dION
FRAGMENTATION-BASED INVARIANCE FOR SELF-SUPERVISED LEARNING OF TANDEM MASS SPECTRA

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a novel invariance for peptide tandem mass spectrometry data, unlocking self-supervised representation learning that improves de novo sequencing of peptides. This invariance exploits the physical relationship between precursor properties (mass and charge) and fragment-ion evidence, without requiring peptide sequence labels. We introduce **dIon**, which adapts the DINO framework with two latent prediction tasks, both recovering a clean teacher representation: one from a spectrum mixture, using the precursor as a selection query, and one from a partial spectrum with the precursor withheld. The first associates precursor information with fragment-ion evidence; the second prevents representational collapse onto that information alone. Mechanistic probes support both effects, and ablations show that the full objective performs best. Under identical end-to-end training, dIon initialization improves de novo peptide precision over training from scratch by 5.5 and 8.4 percentage points on the held-out MassIVE-KB and Kingdoms test sets, and by 2.3 and 4.8 percentage points with a larger supervised training corpus. The resulting models surpass fully supervised state-of-the-art de novo sequencing models on the diverse, multi-species Kingdoms corpus under the same greedy-decoding protocol. Without peptide labels, dIon learns strong native peptide-similarity geometry compared with other learned models; with limited peptide-supervised adaptation, it achieves the best retrieval and pair-discrimination performance across all representation benchmarks.

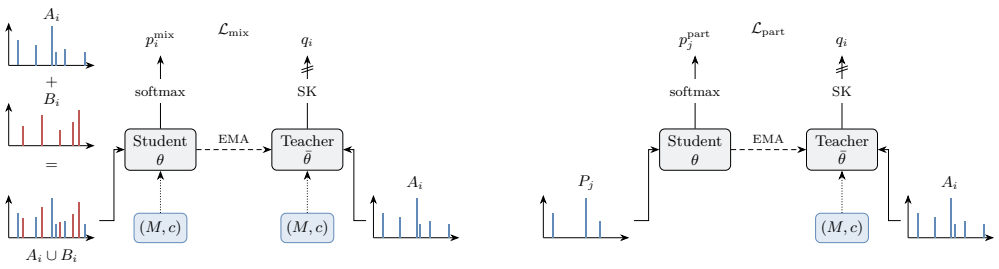
1 INTRODUCTION

Self-supervised learning has become a major approach for learning transferable representations without task labels, using training objectives constructed from the input data itself. Representative approaches include masked modeling (Devlin et al., 2019; He et al., 2022), contrastive learning across augmented views (Chen et al., 2020), and non-contrastive methods that match or predict learned targets across views (Caron et al., 2021; Baevski et al., 2022; Assran et al., 2023). The self-supervised objective determines which structure in the data the model must capture, making objective design especially important when moving to a new modality.

For view-based methods, the construction of the views is part of the self-supervised objective. In SimCLR, two augmented views of the same image define a positive pair, and the choice of augmentations strongly affects the learned representation (Chen et al., 2020). DINO trains a student to match an exponential moving average (EMA) teacher across differently augmented global and local views of the same image (Caron et al., 2021). The views thus define which changes the representation should become invariant to. For images, crops exploit the spatial structure of the pixel grid. The self-distillation machinery itself (student, EMA teacher, stop-gradient) is modality agnostic; the views, however, should express an invariance appropriate for that modality's data.

Peptide tandem mass spectrometry data has a very different structure. A mass spectrum is an unordered set of peaks, each given by its measured mass-to-charge ratio (m/z) and intensity. An intact peptide ion, the precursor, is measured in an MS1 scan and then fragmented, and its fragment ions are measured in an MS2 scan, producing the tandem MS/MS spectrum. Informative fragments may be absent, unrelated or noise peaks may be present, and fragment intensities can differ sub-

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.



(a) Precursor-conditioned mixture task

(b) Precursor-withheld partial-view task

Figure 1: Dual latent-prediction objective. (a) The student sees the mixture $A_i \cup B_i$ with the precursor (M, c) of A and matches the teacher output q_i for the clean view A_i ($\mathcal{L}_{\text{mix}}$). (b) The student sees a partial view P_j of A with the precursor withheld and matches the same target ($\mathcal{L}_{\text{part}}$). The teacher is an EMA of the student, and SK denotes Sinkhorn–Knopp balancing.

stantially between measurements, while the m/z values of detected ions are measured with high precision. An MS/MS spectrum is therefore a precise but incomplete and variable observation of the fragmentation of a precursor with measured mass and charge. Public repositories such as PRIDE (Perez-Riverol et al., 2025) hold vast quantities of MS/MS data, but only a minority of spectra carry peptide labels, pseudo-labels from database search assigned where a spectrum confidently matches a candidate peptide. Supervised training sees only this noisy, non-randomly selected subset, whereas self-supervision can use every spectrum.

Masked modeling would seem a natural first choice. For peptide MS/MS, however, the incompleteness and variability of spectra make the exact peaks and intensities removed by masking a less appealing target for representation learning. We instead turn to latent prediction, where the model predicts a learned target computed from another view rather than the masked observations themselves (Caron et al., 2021; Baevski et al., 2022; LeCun, 2022; Assran et al., 2023). Latent prediction still leaves the question of which relation should be preserved across views. For modalities without an obvious analogue of image crops or spatial masking, that relation can instead come from known structure in the data-generating or measurement process. One way to turn such a relation into a prediction task is to combine two data examples into a single view, provide conditioning information associated with one of them, and ask the model to recover that example's clean representation. Because the conditioning information could itself become a shortcut, a second view withholds it and presents only part of the example. We instantiate this idea for peptide MS/MS using the physical relationship between precursor properties and fragment-ion evidence, with no peptide sequence labels. We call this *fragmentation-based invariance*: the invariant relation is defined by fragmentation physics, while the representation remains sensitive to the fragmentation patterns that distinguish peptides.

We introduce dIon, which adapts DINO self-distillation to peptide MS/MS with two complementary latent-prediction tasks. Our main contributions are:

- We introduce a self-supervised latent-prediction construction that mixes two examples and conditions recovery of a clean representation on information associated with one source, paired with partial views that withhold this information so that it cannot serve as a shortcut.
- We instantiate this construction for peptide MS/MS as *fragmentation-based invariance*, through a *precursor-conditioned mixture task* and a *precursor-withheld partial-view task*, requiring no peptide sequence labels.
- We pretrain dIon as a foundation encoder for peptide MS/MS. Its global spectrum representation serves peptide retrieval and pair discrimination, and its dense peak-token representations transfer to de novo peptide sequencing, which predicts a peptide's amino-acid sequence directly from its spectrum.
- Mechanistic probes support the intended behaviors of both latent-prediction tasks, and objective ablations show that the full dual objective performs best.
- dIon initialization improves de novo peptide precision by 2.3–8.4 percentage points under matched end-to-end training and surpasses the fully supervised Casanovo and InstaNovo models

across all 70 species of the Kingdoms test set. To our knowledge, this is the first demonstration that self-supervised pretraining improves de novo sequencing over matched training from scratch. It also learns strong native peptide-similarity geometry compared with other learned models and improves label-efficient peptide-supervised metric learning.

