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, volume 97, pages 5171–5180. PMLR, 2019.
- [42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139, pages 8748–8763. PMLR, 2021.
- [43] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.
- [44] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.
- [45] D. Schnaus, T. Dagès, D. Cremers, X. Wang, and P. Isola. Shared geometry as a Rosetta Stone: Cross-modal alignment without paired data. arXiv preprint arXiv:2610.09411, 2026.
- [46] E. Soares, E. Vital Brazil, V. Shirasuna, D. Zubarev, R. Cerqueira, and K. Schmidt. An open-source family of large encoder-decoder foundation models for chemistry. *Communications Chemistry*, 8(1):193, 2025.
- [47] M. Sorokina, P. Merseburger, K. Rajan, M. A. Yirik, and C. Steinbeck. COCONUT online: Collection of Open Natural Products database. *Journal of Cheminformatics*, 13(1):2, 2021.

- [48] M. Tjandrasuwita, C. Ekbote, L. Ziyin, and P. P. Liang. Understanding the emergence of multimodal representation alignment. In *International Conference on Machine Learning*, volume 267, pages 59723–59760. PMLR, 2025.
- [49] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 23–30, 2017.
- [50] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou. Going deeper with image transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021.
- [51] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.
- [52] N. Vouitsis, Z. Liu, S. K. Gorti, V. Villecroze, J. C. Cresswell, G. Yu, G. Loaiza-Ganem, and M. Volkovs. Data-efficient multimodal fusion on a single GPU. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27239–27251, 2024.
- [53] F. Wang, J. Liigand, S. Tian, D. Arndt, R. Greiner, and D. S. Wishart. CFM-ID 4.0: More accurate ESI-MS/MS spectral prediction and compound identification. *Analytical Chemistry*, 93(34):11692–11700, 2021.
- [54] R.-X. Wang, R. Wang, and C. W. Coley. GLACIER: Rethinking mass spectrum prediction as an object detection problem. In *Advances in Neural Information Processing Systems*, 2026.
- [55] R. Wang, M. Manjrekar, B. Mahjour, J. Avila-Pacheco, J. Provenzano, E. Reynolds, M. Lederbauer, E. Mashin, S. Goldman, M. Wang, J.-K. Weng, D. L. Plata, C. B. Clish, and C. W. Coley. Neural spectral prediction for structure elucidation with tandem mass spectrometry. bioRxiv preprint, 2025.
- [56] Y. Wang, X. Chen, L. Liu, and S. Hassoun. MADGEN: Mass-spec attends to de novo molecular generation. In *International Conference on Learning Representations*, 2025.
- [57] D. S. Wishart, A. Guo, E. Oler, F. Wang, A. Anjum, H. Peters, R. Dizon, Z. Sayeeda, S. Tian, B. L. Lee, M. Berjanskii, R. Mah, M. Yamamoto, J. Jovel, C. Torres-Calzada, M. Hiebert-Giesbrecht, V. W. Lui, D. Varshavi, D. Varshavi, D. Allen, D. Arndt, N. Khetarpal, A. Sivakumaran, K. Harford, S. Sanford, K. Yee, X. Cao, Z. Budinski, J. Liigand, L. Zhang, J. Zheng, R. Mandal, N. Karu, M. Dambrova, H. B. Schiöth, R. Greiner, and V. Gautam. HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Research*, 50(D1):D622–D631, 2022.
- [58] H.-T. Wu, D. L. Riggs, Y. A. Lyon, and R. R. Julian. Statistical framework for identifying differences in similar mass spectra: Expanding possibilities for isomer identification. *Analytical Chemistry*, 95(17):6996–7005, 2023.
- [59] T. Xie, H. Zhang, Q. Yang, J. Sun, Y. Wang, J. Long, Z. Zhang, and H. Lu. CSU-MS²: A contrastive learning framework for cross-modal compound identification from MS/MS spectra to molecular structures. *Analytical Chemistry*, 97(25):13350–13360, 2025.
- [60] J. Yao, B. Yang, and X. Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15703–15712, 2025.
- [61] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T.-Y. Liu. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems*, volume 34, pages 28877–28888, 2021.
- [62] A. Young, H. Röst, and B. Wang. Tandem mass spectrum prediction for small molecules using graph transformers. *Nature Machine Intelligence*, 6(4):404–416, 2024.
- [63] A. Young, F. Wang, D. S. Wishart, B. Wang, R. Greiner, and H. Röst. FraGNNNet: A deep probabilistic model for tandem mass spectrum prediction. *Transactions on Machine Learning Research*, 2025.

- [64] X. Yu, Y. Li, Y. Zhang, H. Zhao, Y. Yang, H. Tang, Y. Song, X. Hu, C. Qin, S. Yan, and H. Xiong. Anisotropic modality align. arXiv preprint arXiv:2605.07825, 2026.
- [65] X. Yu, Y. Xin, Y. Zhang, W. Zhang, C. Liu, H. Zhao, C. Liu, X. Hu, Z. Qiao, H. Tang, X. Hu, C. Qin, H. Xiong, Y. Qiao, and S. Yan. Modality gap-driven subspace alignment training paradigm for multimodal large language models. arXiv preprint arXiv:2602.07026, 2026.
- [66] H. Zhang, Q. Yang, T. Xie, Y. Wang, Z. Zhang, and H. Lu. MSBERT: Embedding tandem mass spectra into chemically rational space by mask learning and contrastive learning. *Analytical Chemistry*, 96(42):16599–16608, 2024.
- [67] L. Zhang, Q. Yang, and A. Agrawal. Assessing and learning alignment of unimodal vision and language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14604–14614, 2025.

A Extended related work

Frozen-encoder alignment. CLIP trains both towers jointly [42]. FuseMix aligns frozen unimodal encoders using latent-space augmentation [52]; SAIL assesses alignment and achieves CLIP-competitive transfer with substantially less paired data [67]. MSAlign applies this approach to MS/MS [30].

Convergence and its limits. Huh et al. [24] report greater mutual- k NN agreement between more capable models. At larger gallery sizes, Koepke et al. [27] find that agreement falls from 0.135 on 1,024 samples to 0.008 on LAION-15M ($k=10$); within-modality agreement remains much higher, and many-to-many correspondence further reduces cross-modal agreement. Tjandrasuwita et al. [48] analyze when alignment emerges. Schnaus et al. [45] align unpaired embeddings by Wasserstein Procrustes, but obtain near-chance alignment between DreaMS and ECFP4 on MassSpecGym (FOSCTTM 0.47; chance 0.5). Related modality-gap studies examine the relative placement of modalities in a shared space [35, 33, 65, 64].

Spectrum representations. Spec2Vec [22] and MS2DeepScore [23] learn spectral similarity. MS-BERT [66] and DreaMS [6] pretrain transformers on unlabeled spectra; Giné et al. [14] compare featurization strategies.

Molecule retrieval. Methods differ in how they compare a spectrum with candidates. *Library search* matches measured spectra using cosine, entropy [34], or learned similarity [22, 23]; it requires a reference spectrum [32]. *Spectrum-to-fingerprint* methods predict structural features: CSI:FingerID uses fragmentation trees and kernel SVMs [11], and MIST uses subformula-annotated peaks [15, 20]. *Structure-to-spectrum* methods simulate candidates for spectral matching: CFM-ID [53, 3], GrAFF-MS [38], MassFormer [62], ICEBERG [16], GLACIER [54], and FraGNNNet [63]. We use the last two for simulated pretraining. *Joint embedding* methods rank candidates in a shared space: JESTR combines a graph network and spectrum MLP with in-batch InfoNCE and a candidate-list penalty [25]; CSU-MS² uses contrastive spectrum–structure learning [59]; MSAlign adapts frozen encoders [30]; and FLARE uses bidirectional peak–atom MaxSim with subformula-annotated peaks and known parent formula and adduct [7]. De novo methods generate structures beyond a fixed candidate list [5, 2, 56].