2 Related Work

View-based self-supervised learning spans contrastive methods (Chen et al., 2020; He et al., 2020) and non-contrastive ones (Chen & He, 2021; Grill et al., 2020; Bardes et al., 2022; Caron et al., 2021); dIon builds on DINO's teacher–student self-distillation. A second family predicts learned targets from partial observations rather than reconstructing the input: data2vec predicts contextualized teacher representations from masked inputs (Baevski et al., 2022), I-JEPA predicts representations of target image regions from surrounding context (Assran et al., 2023), and Masked Siamese Networks match a masked view to a target computed from the unmasked image (Assran et al., 2022). Dense downstream tasks have motivated explicit local objectives, from DenseCL's local-feature contrast (Wang et al., 2021) to iBOT's masked patch-token prediction (Zhou et al., 2022), retained in DINOv2 (Oquab et al., 2024), and Gram anchoring in DINOv3 (Siméoni et al., 2025).

Mixing-based methods train on mixtures of inputs (Lee et al., 2021; Shen et al., 2022; Alex et al., 2025), but with targets that reflect both sources. dIon also mixes two examples, yet leaves the target unblended: conditioning information associated with one source specifies whose clean teacher output is wanted, and its partial views withhold that information so it cannot serve as a shortcut.

For mass spectra, self-supervision has been explored mainly on small molecules, combining masked-peak prediction with retention-order prediction (Bushuiev et al., 2026) or contrastive learning between masked views (Zhang et al., 2024). Peptide-MS/MS representations have instead been learned with peptide supervision: GLEAMS embeds spectra from database-search assignments (Bittremieux et al., 2022) and is our supervised reference in Section 4.3, Sanders et al. (2025) transfer an encoder pretrained by supervised de novo sequencing, and others learn from identified peptide-spectrum matches or jointly with sequence prediction (Altenburg et al., 2026; Jin et al., 2024; Zhao et al., 2026). Closest to our setting is InstaNovo-FM, concurrent work that pretrains an encoder to reconstruct masked regions of spectra without peptide-sequence labels (Nieuwoudt et al., 2026); on their benchmark the authors report faster convergence but no measurable de novo advantage over training the same architecture directly. dIon learns from teacher targets rather than reconstructing the spectrum, and its initialization improves de novo sequencing over matched training from scratch (Section 4.2).

De novo peptide sequencing has used deep learning since DeepNovo (Tran et al., 2017), and modern Transformer models encode a spectrum into contextualized peak representations that a peptide decoder attends over. Casanovo established this formulation (Yilmaz et al., 2022; 2024), InstaNovo and InstaNovo+ follow it under autoregressive and diffusion-based decoding (Eloff et al., 2025), and Pairwise Attention biases the encoder's attention with pairwise mass differences between peaks (Lapin et al., 2025). That interface is the one dIon's dense representations must be compatible with.

3 Methods

3.1 Spectrum representation and encoder

A tandem mass spectrum is an unordered set of N fragment-ion observations,

$$S = \{(m_i, I_i)\}_{i=1}^{N}, \tag{1}$$

where m_i is the measured m/z of peak i and I_i its intensity, together with precursor information (M, c), where M is the neutral precursor mass and c the discrete precursor charge. Spectra are preprocessed by restricting peaks to $m/z \leq 2500$, retaining the 200 most intense of those, and normalizing intensities by the base peak.

Each peak is embedded from sinusoidal encodings of its m/z and intensity, in the manner commonly used to encode scalar quantities such as position (Vaswani et al., 2017) and standard across Transformer models for tandem mass spectra (Yilmaz et al., 2024; Eloff et al., 2025; Yang et al.,

2024; Lapin et al., 2025) (Appendix B.1). Peaks are distinguished by their continuous m/z rather than by a positional encoding over discrete indices. Precursor information is encoded as a single token formed by summing a sinusoidal encoding of M with a learned embedding of the discrete charge c, and is prepended to the peak tokens.

The spectrum encoder f_θ is a 15-block Transformer of width 1024 with approximately 127M parameters, returning one contextualized representation h_i per input token. We pool these into a global spectrum representation z:

$$\{h_i\}_{i=0}^{N} = f_\theta\big(S, (M, c)\big), \qquad z = \mathrm{Pool}\big(\{h_i\}_{i=0}^{N}\big). \tag{2}$$

Here, Pool is a learned single-query cross-attention operation applied after the final Transformer block. We call h_i the dense peak-token representations and z the global representation, borrowing the distinction between global and dense features drawn by DenseCL (Wang et al., 2021). Remaining architectural details are given in Appendix B.1.

3.2 Self-distillation

dIon is trained by self-distillation, following DINO (Caron et al., 2021; Oquab et al., 2024). Each input is presented to the networks as several views. A view is a certain altered presentation of an input; in DINO augmented crops of an image. A student network must reproduce, from one view, what an exponential moving average (EMA) teacher produces from another view of the same input, without labels. Each network pools its encoder output into a single representation z_v per view v, which a three-layer MLP head g maps to a K-dimensional output. Normalizing that output gives the assignment of the view, and the student's assignment for one view must match the teacher's for another:

$$p_v = \mathrm{softmax}\big(g_\theta(z_v)/\tau_s\big), \qquad q_v = \mathrm{SK}\big(g_{\bar\theta}(z_v)/\tau_t\big), \tag{3}$$

where θ are the student weights and $\bar\theta$ their exponential moving average, τ_s and τ_t are temperatures, and SK denotes Sinkhorn–Knopp balancing of the teacher assignments across the batch, which DINOv2 (Oquab et al., 2024) also adopts from SwAV, in place of the softmax centering used by DINO (Caron et al., 2021). The student is trained by minimizing the cross-entropy $H(q_v, p_{v'})$ between the teacher assignment of one view and the student assignment of another.

The K outputs are equivalently read as scores against K learned vectors in the head's last layer, the prototypes of SwAV and DINOv2 (Caron et al., 2020; Oquab et al., 2024). SK balancing and the low teacher temperature together prevent output collapse, in which the teacher returns the same assignment regardless of the input: balancing spreads a batch's assignment mass across the K outputs, while sharpening by τ_t keeps the assignments from flattening to a uniform distribution, and either mechanism alone encourages the failure the other prevents. No gradient flows through the teacher branch, whose weights change only through the EMA update.

For dIon we used $K = 16{,}384$. Arriving at schedules that train stably on spectra required tuning τ_t and the EMA rate, monitored through the entropy of the teacher assignments, the teacher–student KL, and the across-batch variance of the student assignments (Appendix F.1); the selected values and the remaining hyperparameters are given in Appendix B.2. Avoiding constant output collapse is necessary but not sufficient, since a stable configuration need not learn a useful representation.