Predictive SSL. Equation (2) predicts latent targets, as in data2vec [1], using an EMA encoder [19]. Garrido et al. [12], Ghaemi et al. [13] develop invariant and equivariant representations. Our non-identifiability result is related to, but distinct from, the impossibility of unsupervised disentanglement without inductive biases [37].

B Methods

Table 2 summarizes the three spectrum encoders.

Table 2: The three spectrum encoders. The DreaMS and MS-JEPA columns describe pretraining; the fresh encoder is trained end to end with the recipe of App. D. The last two rows describe retrieval training (App. B.4, B.5).

	DreaMS (reproduced)	MS-JEPA	fresh encoder
Blocks / width / heads / MLP	7 / 1,024 / 8 / 4,096	8 / 768 / 12 / 3,072	8 / 768 / 12 / 3,072
Parameters	95.5M	64.1M encoder; 87.7M as (A)	87.7M, as MS-JEPA (A)
Peaks per spectrum	60 (SSL), 100 (contrastive)	100	100
m/z encoding	fixed Fourier	fixed Fourier, m/z and neutral loss	as MS-JEPA
Attention bias	no own parameters, shared	learned, per head and block	as MS-JEPA
Encoder conditioning	none	FiLM gains: adduct, charge, polarity	as MS-JEPA
Representation	precursor token	[CLS] token, aggregated	as MS-JEPA (A)
Objective	masked m/z , then contrastive	JEPA cosine to an EMA target	InfoNCE to MoLFormer
Training data	GeMS-A10, ≈ 24 M spectra	MSG train, 175,536 spectra	MSG train, 194,119 spectra
Batch / optimizer steps	4,096 spectra / 71,288	256 episodes / 10,240	4×64 spectra / 9,700
Optimizer	Adam 10^{-4} , Noam	AdamW 2×10^{-4} , cosine	AdamW 10^{-4} , cosine
Precision / hardware	fp32, 8 H100	bf16, 4 A100	bf16, 4 A100
Adapter conditions	adduct, CE	none (A); CE, instrument (B)	none
Trainable, frozen / trained	8.42M / 104.0M	7.35M / 95.0M (A)	– / 95.0M

B.1 Notation

Symbols follow §2. Write $z_s = \bar{u}_\theta(f(s, a))$ and $z_m = \bar{v}_\theta(g(m))$ on $\mathbb{S}^{d-1}$, so $\rho_\theta(s, m) = \tau^{-1}\langle z_s, z_m \rangle$ as in (1). Σ_m is the support of $f(s, a)$ over measurements of m under all conditions.

B.2 The spectrum encoder of MS-JEPA

The input contains learned [CLS] and register tokens [10], the precursor (intensity 1.1), and the 100 most intense fragments, normalized to the base peak. Each peak combines three branches: fixed Fourier features of m/z and precursor-minus- fragment mass (12,000 features; periods 1–1,000 Da and 0.1–999.9 mDa), mapped by a shared mass MLP to 256 dimensions each; and four intensity transforms (identity, square root, logarithm, and fraction of total fragment intensity), mapped to 128 dimensions. Layer normalization and a linear map fuse these branches into a 768-dimensional token, with an added token-type embedding.

Eight pre-norm transformer blocks use 12 heads, MLP width 3,072, layer scale initialized at 0.01 [50], and stochastic depth up to 0.1 [21]. Each block learns a per-head attention bias from pairwise mass differences (Fourier features at 16 periods from 1 mDa to 1 kDa), intensities, and token types. By comparison, DreaMS’s Graphormer-style bias [61] uses differences of summed Fourier-branch features, shared across heads and blocks, with no additional parameters.

Adduct, charge, and polarity embeddings are projected to 128 dimensions and modulate each block’s attention and MLP residual branches through FiLM gains $1 + \gamma$. Zero initialization of the gain layer starts training with an unconditioned encoder; replacing the gains by the identity with probability 0.1 permits use without conditions. The trunk has 64.1M parameters.

B.3 MS-JEPA

Episodes. Spectra are grouped by molecule and condition (adduct, instrument family, collision energy). Each episode samples a molecule with at least two condition groups, a target from one group, and context from the others: one spectrum, or three if at least four groups exist. The training fold provides 16,294 eligible molecules and 175,536 spectra (adducts $[M+H]^+$ and $[M+Na]^+$); 5,510 molecules support three contexts.

Prediction. The encoder’s [CLS] output is z_{spec} . A set aggregator (22.0M parameters) concatenates each context representation with its source condition, projects to 768 dimensions, and applies three transformer blocks with a learned pooling token to produce h_{mol} . Every episode uses its first context ($K=1$); three-context episodes also use all three ($K=3$). The condition encoder (1.6M parameters, mostly FiLM gains) embeds adduct, charge, polarity, instrument, and Fourier collision-energy features for the aggregator and predictor. The predictor (3.7M parameters) maps h_{mol} through two residual MLP blocks of hidden width 1,024, FiLM-modulated by the target condition.

The target is the EMA encoder’s z_{spec} under the target spectrum’s condition, with gradients stopped. The loss averages the $1 - \cos$ terms for $K=1$ over all episodes and $K=3$ over three-context episodes. Target encoder, condition encoder, and prefix tokens follow their online counterparts by EMA, with cosine-scheduled momentum from 0.996 to 1.

Optimization. We use AdamW ($\beta = (0.9, 0.999)$), peak learning rate 2×10^{-4} , an 8-epoch warm-up, and cosine decay to 10^{-6} . Layer-wise learning-rate decay is 0.9, with a further $0.5 \times$ factor on the peak encoder. Weight decay rises from 0.04 to 0.40 on a cosine schedule; gradients are clipped at 3. Training uses four A100-80GB GPUs, 64 episodes per GPU (256 per step), bf16, and seed 42: 160 epochs of 64 steps (10,240 total), in under one hour. Validation selects epoch 139, step 8,960, by $K=1$ prediction MRR among the true target and 63 targets sampled uniformly from other validation molecules.

Retrieval. We retain the online encoder, condition encoder, prefix tokens, and aggregator. **(A)** The encoder receives ion conditions; the aggregator processes one spectrum ($K=1$) with the full source condition (adduct, charge, polarity, instrument, Fourier collision-energy features, and validity bit). The adapter receives h_{mol} and no conditions (87.7M encoder parameters). **(B)** The adapter receives z_{spec} , collision energy, and instrument; the encoder sees only ion conditions (65.7M parameters). Results use (A) unless stated otherwise.

B.4 End-to-end and fresh encoders

End-to-end models train the spectrum encoder and both adapters while keeping MoLFormer frozen. Initialization is DreaMS contrastive, MS-JEPA (A) or (B), or a fresh MS-JEPA (A) encoder with identity FiLM gains. The global vector (h_{mol} , z_{spec} , or DreaMS’s precursor token), with the same adapter conditions as in the frozen model, feeds the spectrum adapter (App. B.5). No predictor or extra query/readout tokens are used. DreaMS dropout (0.1), MS-JEPA stochastic depth, and FiLM condition dropout remain active; DreaMS Fourier frequencies remain fixed. Encoder learning rates are 5×10^{-6} , 1.5×10^{-5} , or 5×10^{-5} for pretrained models and 10^{-4} or 3×10^{-4} for fresh models; adapters use 10^{-4} . Other settings follow App. D. Table 1 uses validation-selected configurations (App. J).