3.3 Fragmentation-based invariance

The remaining question is which invariance to impose. It should be one that cannot be satisfied trivially, but require the model to capture structure useful for downstream transfer. The views define this invariance through what they preserve, alter, or remove. DINO constructs views using image transformations that exploit spatial structure, whereas tandem mass spectra are unordered sets of peaks. Peak m/z values cannot be jittered freely because they are measured to within a few parts per million. Intensities vary more across measurements and can therefore be perturbed more naturally, but they carry substantially less information than peak m/z values. We instead derive the views from known structure in the measurement process, instantiated here by the physical relationship between precursor properties and fragment-ion evidence.

For each training example, we call the original spectrum the anchor A. All views are then constructed from A, with an additional mixture view adding peaks from a competing spectrum B.

Clean teacher views. Clean views A_i retain 95% of the peaks of A, sampled independently. They are encoded with the true precursor of A, and provide the targets that every student view must recover.

Precursor-conditioned mixture views. Each clean view A_i is combined with the peaks of an independently sampled competing spectrum B_i, giving the student input $A_i \cup B_i$. The two peak sets are merged rather than concatenated: peaks coinciding within 5 ppm are combined into a single peak with summed intensity. A mixture always combines the anchor with a single competing spectrum. The true anchor precursor is retained, so it indicates which of the two peptides present in the mixture should be represented. For this to be well posed, each mixture view samples its competing spectrum uniformly from the spectra in the batch whose neutral precursor mass differs from the anchor's by more than 10 ppm, which also excludes the same peptide observed at another charge.

Precursor-withheld partial views. Partial views P_j are formed by independently sampling 60% of the peaks of A (swept over $\{30, 60\}\%$, Appendix E.8). Their precursor token is replaced by a learned null embedding, so the student receives no precursor information for these views. The peak restriction is conceptually related to DINO's local crops (Caron et al., 2021): in both cases the student must match a target computed from a more complete teacher view. The constructions themselves differ, since a DINO local crop is a spatially contiguous region of an image, whereas a partial spectrum is a random subset of peaks spanning the full m/z range. Withholding the precursor is a further restriction with no image-domain counterpart, and its purpose is to deny the shortcut: a representation that relied on precursor mass alone would be unable to recover the teacher target here.

3.4 Dual latent-prediction objective

Figure 1 summarizes the objective. The teacher encodes the clean views, producing the targets. The student encodes the mixture views and the partial views, and must recover those same targets: the mixture task requires the student to focus on the signals corresponding to its input precursor while ignoring those of the interfering spectrum, while the partial-view task must recover the target from only a fraction of the anchor's peaks with no precursor at all.

Following Equation 3, we write p_i^{mix} for the student assignment obtained from $A_i \cup B_i$, p_j^{part} for the student assignment obtained from P_j under the null precursor state, and q_i for the teacher assignment obtained from the clean view A_i. In the mixture-task term, a mixture view with the anchor's precursor is matched to a clean teacher target with that same precursor,

$$\mathcal{L}_{\mathrm{mix}} = \operatorname*{avg}_{i} H\big(q_i, p_i^{\mathrm{mix}}\big), \tag{4}$$

whereas each partial view is matched to every clean target,

$$\mathcal{L}_{\mathrm{part}} = \operatorname*{avg}_{i,\,j} H\big(q_i, p_j^{\mathrm{part}}\big), \tag{5}$$

The two prediction tasks receive equal weight,

$$\mathcal{L}_{\mathrm{dIon}} = \tfrac{1}{2}\,\mathcal{L}_{\mathrm{mix}} + \tfrac{1}{2}\,\mathcal{L}_{\mathrm{part}}, \tag{6}$$

where

$$H(q, p) = -q \log p \tag{7}$$

is the cross-entropy between a teacher assignment q and a student assignment p.

The objective is defined on z, and therefore the gradient update is computed through the learned query-pooling of equation 2. This contrasts with DINO, which generates z from a [CLS] token. Thus, the dense representations $\{h_i\}$ are trained by the same objective, and dIon uses no separate dense-token loss of the kind introduced by iBOT (Zhou et al., 2022) and used in DINOv2 (Oquab et al., 2024); Appendix F reports an ablation that adds an iBOT-style dense loss term.

We use two views of each kind. Two is the most that fits in the memory of one GH200 96 GB accelerator at this model scale, and we did not sweep the count. DINO reports that more views help, with diminishing returns (Caron et al., 2021). The head follows the DINOv2 design; Appendix B.2 gives its configuration together with the temperature and EMA schedules.

3.5 DOWNSTREAM TRANSFER

For de novo sequencing, the pretrained encoder is coupled to a freshly initialized autoregressive Transformer decoder (Vaswani et al., 2017), following the encoder–decoder formulation established for peptide MS/MS by Casanovo (Yilmaz et al., 2022; 2024): the decoder cross-attends to the encoder's peak-token memory $\{h_i\}$, so de novo sequencing consumes the dense representations rather than the pooled global one, and both networks are optimized jointly under teacher-forced peptide-token cross-entropy. Spectrum-level binary tasks instead apply a small multilayer perceptron to z, and peptide-supervised metric learning applies a freshly initialized projection head to z trained with a supervised contrastive objective (Khosla et al., 2020). Downstream architectures, optimization, and decoding settings are given in Appendix D and Appendix E.

4 EXPERIMENTS AND RESULTS

4.1 PRETRAINING AND DEVELOPMENT PROTOCOL

dIon is pretrained without peptide labels on the PXD010000 bacterial corpus of Lee et al. (2022) (9,234,199 spectra), holding out a run-disjoint subset of 467,723 spectra for development (Appendix A). The 127M-parameter encoder is trained for 284 epochs on 64 GH200 accelerators in about 75 hours (Appendix B.1). Following DINO, we evaluate and release the EMA teacher. Checkpoints are selected on development data and decoding settings are fixed in advance. Test splits are used only for final evaluation (Appendix B.4).

4.2 DE NOVO SEQUENCE TRANSFER

All models, including the external baselines, are scored with one matcher, which follows the criterion of DeepNovo (Tran et al., 2017) as implemented by Casanovo (Yilmaz et al., 2022): treating a peptide as a sequence of amino-acid tokens, a predicted token matches when its mass and its prefix mass lie within 0.1 and 0.5 Da of the aligned reference token. Peptide precision at full coverage is the fraction of all test spectra whose predicted peptide matches at every token, with spectra lacking a prediction counted as errors (Appendix E.7).

Kingdoms evaluation. dIon initialization improves peptide precision over training from scratch in both training regimes (Table 1). Fine-tuned on dIon-de-novo-labeled-v1 (DNL-v1), a larger multi-source labeled corpus (Appendix A.5), it also exceeds InstaNovo 1.2.2 and Casanovo 5.2.1 across all 70 species (Table 25). When each model ranks its predictions by its own confidence (Figure 6), dIon is the most precise at every coverage level, tied with InstaNovo only at the highest confidence.

Table 1: De novo sequencing on the Kingdoms de novo (all 70 species) and MSKB-final test sets. Within each training regime, scratch and dIon initialization use an identical end-to-end training protocol.