Fine-grained scoring. The spectrum side is the set $S(s)$ of the global vector and the encoder’s last-layer precursor and fragment tokens (at most 100), each with the adapter conditions appended and passed through the spectrum adapter. The molecule side is the set $T(m)$ of MoLFormer’s last-layer token states except BOS, EOS, PAD and MASK, each through the molecule adapter; the unknown-token state is kept, since it stands for an out-of-vocabulary atom. With every vector ℓ_2 -normalized,

$$\rho_{\theta}(s, m) = \frac{1}{\tau |S(s)|} \sum_{u \in S(s)} \max_{v \in T(m)} \langle u, v \rangle,$$

the spectrum-to-molecule direction of MaxSim [7]. With one vector a side it is the cosine score of (1). Token states are computed by the frozen tower on the fly.

B.5 MSAAlign

Our MSAAlign adapters [30] each comprise a linear layer to 2,048 dimensions, LayerNorm, GELU, dropout 0.2, a linear layer to 1,024, and ℓ_2 normalization. Scores are cosine similarities divided by a learned temperature initialized at 0.07. The spectrum input concatenates the encoder output (1,024 dimensions for DreaMS, 768 for MS-JEPA) with 128-dimensional embeddings of conditions absent from the encoder: adduct and collision energy for DreaMS; collision energy and instrument for MS-JEPA (B); none for (A). Adduct and instrument use learned embeddings. Collision energy uses sine and cosine at 16 log-spaced frequencies (0.01–10 rad/eV), a validity bit, and a two-layer MLP.

MoLFormer-XL-both-10pct [43] and ChemBERTa-zinc-base-v1 [8] are frozen at pinned snapshots in evaluation mode, with fixed random attention features for MoLFormer. Both output 768 dimensions. Global vectors are MoLFormer’s mean over attended tokens and ChemBERTa’s $\langle s \rangle$ state.

Training. We use in-batch or candidate InfoNCE (Eq. (4)) and the recipe in App. D; simulated pairs share identity when they share a store molecule (App. K). With frozen encoders, trainable parameters total 8.42M for DreaMS (4.73M spectrum adapter, 3.68M molecule adapter, 21k condition embeddings, and temperature), 7.35M for MS-JEPA (A), and 7.90M for (B).

B.6 DreaMS: deviations from the published recipe

We retrain both stages because the released model’s contrastive stage uses MoNA, a source of MassSpecGym spectra.

Self-supervised stage. We run the released checkpoint’s configuration on GeMS-A10, changing hardware to eight H100-80GB GPUs (from A100-40GB), validation to 0.5% of LC-MS runs (from 1% of spectra), mask sampling to distribution-equivalent GPU Gumbel-top- k , the Noam step index by one, and checkpoint frequency to every 2,750 steps. We use step 71,288.

Contrastive stage. We apply the released MassSpecGym fine-tuning settings to the training fold: Adam at 3×10^{-6} without scheduling, 32 anchors per step, cosine-triplet margin 0.1, positives sharing an InChIKey connectivity block, negatives within 0.05 Da neutral mass, 100 peaks, fp32, seed 3407, and an unfrozen backbone throughout. Anchors are restricted to $[M+H]^+$ to prevent adduct-based separation of mass-window negatives. We train for 25,000 steps and use step 24,000.

Retrieval. Each checkpoint retains its training peak budget (60 SSL, 100 contrastive), padded to that length because the released attention does not mask padded keys. The representation is the final precursor token.

C Proofs

Proof of Proposition 1. Write $L(f) = \inf_h L(f, h)$, where

$$L(f, h) = \mathbb{E} \ell \left(\left\langle \overline{h(X_C, A_C, a_t)}, \overline{f(s_t, a_t)} \right\rangle \right), \quad X_C = \{f(s_i, a_i)\}_{i \in C},$$

and the context and target spectra come from the same molecule.

Let $f' = R \circ f$ and $\Sigma'_m = Q_m \Sigma_m$. Since the transformed supports are pairwise disjoint, any nonempty same-molecule context uniquely determines m . For a valid transformed context $x_C \subset \Sigma'_m$, define

$$h'(x_C, A_C, a_t) = Q_m h(Q_m^{-1} x_C, A_C, a_t),$$

where $Q_m^{-1} x_C = \{Q_m^{-1} x_i : x_i \in x_C\}$. Set h' to a fixed nonzero vector on other contexts. Since the molecule family is finite or countable and the supports are Borel, this defines a measurable predictor.

For a context–target sample from molecule m ,

$$\overline{h'(X'_C, A_C, a_t)} = Q_m \overline{h(X_C, A_C, a_t)}, \quad \overline{f'(s_t, a_t)} = Q_m \overline{f(s_t, a_t)}.$$

Orthogonality of Q_m leaves the cosine unchanged, so $L(f', h') = L(f, h)$. Hence $L(f') \leq L(f)$. Applying the same construction to the inverse of R on $\bigcup_m \Sigma'_m$, which acts as Q_m^{-1} on each Σ'_m , gives the reverse inequality and therefore

$$L(R \circ f) = L(f).$$

For the second claim, let $\{\bar{g}(m)\} \subset \mathbb{S}^{d-1}$, $d \geq 2$, be distinct, suppose that $m(s, a)$ is measurable, and let each $\mathcal{K}(m)$ contain m and at least one other molecule. Set

$$f(s, a) = \bar{g}(m(s, a)).$$

Then $\Sigma_m = \{\bar{g}(m)\}$, and the score

$$\rho_f(s, m') = \langle f(s, a), \bar{g}(m') \rangle$$

is uniquely maximized at $m' = m(s, a)$.

For each m , choose $\sigma(m) \in \mathcal{K}(m) \setminus \{m\}$ and let

$$\delta_m = \angle(\bar{g}(m), \bar{g}(\sigma(m))) > 0.$$

Because $d \geq 2$ and the molecule family is at most countable, we may choose pairwise distinct unit vectors y_m such that

$$\angle(y_m, \bar{g}(\sigma(m))) < \delta_m/3.$$

Choose an orthogonal Q_m satisfying $Q_m \bar{g}(m) = y_m$, and define $R|_{\Sigma_m} = Q_m$. The transformed supports $\{y_m\}$ are pairwise disjoint. Moreover, by the reverse triangle inequality,

$$\angle(y_m, \bar{g}(m)) \geq \delta_m - \angle(y_m, \bar{g}(\sigma(m))) > 2\delta_m/3 > \angle(y_m, \bar{g}(\sigma(m))).$$

Thus, under $f' = R \circ f$, the candidate $\sigma(m)$ has strictly larger cosine similarity than the true molecule m . Hence direct cosine retrieval with f' never ranks the true molecule first, whereas retrieval with f always does, despite $L(f') = L(f)$. $\square$

Proposition 2 (Raw kernel agreement contracts each support). *Let $f(s, a) \in \mathbb{S}^{d_s-1}$ and $g(m) \in \mathbb{S}^{d_m-1}$, and let $\varepsilon \geq 0$. Suppose*

$$|\langle x, y \rangle - \langle g(m), g(n) \rangle| \leq \varepsilon$$

for all m, n and all $x \in \Sigma_m, y \in \Sigma_n$. Then $\text{diam}(\Sigma_m) \leq \sqrt{2\varepsilon}$ for every m . In particular, if $\varepsilon = 0$, every nonempty Σ_m is a singleton.