	Peptide precision	
Training / initialization	Kingdoms	MSKB-final
MSKB-final training		
Scratch, end-to-end	0.449	0.747
dIon init., end-to-end	0.533	0.802
dIon-de-novo-labeled-v1 training		
Scratch, end-to-end	0.554	0.819
dIon init., end-to-end	**0.602**	**0.842**
External reference		
Casanovo 5.2.1	0.492	0.783
InstaNovo 1.2.2	0.570	—

MSKB-final evaluation. The same holds on MSKB-final: dIon initialization improves over scratch in both training regimes, and every dIon-initialized model exceeds Casanovo 5.2.1 (Table 1). InstaNovo 1.2.2 is omitted there because it was trained on MassIVE-KB v1, the source of this test set, and its splits are unpublished.

Larger peak cap. For applications where prediction quality matters more than compute, we also provide dIon-initialized models fine-tuned with a 1,000-peak cap. Although the encoder is pre-trained with at most 200 peaks per spectrum, the larger cap improves peptide precision further on both benchmarks. Trained on DNL-v1, it reaches 0.607 on Kingdoms and 0.856 on MSKB-final (Tables 24 and 26).

Unlabeled sequencing on yeast MS2 data. We additionally evaluate on a public yeast run without peptide labels, using proteome matching with a decoy-based false-discovery control (Appendix E.3). On emitted predictions, the three models fall within about one percentage point of each other (Table 23).

4.3 Peptide-representation geometry

Retrieval. Each spectrum queries the other spectra of its benchmark, and a retrieved spectrum is relevant when it carries the query's peptide. Over the full gallery, precursor mass and charge alone separate most candidates, so a high *broad* retrieval score does not necessarily show that a representation has learned anything from the fragment ions. In practice these metadata are known for every spectrum and are used to prefilter candidates, as GLEAMS does by charge and precursor mass before clustering (Bittremieux et al., 2022). What a representation must then contribute is to tell apart the remaining near-isobaric peptides, which differ only in their fragment-ion patterns. We therefore evaluate retrieval within such a prefilter, restricting the gallery to spectra with the query's charge and a precursor m/z within 10 ppm, and report mean average precision (mAP) in this *strict* setting (Appendix C).

Pair discrimination. Strict pairs are a fixed, balanced set of same-peptide and different-peptide pairs under the same constraint, shared by all models. Retrieval asks whether each query ranks its own peptide first. Pair ROC-AUC asks whether one similarity scale separates same-peptide from different-peptide pairs across the benchmark, as clustering with a single threshold requires. Table 2 reports both strict metrics.

Without peptide labels, frozen dIon exceeds InstaNovo-FM on every benchmark and metric, and Casanovo 4.0, trained with full peptide supervision for de novo sequencing, on four of the six (Table 2). GLEAMS (Bittremieux et al., 2022), trained supervised end to end for peptide similarity on about 30 million labeled MassIVE-KB v1 spectra, scores highest of these models and is the only one of them to match the binned spectral angle. Supervised metric learning changes the ranking. It fine-tunes the encoder with a fresh projection head and a supervised contrastive loss over same-peptide positives, using 1% or 10% of the training peptides (Appendix D.2). With 10% of the labels, dIon initialization is best on every benchmark and metric, and with 1% it already exceeds GLEAMS on the Bacterial and Kingdoms benchmarks. At both label budgets it matches or improves on scratch initialization.

Table 2: Representation geometry across test benchmarks (Kingdoms: all 70 species).

	Bacterial		NineSpecies		Kingdoms	
Method	Retrieval mAP	Pairs AUROC	Retrieval mAP	Pairs AUROC	Retrieval mAP	Pairs AUROC
Binned spectral angle	0.915	0.943	0.975	0.954	0.905	0.871
Peptide-supervised						
Casanovo 4.0	0.881	0.898	0.932	0.859	0.838	0.746
GLEAMS	0.916	0.948	0.976	0.962	0.903	0.876
Self-supervised						
InstaNovo-FM	0.847	0.864	0.874	0.763	0.805	0.708
dIon	0.895	0.928	0.926	0.865	0.821	0.748
Peptide-supervised contrastive						
Scratch init. (1%)	0.909	0.934	0.972	0.946	0.906	0.873
dIon init. (1%)	0.922	0.953	0.972	0.953	0.911	0.887
Scratch init. (10%)	0.928	0.960	**0.978**	0.957	0.930	0.902
dIon init. (10%)	**0.933**	**0.968**	**0.978**	**0.965**	**0.940**	**0.926**

4.4 MECHANISTIC PROBES

Two probes on the frozen encoder test the behaviours the dual objective was designed to induce (Table 3). In the first, a fixed mixture $A \cup B$ is embedded under the precursor of A or of B, and a trial is correct when the representation lies closer to the clean representation of the conditioned source than to that of the other. In the second, two random 60% peak subsets of A are embedded with the precursor withheld, and a view is correct when it lies closer to the clean representation of A than to that of a different-peptide spectrum. Appendix C.2 gives other readouts.

Table 3: Source recovery by dIon on the NineSpecies *validation* set, by cosine similarity of the global representation.

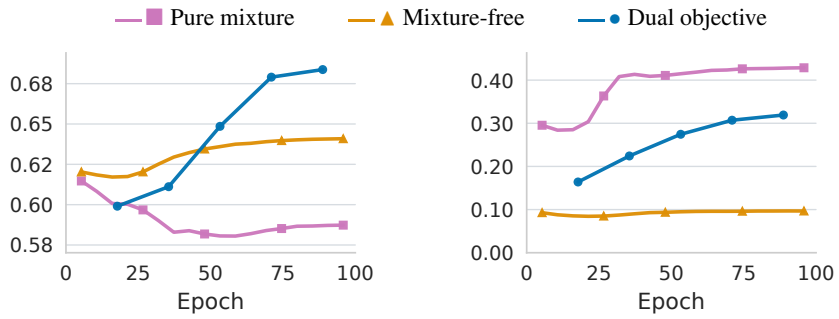
	Correct-source accuracy		
Probe	First	Second	Both
Mixture $A \cup B$, precursor of A / B	0.997	0.929	0.926
Partial views P_1 / P_2 of A, null precursor	0.987	0.984	0.978

Conditioning on the anchor precursor selects the anchor almost always. The competing precursor also selects its own source in the great majority of trials, though less often, since it is capped at the anchor's peak count (Equation 13) and is never itself a target. The goal of dIon is to learn representations rather than to serve as a final decomposition model. One aimed at decomposition would drop the cap and add a symmetric term.

4.5 OBJECTIVE ABLATIONS

We compare the dual objective of Section 3 against two baselines that each retain one half of it. All three are pretrained for only 100 epochs under otherwise identical settings. The pure mixture objective uses only the mixture views. The mixture-free control uses only clean views and swaps which views receive the precursor, so that the global views must infer it from the fragments (Appendix C.1). Neither matches the dual objective.

The pure mixture objective reaches the highest broad retrieval mAP but the weakest strict pair discrimination, and its strict pair AUROC falls as training proceeds (Figure 2). The same ordering holds on an auxiliary probe of the tryptic C-terminal residue (Appendix C.1). This is the representational collapse onto precursor information that the precursor-withheld partial views are designed to prevent: without them, the representation organizes spectra by the precursor metadata that broad retrieval rewards (Section 4.3). The mixture-free control avoids this collapse, and the dual objective, which adds the mixture views, pulls clearly ahead on strict pairs in the second half of pretraining.