Proof. Fix m and $x, y \in \Sigma_m$. Taking $n = m$ and using $\langle g(m), g(m) \rangle = 1$ gives $\langle x, y \rangle \geq 1 - \varepsilon$. Because x and y are unit normalized,

$$\|x - y\|_2^2 = 2 - 2\langle x, y \rangle \leq 2\varepsilon.$$

Taking the supremum over $x, y \in \Sigma_m$ proves the claim. $\square$

D Evaluation protocol

Data. We use MassSpecGym v1.5 [5, 36] and its official MCES split: 194,119 / 19,429 / 17,556 train / validation / test spectra from 25,046 / 3,386 / 3,170 query molecules (2,998 test InChIKey-14 blocks). Each query uses its official mass-based list, including the true molecule (12–256 candidates; median 256). Formula-based lists are used only for the leakage audit (App. K). Every tower reads RDKit canonical isomeric SMILES, which also identify the true candidate [36]. We drop 66 invalid candidate entries; canonicalization produces no duplicates within lists.

Metrics. The rank of the true molecule is 1 plus the number of candidates scored strictly higher (fp32 scores), so ties favor the true molecule; no run has ties on more than 4 of the 17,556 test spectra. Mean reciprocal rank

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^N \frac{1}{\text{rank}_i}$$

and $\text{Recall}@ \{1, 5, 20\}$ are averaged over test spectra. Uniform random ranking gives $\mathbb{E}[1/\text{rank}] = H_K/K$ for a list of K , MRR 0.024 on the test fold.

Training recipe. Except JESTR (App. I.1), all retrieval models share this recipe. Frozen encoders remain in evaluation mode with learning rate 0. Each epoch samples one spectrum per training query molecule in random order. Four A100-80GB GPUs each process 64 spectra per batch: 97 steps per epoch, dropping 214 molecules, for 100 epochs (9,700 steps).

Augmentation independently drops peaks with probability 0.2 (retaining at least one), multiplies intensities by $e^{0.3\varepsilon}$, and multiplies m/z by $1 + 5 \times 10^{-6}\varepsilon$, with $\varepsilon \sim \mathcal{N}(0, 1)$; the precursor is unchanged. Spectra are then restricted to the encoder’s peak budget and normalized to the base peak: 100 peaks by default, 60 for DreaMS SSL, and 128 for the MS-JEPA ablations in App. H.

AdamW ($\beta = (0.9, 0.999)$) uses learning rate 10^{-4} for adapters and condition embeddings and the rates in App. B.4 for encoders. Weight decay is 0.02 on parameters with at least two dimensions. We use 5% linear warm-up, cosine decay to zero, and gradient clipping at 1. Temperature is learned as $\log \tau$, initialized at $\tau = 0.07$, with learning rate 10^{-2} , no weight decay, and $\tau \geq 0.01$ enforced after each step. Forward passes use bf16 autocast; scores use fp32. Validation on official lists runs every 200 steps (485, or five epochs, for MaxSim) and at the final step; the highest MRR selects the checkpoint. App. J lists deviations.

Statistics. Models use one training seed (42), except the JESTR repeats. Checkpoints and configurations are selected by validation MRR, then evaluated once on test; JESTR uses the last checkpoint. We report 95% percentile intervals from 1,000 bootstrap samples of the 2,998 test InChIKey-14 blocks, recomputing metrics over their spectra. Model differences use paired samples. These intervals capture test-sample uncertainty, not training variance.

E Geometry probes

Within-modality probes use the 17,556 test spectra of 2,998 molecules (InChIKey-14); cross-modal probes use the 2,994 of them that every representation covers.

Same-molecule silhouette. For each spectrum, let a be its mean cosine distance to other spectra of the same molecule and b its smallest mean distance to another molecule. We average $(b - a)/\max(a, b)$ [44] over spectra with at least two replicates per molecule.

k NN purity. Purity is the same-molecule fraction among the k nearest other spectra. For a molecule with n spectra, its ceiling is $\min(k, n - 1)/k$ (fold mean 0.676 at $k=10$). We normalize by the per-spectrum ceiling for comparison across k and corpora. Chance is 0.0021.

Scatter summaries. $\text{tr } S_W$ is the mean squared distance to each molecule’s centroid; $\text{tr } S_B$ is the size-weighted mean squared distance of centroids to the global mean. Their ratio is the Fisher ratio; multiplying by $(N - C)/(C - 1)$ gives Calinski–Harabasz. Davies–Bouldin averages each molecule’s largest ratio of summed within-molecule spread to between-centroid distance. Alignment is the mean squared distance between same-molecule pairs; uniformity is the log mean Gaussian potential over all pairs.

Cross-modal agreement. Mutual- k NN [24] measures overlap between a molecule’s k nearest neighbors in two representations. Spectrum embeddings are averaged and normalized per molecule. For n molecules, chance overlap is $k/(n - 1)$ when the query is excluded. Linear CKA [28] is $\|Y^\top X\|_F^2 / (\|X^\top X\|_F \|Y^\top Y\|_F)$ for column-centered matrices. Within-modality references help distinguish poor agreement from scale effects in the estimator [27]. The two DreaMS stages agree at 0.372 but share architecture and most weights; App. G supplies independent molecule-encoder references.

Table 3: Frozen-geometry panel, MassSpecGym test fold. MS-JEPA is the aggregated representation; MS-JEPA [CLS] is the encoder’s [CLS] token, which sees only the ion condition. Arrows give the preferred direction. Frozen MRR: in-batch MoLFormer adapter (App. D); the adapter receives adduct and collision energy for DreaMS, no conditions for MS-JEPA (its encoder reads them), and collision energy and instrument for [CLS].

Metric	DreaMS SSL	DreaMS contr.	MS-JEPA	MS-JEPA [CLS]
<i>Within-modality (partition)</i>				
Same-molecule silhouette ($\uparrow$)	+0.141	+0.282	+0.490	+0.381
k NN purity@1 ($\uparrow$)	0.878	0.922	0.971	0.965
k NN purity@10, normalized ($\uparrow$)	0.683	0.782	0.905	0.881
Fisher trace ratio ($\uparrow$)	1.85	2.12	6.45	3.56
Calinski–Harabasz ($\uparrow$)	10.3	11.9	36.1	20.0
Davies–Bouldin ($\downarrow$)	1.463	1.346	1.140	1.221
Alignment ($\downarrow$)	0.510	0.448	0.142	0.277
Uniformity	−3.293	−3.386	−2.729	−3.169
<i>Cross-modal (metric)</i>				
cos–Tanimoto (mass null 0.370)	0.626	0.661	0.548	–
Mutual- k NN $\rightarrow$ MoLFormer, $k=10$ ($\uparrow$)	0.091	0.090	0.033	0.037
Mutual- k NN $\rightarrow$ ChemBERTa, $k=10$ ($\uparrow$)	0.061	0.060	0.024	0.027
Linear CKA $\rightarrow$ MoLFormer	0.265	0.209	0.260	0.241
Linear CKA $\rightarrow$ ChemBERTa	0.243	0.170	0.280	0.253
Frozen retrieval MRR	0.227	0.227	0.121	0.170

F Cosine–Tanimoto correlation

Mass null. DreaMS’s fixed precursor Fourier features define a parameter-free 11,994-dimensional representation of precursor m/z alone. Its cosine similarity closely tracks log mass difference ($r = -0.916$). On the same molecule pairs as the learned encoders, its cosine–Tanimoto correlation is 0.370 under the DreaMS protocol (56% of the contrastive encoder’s 0.661) and 0.189 under uniform sampling, where the encoders reach 0.185–0.228.

Sampling. The DreaMS protocol uses Morgan fingerprints (radius 2, 2,048 bits), one random spectrum per SMILES, and 2,500 pairs sampled with replacement from each of 40 Tanimoto bins (width 0.025). This raises mean Tanimoto from 0.108 to 0.501 and all correlations, including the mass null, by 2.0–3.6 $\times$. On this test fold, 28 bins contain fewer unique pairs than the quota, requiring repeated pairs at high similarity. Uniformly sampled pairs concentrate below Tanimoto 0.3 (Fig. 2).

G Extended comparison between encoders

Figure 3 scores every pair among five molecule and four spectrum representations: a Morgan fingerprint (radius 2, 4,096 bits), MoLFormer [43], ChemBERTa [8], SMI-TED [46] (289M parameters) and MiniMol [26] (10M, graph-based) on the molecule side; DreaMS SSL, DreaMS contrastive, MS-JEPA and spectral entropy similarity [34] on the spectrum side. Spectrum encoders enter as one normalized centroid per molecule. The entropy kernel has no embedding, so its molecule-level entry is the most similar pair of spectra between two molecules, and its CKA uses that similarity matrix directly.

Cross-modal mutual- k NN is at most 0.099 across all twenty molecule–spectrum pairs, compared with 0.11–0.29 among molecule representations and 0.37 between DreaMS stages. Fingerprints and MiniMol align best with DreaMS (0.096–0.099), followed by MoLFormer (0.090–0.091), SMI-TED

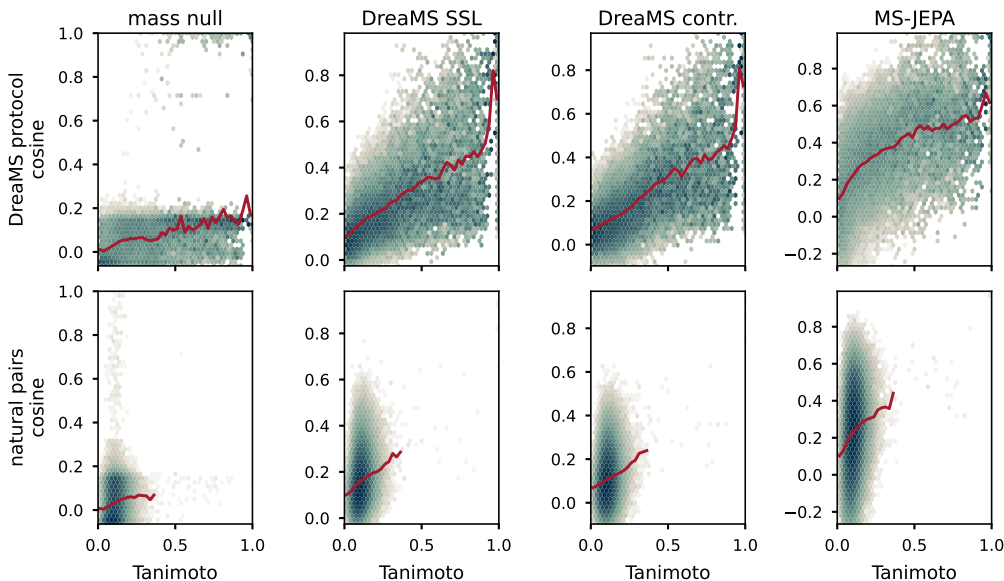


Figure 2: **Cosine against Morgan Tanimoto** for four representations of the same 3,170 test molecules (one random spectrum each), as a pair density with the mean cosine per Tanimoto bin (line). Top: DreaMS’s protocol, 40 Tanimoto bins with 2,500 pairs each drawn with replacement; Pearson 0.370 (mass null), 0.626 (DreaMS SSL), 0.661 (DreaMS contrastive), 0.548 (MS-JEPA). Bottom: 200,000 uniformly drawn molecule pairs; Pearson 0.189, 0.228, 0.185, 0.200 in the same order.

(0.063–0.067), and ChemBERTa (0.060–0.061). Fingerprints and MiniMol also agree most strongly with each other (0.285).

Without learning, entropy similarity reaches 0.076 against fingerprints: 79% of DreaMS SSL’s 0.096, and above its agreement with ChemBERTa (0.040) or SMI-TED (0.045). MS-JEPA agrees least with every molecule representation (0.024–0.037) and with entropy similarity (0.052, versus DreaMS’s 0.130–0.142). CKA gives a different ordering: ChemBERTa–SMI-TED scores 0.535 despite mutual- k NN of 0.113, and MS-JEPA exceeds DreaMS contrastive against four of five molecule representations.

H The molecule-tower term

We adapt VA-VAE’s foundation-alignment loss [60]: cosine registration to MoLFormer through a learned linear projection (*mcos*), plus matching of within-batch pairwise similarities (*mdms*). Total weight is 1, *mdms* weight 0.25, and margins 0. Each intervention has a matched control. Encoders are then frozen and evaluated with the in-batch MoLFormer adapter; test-molecule differences are paired.

MS-JEPA. The term acts on h_{mol} . Control, *mcos*, and *mcos+mdms* use one GPU, batch 64, 160 epochs (40,800 steps), and final checkpoints. The control’s frozen MRR is 0.105, below the main encoder’s 0.121. MRR rises 0.105 $\rightarrow$ 0.150 $\rightarrow$ 0.192 and R@1 0.038 $\rightarrow$ 0.065 $\rightarrow$ 0.091: *mcos* adds 0.045 [0.026, 0.069], and *mdms* adds 0.042 [0.025, 0.060]. Mutual- k NN rises 0.029 $\rightarrow$ 0.052 $\rightarrow$ 0.064; silhouette changes 0.487 $\rightarrow$ 0.290 $\rightarrow$ 0.385.

DreaMS. The term acts on the contrastive stage’s pre-head precursor token through a 1,024 $\rightarrow$ 768 projection. Frozen retrieval changes little: MRR 0.227 $\rightarrow$ 0.230 (difference +0.003 [−0.012, 0.017]), R@1 0.124 $\rightarrow$ 0.119, and mutual- k NN 0.088 $\rightarrow$ 0.092. Same-molecule alignment changes 0.437 $\rightarrow$ 0.044, CKA 0.203 $\rightarrow$ 0.455, and silhouette 0.298 $\rightarrow$ 0.157.