(a) Strict pair AUROC

(b) Broad retrieval mAP

Figure 2: *Validation* probes during pretraining of the 100-epoch ablation checkpoints. Strict is defined in Section 4.3.

De novo transfer. With the encoder frozen, the dual objective ranks first in validation peptide precision (0.675), ahead of the pure mixture objective (0.638), the mixture-free control (0.610) and a randomly initialized encoder (0.582). Appendix C.1 also reports end-to-end fine-tuning.

5 DISCUSSION

The self-distillation machinery of DINO transfers to peptide tandem mass spectra unchanged; what does not transfer is the view construction built on image geometry. The fragmentation-based invariance is that adaptation: it derives the invariant relation from the structure of the measurement, and the dual objective imposes it without labels. The failure mode it avoids is not necessarily specific to this modality. Conditioning information supplied alongside an observation is a shortcut, and an objective that always provides it will be solved using it, much as SimCLR needs color distortion because crops of one image can be matched by their color histograms alone (Chen et al., 2020). Withholding the precursor in a second view family denies the analogous shortcut here, and neither branch alone reproduces the full result (Figure 2).

Unlike masked reconstruction, which gave no measurable de novo advantage in InstaNovo-FM (Nieuwoudt et al., 2026), predicting learned teacher targets improves de novo sequencing over matched training from scratch, consistent with our argument that reconstructing incomplete and variable peaks is a poor pretraining target for peptide MS/MS. Although the objective is defined only on the pooled global representation, the dense per-token representations benefit a decoder that attends over them, and an explicit token-level objective of the kind used by iBOT and DINOv2 did not help (Appendix F). We attribute this to the pooling and the partial views. Because z is pooled from the token outputs by learned-query attention rather than read from a [CLS] token, the objective acts on the tokens themselves, and because a partial view holds only a random subset of peaks, the tokens must stay informative without the full spectrum. Gram anchoring, which DINOv3 introduced against dense-feature degradation (Siméoni et al., 2025), helps here too: anchoring to a 30% partial-view checkpoint with stronger dense probes improves frozen but not fine-tuned de novo transfer (Appendix E.9). The dense information is therefore present without anchoring and is recovered when the encoder is adapted.

Frozen, the training-free binned spectral angle outperforms every learned representation not trained for peptide similarity, including the peptide-supervised Casanovo encoder (Table 2). With supervised metric learning on 10% of the labels, dIon initialization outperforms every method on every benchmark and metric, GLEAMS included. On three auxiliary spectrum-level tasks, fine-tuning from dIon likewise exceeds training from scratch on all three, and every other evaluated representation on spectrum quality and methionine oxidation. Only on these auxiliary tasks does its frozen representation trail both Casanovo and InstaNovo-FM, and on spectrum quality only negligibly (Appendix D.1).

A limitation we share with the field is that our evaluations score predictions against peptide labels assigned by a search engine, so they inherit its errors and selection bias. Only pretraining is free of these labels. The one exception is the yeast evaluation, which instead checks whether each predicted peptide occurs in the yeast proteome (Appendix E.3). Pretraining at this scale is also costly (Appendix B.3). Although not trained as a decomposition model, dIon can decompose a mixture given each component's precursor, so it could potentially be used for data-independent acquisition (DIA), where wide isolation windows co-fragment several peptides. Appendix E.6 reports a pilot on real DIA spectra.

Because the objective needs no labels, pretraining can in princple draw on every acquired spectrum, including the majority that search engines leave unassigned.

6 CONCLUSION

View-based self-distillation can be carried to a new modality by deriving its views from a relation that the measurement itself supplies. For peptide tandem mass spectra that relation is the physical link between a precursor and the fragments it produces, and imposing it needs no peptide labels. Under identical end-to-end training, dIon initialization improves de novo peptide precision over training from scratch by between 2.3 and 8.4 percentage points across both test sets and both supervised training corpora, and exceeds the supervised state of the art on Kingdoms de novo. Without labels it learns strong native peptide-similarity geometry compared with other learned models, and as an initialization for peptide-supervised metric learning it exceeds training from scratch at both label budgets and, with 10% of the labels, outperforms every compared method. With dIon pretraining, useful global and dense representations emerge from unlabeled spectra alone.

AI USE STATEMENT

We used generative AI tools in this work and describe their role below. All AI-assisted output was reviewed by the authors, and we take responsibility for the final content of this paper, including its text, claims, figures and released artifacts.

Research ideation and experimental design. The contribution of this paper, the fragmentation-based invariance and the dual latent-prediction objective that imposes it, was conceived by the authors. We used large language models, through Anthropic's Claude Code and OpenAI's Codex, to help design some of the auxiliary spectrum-level tasks reported in Section 4 and Appendix D.1. The authors selected, specified and validated the final task definitions.

Code. The model, training and evaluation code was written by the authors. We used the same tools to refactor, document and debug that code, and to implement dataset format conversions and preprocessing. All AI-assisted code was reviewed and tested by the authors, and every reported number comes from a run of the reviewed code.

Writing. We used the same tools for drafting and revision, including reorganizing sections, condensing text to the page limit and rewriting passages, and for language correction alongside Grammarly. Every sentence of the submitted manuscript was read and approved by the authors, and no text was included that the authors had not checked against the underlying results.

Related work. Most of the work we cite was known to us from our own research in these fields. In a few cases, we used the same tools to find related work.

We did not use generative AI to produce, alter or select experimental results, and no AI system is an author of this work.

ETHICS STATEMENT

Every corpus used here was collected, approved and publicly released by an earlier study, and we accessed each one through its public repository under the terms of that release. We collected no new samples and performed no experiments on biological material. Some corpora derive from human samples; we use only the released fragment spectra, we make no attempt to recover donor identity, and we add no information that the original depositors did not publish. Accessions and the originating publications are given in Appendix A so that credit follows the data.

We see a wider benefit in this mode of work. Public repositories already hold far more tandem mass spectra than database search can confidently identify, and the unassigned majority is ordinarily discarded. Learning without peptide labels extracts further value from measurements that have already been made and paid for, rather than requiring new acquisition, and the resulting encoder is reused across several downstream tasks. Against that we set the cost of pretraining, which used 64 accelerators for approximately three days; we report that cost in Appendix B and note that it is incurred once and amortized over the transfer experiments reported here and any later reuse of the released model.

De novo peptide sequencing is a general analytical capability, and we are not aware of a misuse pathway specific to this work beyond those already present in computational proteomics.

REPRODUCIBILITY STATEMENT

The pretraining objective and both latent-prediction tasks are specified in Section 3 and Appendix B.2. Encoder architecture, optimization, and the full training configuration are given in Appendix B, with compute in Appendix B.3. Dataset manifests, split construction and the filtering rules separating PXD010000 from every evaluation corpus are in Appendix A. The evaluation protocol, readouts and decoding settings are in Appendices C and E.7, and the handling of the external de novo models in Appendix E.4. Anonymous code is available at `https://anonymous.4open.science/r/bedrock-6CD8`. On acceptance, we will publicly release the pretrained dIon encoder and the fine-tuned de novo checkpoints.