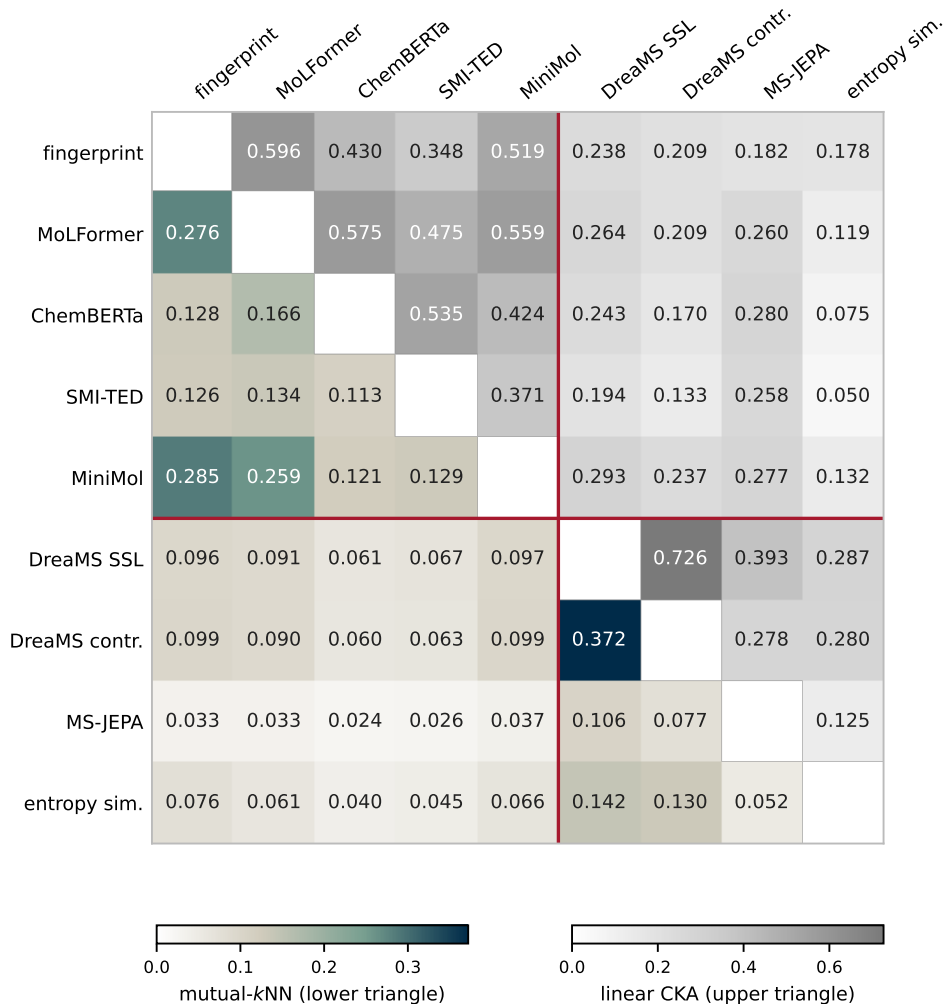


Figure 3: **Representational alignment across nine representations** on the 2,994 MassSpecGym test molecules covered by all of them. Lower triangle: mutual- k NN at $k=10$ (chance 0.003); upper triangle: linear CKA. Crimson lines separate molecule (top left) from spectrum (bottom right) representations; each triangle is colored on its own scale.

I Candidate-aware objectives

The two objectives differ only in the negatives. With a per-GPU batch $\{(s_i, m_i)\}_{i=1}^n$, $n = 64$, and $N_i = \{i\} \cup \{j : m_j \neq m_i\}$ (molecules compared by the first InChIKey block),

$$\mathcal{L}_{\text{batch}} = -\frac{1}{n} \sum_i \log \frac{\exp \rho_{\theta}(s_i, m_i)}{\sum_{j \in N_i} \exp \rho_{\theta}(s_i, m_j)}, \quad \mathcal{L}_{\text{cand}} = -\mathbb{E}_{(s, m^*)} \left[\log \frac{\exp \rho_{\theta}(s, m^*)}{\sum_{m' \in \mathcal{K}_{128}(s)} \exp \rho_{\theta}(s, m')} \right], \quad (4)$$

where $\mathcal{K}_{128}(s)$ is m^* and 127 molecules drawn uniformly without replacement from the query’s official list $\mathcal{K}(s)$ (at most 256), redrawn every step; a list with fewer than 128 entries (0.8% of training queries) is used whole.

Candidate construction. Mass-based lists are conditioned on a window around the true molecule’s mass, using the same procedure in training and evaluation. The concern here is the transferable difference between library structures and decoys.

The optimum. With K negatives drawn i.i.d. from a proposal r and the positive from $p(\cdot | s)$, the population minimizer of the InfoNCE risk is the log density ratio [51, 41]

$$\rho_r^*(s, m) = \log \frac{p(m | s)}{r(m)} + c(s) = \underbrace{\log \frac{p(m)}{r(m)}}_{A_r(m)} + \underbrace{\log \frac{p(s | m)}{p(s)}}_{\text{PMI}(s;m)} + c(s). \quad (5)$$

For independent in-batch negatives $r = p$, the prior term vanishes. For an independent candidate proposal $r = q \neq p$, it remains. This idealization explains the incentive to learn a library–decoy prior. Actual benchmark negatives are query-dependent and sampled without replacement. Property-matched decoys [18] reduce differences on the matched properties.

Blind controls. The primary control keeps the trained adapters and ranks each official list by $\bar{\rho}(m) = \mathbb{E}_s[\rho_\theta(s, m)]$ over the 194,119 training spectra, each with its own conditions. It uses neither query spectra nor query conditions. For global scores, $\bar{\rho}(m) = \tau^{-1} \langle \mathbb{E}_s[z_s], z_m \rangle$ exactly; this differs from adapting the mean encoder output because the adapter is nonlinear. For MaxSim, we estimate the expectation with 16,384 training spectra sampled without replacement.

The secondary control (global scores only) replaces the encoder output by its training mean but retains query adduct and collision energy at the adapter. We report $\Delta_0 = \text{MRR}_{\text{control}} - 0.024$ and $\Delta_s = \text{MRR}_{\text{full}} - \text{MRR}_{\text{control}}$.

Table 4: Spectrum-blind controls, frozen DreaMS contrastive, MassSpecGym test fold. Expected score: primary control (Table 1, left); mean embedding: secondary control, conditions kept.

Tower	Negatives	MRR	expected score			mean embedding		
			blind	Δ_0	Δ_s	blind	Δ_0	Δ_s
ChemBERTa	in-batch	0.147	0.020	−0.004	0.127	0.039	+0.015	0.108
ChemBERTa	candidate	0.211	0.147	+0.122	0.064	0.155	+0.131	0.056
MoLFormer	in-batch	0.227	0.018	−0.006	0.209	0.034	+0.009	0.193
MoLFormer	candidate	0.303	0.169	+0.144	0.135	0.160	+0.136	0.143

At fixed towers (Table 4), candidate training raises expected-score blind MRR by 0.151 [0.125, 0.178] for MoLFormer and 0.126 [0.105, 0.148] for ChemBERTa, versus full-MRR gains of 0.077 [0.059, 0.095] and 0.064 [0.049, 0.078]. Thus Δ_s falls by 0.074 [0.050, 0.097] and 0.063 [0.042, 0.082]. The blind gain accounts for 0.52 [0.45, 0.59] and 0.66 [0.56, 0.75] of full MRR above random.

The mean-embedding control gives similar candidate-training gains (0.127 [0.107, 0.148] and 0.115 [0.094, 0.136]). Its in-batch scores are slightly above chance ($\Delta_0 = 0.009$ [0.007, 0.012] and 0.015 [0.011, 0.020]), whereas expected-score controls are not (−0.006 [−0.010, 0.000] and −0.004 [−0.007, 0.001]).

I.1 Other candidate-trained models

JESTR [25] trains a spectrum MLP and a molecular-graph GCN from scratch with in-batch InfoNCE in both directions, $\mathcal{L}_{\text{NCE}}$, on cosine scores. For the last 3% of its epochs it adds a penalty on L entries of each training query’s mass candidate list, sorted by Tanimoto similarity to the query and visited in turn; JESTR-NR omits it. With unit spectrum embeddings $\bar{s}_i$ of a batch of n queries and unit embeddings $\bar{z}_{ij}$ of their listed decoys,

$$\mathcal{L} = (1 - w) \mathcal{L}_{\text{NCE}} + \frac{w}{nL} \sum_{i=1}^n \sum_{j=1}^L \langle \bar{s}_i, \bar{z}_{ij} \rangle, \quad w = 0.1. \quad (6)$$

We train both on our split with the official mass lists and the evaluation of App. D. The blind control is the expected score, $\langle \mu, \bar{z}_m \rangle$ with $\mu = \mathbb{E}_s[\bar{s}]$ over the training spectra (Table 5).