REFERENCES

Kumar K. Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-ReQ: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17626–17638. Curran Associates, Inc., 2022. doi: 10.52202/068431-1281. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/70596d70542c51c8d9b4e423f4bf2736-Abstract-Conference.html.

Tony Alex, Sara Atito, Armin Mustafa, Muhammad Awais, and Philip J. B. Jackson. SSLAM: Enhancing self-supervised models with audio mixtures for polyphonic soundscapes. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=odU59TxdiB.

Tom Altenburg, Thilo Muth, Patrick van Zalm, Hanno Steen, and Bernhard Y. Renard. Foundation model enables interpretable open and error-tolerant searching for mass spectrometry-based proteomics. *Bioinformatics*, 42(Supplement_1):btag297, 2026. doi: 10.1093/bioinformatics/btag297.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked Siamese networks for label-efficient learning. In *European Conference on Computer Vision (ECCV)*, pp. 456–473, 2022. doi: 10.1007/978-3-031-19821-2_26.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, June 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/baevski22a.html.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Wout Bittremieux. MassIVE-KB v1 30 million PSMs training/validation/test splits, 2025.

Wout Bittremieux, Damon H. May, Jeffrey Bilmes, and William Stafford Noble. A learned embedding for efficient joint analysis of millions of mass spectra. *Nature Methods*, 19:675–678, 2022. doi: 10.1038/s41592-022-01496-1.

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

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, 2020.

Xinlei Chen and Kaiming He. Exploring simple Siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, June 2021.

Vadim Demichev, Christoph B. Messner, Spyros I. Vernardis, Kathryn S. Lilley, and Markus Ralser. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, 17(1):41–44, 2020. doi: 10.1038/s41592-019-0638-x.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186, 2019.

Kevin Eloff, Konstantinos Kalogeropoulos, Amandla Mabona, Oliver Morell, Rachel Catzel, Esperanza Rivera-de Torre, Jakob Berg Jespersen, Wesley Williams, Sam P. B. van Beljouw, Marcin J. Skwark, Andreas Hougaard Laustsen, Stan J. J. Brouns, Anne Ljungars, Erwin M. Schoof, Jeroen Van Goey, Ulrich auf dem Keller, Karim Beguir, Nicolas Lopez Carranza, and Timothy P. Jenkins. InstaNovo enables diffusion-powered de novo peptide sequencing in large-scale proteomics experiments. *Nature Machine Intelligence*, 7:565–579, 2025. doi: 10.1038/s42256-025-01019-5.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/garrido23a.html`.

Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, Ulf Reimer, Hans-Christian Ehrlich, Stephan Aiche, Bernhard Kuster, and Mathias Wilhelm. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*, 16: 509–518, 2019. doi: 10.1038/s41592-019-0426-7.

Vladimir Gorshkov, Julia A. Bubis, Elizaveta M. Solovyeva, Mikhail V. Gorshkov, and Frank Kjeldsen. Protein corona formed on silver nanoparticles in blood plasma is highly selective and resistant to physicochemical changes of the solution. *Environmental Science: Nano*, 6(4):1089–1098, 2019. doi: 10.1039/C8EN01054D.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21271–21284, 2020.

Bobby He and Mete Ozay. Exploring the gap between collapsed & whitened features in self-supervised learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8613–8634. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/he22c.html`.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Stephane Houel, Robert Abernathy, Kutralanathan Renganathan, Karen Meyer-Arendt, Natalie G. Ahn, and William M. Old. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *Journal of Proteome Research*, 9(8):4152–4160, 2010. doi: 10.1021/pr1003856.

Zhi Jin, Sheng Xu, Xiang Zhang, Tianze Ling, Nanqing Dong, Wanli Ouyang, Zhiqiang Gao, Cheng Chang, and Siqi Sun. ContraNovo: A contrastive learning approach to enhance de novo peptide sequencing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 144–152, 2024. doi: 10.1609/aaai.v38i1.27765.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 18661–18673, 2020.

Dongwan Kim, Young-suk Lee, Soo-Jin Jung, Jinah Yeo, Jenny J. Seo, Young-Yoon Lee, Jaechul Lim, Hyeshik Chang, Jaewon Song, Jihye Yang, Jong-Seo Kim, Guhung Jung, Kwangseok Ahn, and V. Narry Kim. Viral hijacking of the TENT4–ZCCHC14 complex protects viral RNAs via mixed tailing. *Nature Structural & Molecular Biology*, 27(6):581–588, 2020. doi: 10.1038/s41594-020-0427-3.

Joel Lapin, Alfred Nilsson, Mathias Wilhelm, and Lukas Käll. Pairwise attention: Leveraging mass differences to enhance de novo sequencing of mass spectra. *Journal of Proteome Research*, 24 (7):3722–3730, 2025. doi: 10.1021/acs.jproteome.5c00063.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview, 2022. URL `https://openreview.net/forum?id=BZ5a1r-kVsf`. Version 0.9.2, released June 27, 2022.

Joon-Yong Lee, Hugh D. Mitchell, Meagan C. Burnet, Ruonan Wu, Sarah C. Jenson, Eric D. Merkley, Ernesto S. Nakayasu, Carrie D. Nicora, Janet K. Jansson, Kristin E. Burnum-Johnson, and Samuel H. Payne. Uncovering hidden members and functions of the soil microbiome using *De Novo* metaproteomics. *Journal of Proteome Research*, 21(8):2023–2035, 2022. doi: 10.1021/acs.jproteome.2c00334.

Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. i-Mix: A domain-agnostic strategy for contrastive representation learning. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=T6AxtOaWydQ`.

Cathena Meiling Li, Jaemin Kang, Jongyeon Baek, Youbin Kim, Heemin Park, and Yong-Keun Jung. Cytosolic FKBPL and ER-resident CKAP4 co-regulates ER-phagy and protein secretion. *Nature Communications*, 15, 2024. doi: 10.1038/s41467-024-52188-7.

JinA Lim, Namseok Lee, Seonmin Ju, Jeesoo Kim, Subin Mun, Moonhyeon Jeon, Yong-ki Lee, Seok-Hoon Lee, Jayoung Ku, Sujin Kim, Sangsu Bae, Jong-Seo Kim, and Yoosik Kim. Cellular dsRNA interactome captured by K1 antibody reveals the regulatory map of exogenous RNA sensing. *Communications Biology*, 8, 2025. doi: 10.1038/s42003-025-07807-4.

Carlo Melendez. Data for "accounting for digestion enzyme bias in Casanovo", 2024.

Carlo Melendez, Justin Sanders, Melih Yilmaz, Wout Bittremieux, William E. Fondrie, Sewoong Oh, and William Stafford Noble. Accounting for digestion enzyme bias in Casanovo. *Journal of Proteome Research*, 23(10):4761–4769, 2024. doi: 10.1021/acs.jproteome.4c00422.

Annette Michalski, Jürgen Cox, and Matthias Mann. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *Journal of Proteome Research*, 10(4):1785–1793, 2011. doi: 10.1021/pr101060v.

Johannes B. Müller, Philipp E. Geyer, Ana R. Colaço, Peter V. Treit, Maximilian T. Strauss, Mario Oroshi, Sophia Doll, Sebastian Virreira Winter, Jakob M. Bader, Niklas Köhler, Fabian Theis, Alberto Santos, and Matthias Mann. The proteome landscape of the kingdoms of life. *Nature*, 582(7813):592–596, 2020. doi: 10.1038/s41586-020-2402-x.

Mechiel Nieuwoudt, Marco Reverenna, Divanisha Patel, Rachel Catzel, Isaac H. J. Houngue, Jemma Daniel, Kevin Eloff, Alberto Santos, Nicolas Lopez Carranza, Timothy P. Jenkins, Jeroen Van Goey, and Konstantinos Kalogeropoulos. Learning from tandem mass spectra at scale with a self-supervised foundation model for proteomics. *bioRxiv*, 2026. doi: 10.64898/2026.09.03.747733.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Yasset Perez-Riverol, Chakradhar Bandla, Deepti J. Kundu, Selvakumar Kamatchinathan, Jingwen Bai, Suresh Hewapathirana, Nithu Sara John, Ananth Prakash, Mathias Walzer, Shengbo Wang, and Juan Antonio Vizcaíno. The PRIDE database at 20 years: 2025 update. *Nucleic Acids Research*, 53(D1):D543–D553, 2025. doi: 10.1093/nar/gkae1011.

Justin Sanders, Melih Yilmaz, Jacob H. Russell, Wout Bittremieux, William E. Fondrie, Nicholas M. Riley, Sewoong Oh, and William Stafford Noble. Foundation model for mass spectrometry proteomics. *arXiv preprint arXiv:2505.10848*, 2025.

Zhiqiang Shen, Zechun Liu, Zhuang Liu, Marios Savvides, Trevor Darrell, and Eric Xing. Un-Mix: Rethinking image mixtures for unsupervised visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2216–2224, 2022. doi: 10.1609/aaai.v36i2.20119.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences*, 114(31):8247–8252, 2017. doi: 10.1073/pnas.1705691114.

Anton Tsitsulin, Marina Munkhoeva, and Bryan Perozzi. Unsupervised embedding quality evaluation. In Timothy Doster, Tegan Emerson, Henry Kvinge, Nina Miolane, Mathilde Papillon, Bastian Rieck, and Sophia Sanborn (eds.), *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*, volume 221 of *Proceedings of Machine Learning Research*, pp. 169–188. PMLR, 28 Jul 2023. URL https://proceedings.mlr.press/v221/tsitsulin23a.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Mingxun Wang, Jian Wang, Jeremy Carver, Benjamin S. Pullman, Seong Won Cha, and Nuno Bandeira. Assembling the community-scale discoverable human proteome. *Cell Systems*, 7(4):412–421.e5, 2018. doi: 10.1016/j.cels.2018.08.004.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3024–3033, 2021.

Tingpeng Yang, Tianze Ling, Boyan Sun, Zhendong Liang, Fan Xu, Xiansong Huang, Linhai Xie, Yonghong He, Leyuan Li, Fuchu He, Yu Wang, and Cheng Chang. Introducing π-HelixNovo for practical large-scale de novo peptide sequencing. *Briefings in Bioinformatics*, 25(2):bbae021, 2024. doi: 10.1093/bib/bbae021.

Melih Yilmaz, William Fondrie, Wout Bittremieux, Sewoong Oh, and William S Noble. De novo mass spectrometry peptide sequencing with a transformer model. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25514–25522. PMLR, 17–23 Jul 2022.

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.

Hailiang Zhang, Qiong Yang, Ting Xie, Yue Wang, Zhimin Zhang, and Hongmei Lu. MSBERT: Embedding tandem mass spectra into chemically rational space by mask learning and contrastive learning. *Analytical Chemistry*, 96(42):16599–16608, 2024. doi: 10.1021/acs.analchem.4c02426.

Jiale Zhao, Peng-Zhi Mao, Kaifei Wang, Yi-Ming Li, Ya-Ping Peng, Ranfei Chen, Shuqi Lu, Xiaohong Ji, Jiaxiang Ding, Xin Zhang, Yucheng Liao, Weinan E, Han Wen, Wei-Jie Zhang, and Hao Chi. A large-scale unified deep learning model for peptide mass spectrum interpretation trained on multimodal data. *Nature Machine Intelligence*, 8:997–1011, 2026. doi: 10.1038/s42256-026-01234-8.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.

Daniel P. Zolg, Mathias Wilhelm, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Bernard Delanghe, Derek J. Bailey, Siegfried Gessulat, Hans-Christian Ehrlich, Maximilian Weininger, Peng Yu, Judith Schlegl, Karl Kramer, Tobias Schmidt, Ulrike Kusebauch, Eric W. Deutsch, Ruedi Aebersold, Robert L. Moritz, Holger Wenschuh, Thomas Moehring, Stephan Aiche, Andreas Huhmer, Ulf Reimer, and Bernhard Kuster. Building ProteomeTools based on a complete synthetic human proteome. *Nature Methods*, 14(3):259–262, 2017. doi: 10.1038/nmeth.4153.

A Datasets, splits, and provenance

A.1 Overview of datasets and roles

Table 4 lists every dataset with its role, split sizes and split criterion.

Table 4: Datasets used in the principal experiments, with spectrum counts per split and the split criterion.

Dataset	Use	Train	Validation	Test	Split
PXD010000	SSL pretraining	9,234,199	467,723	—	Acquisition project / run
PXD010613	Retrieval	—	51,770	518,804	Run held out; exact sequence filtered; development from PXD010000 validation
NineSpecies	Retrieval	—	6,119	182,393	Reserved test; test peptides excluded from development
Kingdoms	Retrieval	—	658,738	4,112,590	Acquisition batch; run-disjoint where available; 70 species
MSKB-final	De novo	1,989,996	10,004	200,000	Casanovo split; canonical peptidoform
DNL-v1	De novo training	5,275,467	170,140	345,414	Global peptidoform separation
Kingdoms de novo	De novo test	—	—	4,926,232	Run-disjoint; all 70 species, at most 100k spectra each
PXD026806	Label-free proteome	—	—	39,460	Single yeast acquisition run
SQA	Spectrum quality	167,778	117,082	151,256	Released project split
HYE	Chimericity	79,940	45,886	47,411	HYE partition
Bacterial chim.	Chimericity transfer	276,834	92,603	287,362	Bacterial species
OxMet	Oxidized Met	135,926	60,511	181,479	Bacterial species

A.2 SSL PRETRAINING AND DEVELOPMENT CORPUS

dIon is pretrained on the PXD010000 bacterial corpus (Lee et al., 2022), which covers the proteomes of 51 bacterial organisms and comprises 9,234,199 charge-1–10 training spectra. The held-out 467,723-spectrum subset is used for development. The underlying peptide-spectrum matches provide corpus construction metadata, but the SSL objective receives only peaks and precursor information: peptide sequence, identity, PSM assignment, and fragment annotation are withheld.