Published recipe. We use the public architecture and hyperparameters: spectrum dropout 0.4; GCN with BatchNorm, dropout 0.2, and max pooling; $\tau = 0.05$; Adam at 5×10^{-4} ; batch 32; and $L = 32$ decoys in a separate train-mode forward pass. Our deviations are all-adduct training for 1,000 epochs, one random spectrum per query per epoch, masking same-molecule negatives, removing the query

from its candidate list, and using the last checkpoint. The public code restricts adducts to $[M+H]^+$ and stops early on training loss. Both arms share the first 970 epochs.

The penalty changes full MRR by -0.012 $[-0.017, -0.007]$ and blind MRR by -0.033 $[-0.047, -0.020]$. Nine repeats of the last 30 epochs from a second run give -0.004 ± 0.001 full and $+0.005 \pm 0.002$ blind MRR (mean $\pm$ s.e.); three further runs give blind changes of $+0.011$, -0.007 , and -0.004 . The effect is small and inconsistent. Without the penalty, JESTR-NR has blind MRR 0.129 ($\Delta_0 = 0.105$), but 94% of this gain comes from 153 molecules with at least 20 spectra (39% of test spectra). Molecule-averaged blind MRR is 0.043 (random 0.024). Mean normalized spectrum embeddings have norms 0.14 and 0.13 ; using mean unnormalized embeddings instead gives blind MRR 0.037 and 0.036 .

Simplified recipe. We retain Eq. (6) but use a 1000–512–512–512 spectrum MLP and a three-layer, 512-wide GCN with mean pooling, neither with dropout nor BatchNorm. Training uses unmasked InfoNCE, $\tau = 0.07$, AdamW at 10^{-3} , weight decay 0.01 , and 60 epochs over all spectra. The penalty acts in the last two epochs on $L = 5$ of the 20 candidates most Tanimoto-similar to the query. JESTR-NR is at chance under the blind control. The penalty improves blind MRR in all five seeds, adding 0.056 ± 0.016 blind MRR (0.024 ± 0.003 per molecule) and 0.013 ± 0.002 full MRR.

At penalty onset, its gradient-norm ratio to $(1 - w)\mathcal{L}_{\text{NCE}}$ is 0.0059 in the published recipe versus 0.073 in the simplified recipe. This measures relative gradient strength before Adam’s adaptive scaling. Raising w to 0.58 restores the ratio; encoding decoys with positives in the same BatchNorm batch then gives $+0.068$ blind and $+0.020$ full MRR in three of three seeds. Separate batches allow BatchNorm statistics to distinguish positives from decoys and cause collapse in two of three seeds.

Why the penalty moves the blind scorer. Write $\bar{s}_i = \mu + \varepsilon_i$. The penalty splits into the listed decoys’ blind scores and a spectrum-specific remainder,

$$\frac{w}{nL} \sum_{i,j} \langle \bar{s}_i, \bar{z}_{ij} \rangle = \frac{w}{nL} \sum_{i,j} \langle \mu, \bar{z}_{ij} \rangle + \frac{w}{nL} \sum_{i,j} \langle \varepsilon_i, \bar{z}_{ij} \rangle,$$

so the negative gradient with respect to a decoy’s direction is proportional to $-(I - \bar{z}_{ij}\bar{z}_{ij}^\top)(\mu + \varepsilon_i)$: it points away from the listing spectrum, with a shared component along $-\mu$. Candidate InfoNCE has the same directional contribution, weighted by $p_\theta(m' | s)/\tau$ instead of $w/(nL)$. Under the independent-proposal idealization, its expected score equals $A_q(m') + \mathbb{E}_s[\text{PMI}(s; m')]$ up to a constant (Eq. (5)); the blind control therefore need not isolate A_q alone.

A test-set blind gain requires the molecule encoder to transfer this effect more strongly to decoys than to positives. In units of $\langle \mu, \bar{z} \rangle / \|\mu\|$, the simplified GCN lowers positive and decoy scores by 0.089 and 0.168 on training queries, and by 0.086 and 0.165 on test queries (gaps 0.079 and 0.080). The published GCN in the otherwise simplified recipe produces gaps of 0.024 on training and 0.001 on test, indicating little transfer of this separation.

Table 5: Effect of JESTR’s candidate-list penalty, MassSpecGym test fold. Change: JESTR minus JESTR-NR, averaged over paired runs (the published runs share their first 970 epochs). Blind: expected-score control (random 0.024). Reweighted: published recipe with $w = 0.58$ and the decoys in the positives’ BatchNorm batch.

Recipe	Runs	JESTR-NR		Change with penalty	
		MRR	Blind	MRR	Blind
Published	9	0.230	0.118	-0.004	$+0.005$
Simplified	5	0.163	0.029	$+0.013$	$+0.056$
Reweighted	3	0.235	0.118	$+0.020$	$+0.068$

J Improved baselines

Unless stated otherwise, all arms use frozen MoLFormer, in-batch negatives, 100 molecule epochs, validation-selected checkpoints, and seed 42 (App. D). Table 6 reports results.

Initialization and learning rates. DreaMS contrastive is unfrozen from the first step at 5×10^{-6} , 1.5×10^{-5} , or 5×10^{-5} (104.0M trainable parameters, versus 8.4M frozen). MS-JEPA (A) and (B)

use the same rates. Validation selects (A) at 5×10^{-5} (MRR 0.244); (B) at that rate has higher test MRR by 0.014 [0.001, 0.027]. For fresh encoders, validation selects 10^{-4} over 3×10^{-4} (0.235 versus 0.230). These selected rates are retained throughout Table 1.

Ablations. *MaxSim* uses the scoring rule in App. B.4. *Simulated pretraining* uses 46,000 steps of 256 simulated spectra, the shared optimizer, and warm-up and cosine decay over those steps. Real-data validation runs every 2,000 steps. The best checkpoint initializes the encoder, adapters, and temperature for real-data fine-tuning, with a fresh optimizer and schedule. Global and *MaxSim* scoring have separate pretraining runs. *Candidate training* replaces in-batch negatives during fine-tuning from the *MaxSim* pretraining checkpoint (Eq. (4)).

Table 6(b) reports successive changes. After simulated pretraining, pretrained MS-JEPA exceeds the fresh encoder by 0.012 [0.001, 0.024] MRR with *MaxSim* but trails by 0.024 [0.001, 0.049] with candidate training. Candidate training raises blind MRR from 0.013 to 0.083 (fresh) and 0.098 (pretrained).

Training length. Fresh encoders also run for 48 and 200 molecule epochs (4,656 and 19,400 steps), with global and *MaxSim* scoring. In all real-only end-to-end runs, training loss falls from $\ln 64 \approx 4.2$ to 0.02–0.07 (last 500 steps), while validation peaks before training ends. The 200-epoch runs peak at steps 4,800 (global, epoch 49) and 11,640 (*MaxSim*, epoch 120), then lose 0.031 and 0.020 validation MRR. Best validation MRR still rises with training length: 0.215, 0.235, 0.253 globally and 0.234, 0.240, 0.248 with *MaxSim* at 48, 100, 200 epochs. Test MRR shows no clear gain beyond 100 epochs. After simulated pretraining, fresh *MaxSim* validation peaks at the last step.

K Simulated spectra

The pretraining corpus of §6 is built in three steps: collect molecules and remove those near the test fold, simulate their spectra, and remove every molecule that the benchmark’s retrieval ranks.

Molecules. We pool MOSES [40], COCONUT [47], DSSTox [17], HMDB [57], MassSpecGym’s 3.92M PubChem structures [5], and the 2.50M pretraining structures of Liu et al. [36]: 7,627,340 distinct canonical isomeric SMILES. We remove, in order, structures that

1. share 2D connectivity with any MassSpecGym molecule (65,410);
2. have the Morgan or RDKit fingerprint of a test molecule (1,734);
3. cannot form $[M+H]^+$ or $[M+Na]^+$, or are unstable on the LC time scale (350,568), by rules that exclude no MassSpecGym molecule;
4. are not proven to lie at MCES distance ≥ 10 , MassSpecGym’s threshold, from every test molecule (265,749), using lower bounds or exact myopic MCES [29].

The 6,943,879 remaining structures and the 28,929 MassSpecGym molecules form 6.19M molecules (InChIKey first blocks).

Simulators. GLACIER [54] is the authors’ non-contrastive checkpoint, reported as trained on the official MassSpecGym split. FraGNNNet [63] is retrained on the training-fold spectra. Both models simulate $[M+H]^+$ spectra at eight normalized collision energies, the medians of equal-count bins of training $[M+H]^+$ NCE. Two simulators with different errors act as domain randomization [49].

Removing retrieval candidates. MassSpecGym molecules keep their folds; the others are split 98/1/1 by hash, and only the training split is used (6,064,935 molecules). From it we drop every validation or test query and every molecule in any of their mass- or formula-based candidate lists: 187,223 molecules and 3.48M spectra, 1,798 of them MassSpecGym training molecules. The remaining 5,877,712 molecules carry 98.2M spectra (52.1M GLACIER, 46.1M FraGNNNet).

Sim-to-real differences. Three mismatches are corrected. (i) GLACIER predicts square-root intensities and its spectra are too flat; we square them and drop peaks below 6.1×10^{-4} of the base peak, a floor fit to real peak counts on energy-matched validation pairs. Spectra then average 27.9 (GLACIER) and 29.9 (FraGNNNet) peaks, against 31.5 measured. (ii) The pooled molecules are heavier than MassSpecGym’s; reweighting by 25 Da precursor bins lowers the 1-Wasserstein distance to measured precursor masses from 38.4 (GLACIER) and 34.7 Da (FraGNNNet) to 0.94 Da, with bins absent from the measured data, almost all above 1,000 Da, given zero weight. Each simulator receives half the probability. (iii) Simulated m/z values are exact, whereas measured ones carry

Table 6: **Improved baselines**, MassSpecGym test fold, one seed each. Δ : paired MRR difference with a 95% interval from 1,000 bootstrap draws over test molecules (InChIKey-14). (a) Arms trained on real spectra; Δ against the reference named in each block. (b) One-change steps of Table 1 (right) and pretraining on simulated spectra alone (sim only), for the fresh encoder and the pretrained MS-JEPA (A) encoder.

(a)			
Arm	MRR	R@1	Δ MRR [95% CI]
<i>DreaMS contrastive, backbone LR; Δ vs. frozen</i>			
frozen	0.227	0.122	
5×10^{-6}	0.255	0.144	+0.028 [0.016, 0.041]
1.5×10^{-5}	0.255	0.136	+0.028 [0.016, 0.041]
5×10^{-5}	0.244	0.130	+0.018 [0.005, 0.029]
<i>MS-JEPA (A), aggregated, encoder LR; Δ vs. frozen</i>			
frozen	0.121	0.047	
5×10^{-6}	0.210	0.107	+0.089 [0.075, 0.104]
1.5×10^{-5}	0.236	0.124	+0.115 [0.098, 0.132]
5×10^{-5} (ladder)	0.243	0.128	+0.121 [0.105, 0.140]
<i>MS-JEPA (B), [CLS], encoder LR; Δ vs. frozen</i>			
frozen	0.170	0.077	
5×10^{-6}	0.219	0.116	+0.048 [0.038, 0.059]
1.5×10^{-5}	0.235	0.128	+0.065 [0.053, 0.078]
5×10^{-5}	0.257	0.143	+0.087 [0.070, 0.104]
<i>Fresh encoder, encoder LR; Δ vs. MS-JEPA (A) at 5×10^{-5}</i>			
1×10^{-4} (ladder)	0.232	0.124	-0.011 [-0.025, 0.003]
3×10^{-4}	0.212	0.111	-0.031 [-0.045, -0.017]
<i>Fresh encoder, training length; Δ vs. 100 epochs (0.232 global, 0.241 MaxSim)</i>			
global, 48 epochs	0.216	0.113	-0.016 [-0.029, -0.003]
global, 200 epochs	0.237	0.121	+0.005 [-0.011, 0.020]
MaxSim, 48 epochs	0.244	0.126	+0.002 [-0.009, 0.014]
MaxSim, 200 epochs	0.239	0.135	-0.003 [-0.018, 0.012]

(b)		
Step	fresh	pretrained
<i>ΔMRR [95% CI]</i>		
real: global $\rightarrow$ MaxSim	+0.010 [-0.008, 0.027]	-0.009 [-0.024, 0.006]
global: real $\rightarrow$ sim $\rightarrow$ real	+0.137 [0.117, 0.157]	+0.131 [0.112, 0.151]
sim $\rightarrow$ real: global $\rightarrow$ MaxSim	+0.094 [0.075, 0.116]	+0.102 [0.081, 0.123]
sim $\rightarrow$ real, MaxSim: in-batch $\rightarrow$ candidate	+0.103 [0.080, 0.128]	+0.067 [0.044, 0.088]
same step, spectrum-blind MRR	+0.070 [0.057, 0.085]	+0.085 [0.070, 0.102]
global: real $\rightarrow$ sim only	+0.005 [-0.019, 0.029]	+0.017 [-0.006, 0.038]
MaxSim: real $\rightarrow$ sim only	+0.105 [0.077, 0.134]	+0.110 [0.082, 0.141]
<i>sim only, MRR / R@1</i>		
global	0.236 / 0.140	0.260 / 0.158
MaxSim	0.347 / 0.251	0.344 / 0.249

ppm-scale error; the augmentation shared with measured spectra (peak dropout 0.2, log-normal intensity noise $\sigma = 0.3$, 5 ppm m/z jitter) blurs them. Other differences remain: one adduct, one instrument and eight energies, where measured spectra also include $[M+Na]^+$ and QTOF, and no noise or contaminant peaks.

Before fine-tuning. Without measured training spectra (measured validation spectra select checkpoints), global scoring matches real-only MRR (Table 6(b)), with higher R@1 (+0.016 [-0.005, 0.035] fresh, +0.029 [0.007, 0.050] pretrained) but lower R@20 (-0.067 [-0.113, -0.022], -0.069 [-0.111, -0.029]). MaxSim reaches 0.347 (fresh) and 0.344 (pretrained) MRR, 0.105 and 0.110 above real-only MaxSim training. Pretrained initialization helps global scoring (+0.023 [0.016, 0.031]) but not MaxSim (-0.002 [-0.008, 0.002]). Fine-tuning on measured spectra adds 0.114–0.132 MRR.