A.3 NATIVE-GEOMETRY EVALUATION DATASETS

We construct the bacterial native-geometry benchmark from PXD010000 validation spectra for development and PXD010613 spectra for held-out testing. The derived test retrieval population contains 518,804 spectra and is run-disjoint from the PXD010000 pretraining data, and test spectra whose exact modified peptide sequence occurs in PXD010000 training or validation are removed. The two corpora were released together as the training and test sets of a de novo model for soil metaproteomics (Lee et al., 2022), and they share no organism: PXD010613 covers four bacterial species absent from the 51 organisms of PXD010000.

Table 5: Held-out retrieval spectra. The charge-2–4 column is the matched subset used for external-reference comparisons.

Corpus	Full-charge	Charge-2–4	Split definition
PXD010613	518,804	499,617	PXD010613 runs are held out from PXD010000 development data; exact modified sequences overlapping PXD010000 train/validation are removed.
NineSpecies	182,393	181,852	Reserved fixed paper test; development excludes normalized peptide backbones appearing in this test.
Kingdoms	4,112,590	4,087,062	For 10 multi-batch species, validation holds out one deterministic acquisition batch; test uses remaining files plus all other species. Run-disjoint where batch structure exists, not universally project-disjoint.

NineSpecies V2 (MassIVE accession MSV000090982; Yilmaz et al., 2024) evaluates nine organisms using separate development and reserved evaluation populations; the latter contains 182,393 retrieval spectra. Peptide backbones assigned to the reserved population are excluded from development using uppercase amino-acid normalization. The Kingdoms proteome landscape (Müller et al., 2020) provides a larger cross-species setting. Its retrieval, pair and de novo test sets all cover the same 70 species, with 4,112,590 retrieval spectra in the held-out population. Its validation and test sets are separated by acquisition run but are intentionally not peptidoform-disjoint.

Retrieval and static-pair evaluation are separate views of each corpus. Mass-controlled comparisons require equal precursor charge and precursor mass within 10 ppm. GLEAMS comparisons use deterministic charge-2–4 subsets on which every method has an embedding; these are reported separately from the charge-complete populations. The restriction is small: the charge-2–4 cohort retains 96.3% of PXD010613, 99.7% of NineSpecies, and 99.4% of Kingdoms, or 99.1% of the three benchmarks combined.

A.4 SUPERVISED METRIC-LEARNING DATA

Metric learning uses peptide identities from the bacterial corpus after SSL pretraining. Label budgets are sampled at the peptide-identity level, with all eligible spectra for a selected identity retained; the 1% subset is nested within the 10% subset. Evaluation populations and the supervised contrastive head are held fixed across scratch and dIon initialization, isolating the contribution of SSL pretraining.

A.5 De novo datasets and split construction

The MSKB-final de novo corpus is the released two-million-PSM tryptic/non-tryptic MassIVE-KB corpus used to train Casanovo 4.2 and 5.0 (Melendez et al., 2024; Melendez, 2024). Tryptic examples originate from MassIVE-KB v1, while the multi-enzyme component comes from the corresponding MassIVE-KB v2.0.15 subset. We preserve the 200,000-spectrum Casanovo test split and draw a seed-42 validation subset of 10,004 spectra from the remaining data. Sampling is grouped by canonical peptidoform without charge, preventing train–validation peptidoform overlap.

The dIon-de-novo-labeled-v1 (DNL-v1) corpus broadens the peptide, enzyme and charge coverage of MSKB-final with 5,791,021 labeled spectra from eight sources (Table 6). It contains all of MSKB-final, with its test split kept as test and a validation subset carved out of its training split. Synthetic human peptides come from ProteomeTools Parts I–III (Zolg et al., 2017; Gessulat et al., 2019), where Part III adds HLA class I and II and non-tryptic peptides. Their PSMs are kept at a posterior error probability of at most 0.01, with reverse and contaminant hits removed. MassIVE-KB v2.0.15 (Wang et al., 2018) contributes PSMs at 1% FDR from public MassIVE datasets, with chimeric spectra removed. The PXD010000 bacterial corpus contributes its database-search PSMs, and the non-enzymatic Casanovo training set (Yilmaz et al., 2024) contributes MassIVE-KB and PROSPECT PSMs balanced over C-terminal residues. A high-charge collection raises the share of precursors with charge 5 or more. It combines PXD053100, PXD018061, PXD007648 and PXD034980 (Lim et al., 2025; Kim et al., 2020; Gorshkov et al., 2019; Li et al., 2024) with additional high-charge spectra that will be released upon acceptance. The added sources are subsampled with per-charge quotas and, for all but the high-charge collection, a cap of 3,000 spectra per run. The corpus contains no Kingdoms data, and it shares no acquisition run with the Kingdoms de novo test set.

Peptidoforms outside MSKB-final are split 90/5/5 into training, validation and test. A global test–validation–train precedence rule then assigns each canonical peptidoform to one split, and exact peak-list duplicates are collapsed. We checked the finished corpus for peptidoform overlap: its splits share no canonical peptidoform, and no MSKB-final test peptidoform occurs in its training or validation splits. The duplicate collapse removes eight spectra from the MSKB-final test component, so the DNL-v1 test set is closely related to, but not identical to, the released MSKB-final test split.

Table 6: Sources of DNL-v1, with spectra per split and precursor-charge range.

Source	Train	Validation	Test	Charges
MSKB-final	1,974,956	25,002	199,992	2–5
PXD010000 bacterial corpus	863,005	48,273	48,073	1–10
ProteomeTools Part III (PXD021013)	788,303	37,086	37,426	1–7
ProteomeTools Part I (PXD004732)	508,466	12,662	12,939	2–7
MassIVE-KB v2.0.15	416,182	17,891	17,472	1–8
ProteomeTools Part II (PXD010595)	382,417	16,881	16,576	2–6
Casanovo non-enzymatic set	223,952	6,838	7,044	1–7
High-charge collection	118,186	5,507	5,892	2–7
Total	5,275,467	170,140	345,414	1–10

Kingdoms capped de novo evaluation. We derive the Kingdoms run-disjoint de novo test set by applying a deterministic species-stratified cap to the existing Kingdoms run-disjoint test split. The original test population has 14,519,567 spectra across 70 species; the cap retains 4,926,232 spectra (33.93%). Each species contributes at most 100,000 spectra: 33 species are capped and smaller species are retained in full. Within species, each spectrum receives a pseudo-random number, the BLAKE2b hash of its species, peak file, scan ID and source-row index with seed 20260915, and the 100,000 spectra with the smallest numbers are retained. This gives a uniform random sample that is deterministic and independent of row order. Validation is copied unchanged. The resulting cohort remains run-disjoint and is not peptidoform-disjoint.

Casanovo 4.0 and MSKB-final serve different purposes in this study. The Casanovo 4.0 encoder is the representation reference examined in Casanovo Foundation (Yilmaz et al., 2024; Sanders et al., 2025). It was trained on 30 million MassIVE-KB v1 PSMs, retaining up to 100 PSMs per peptidoform–charge precursor (Bittremieux, 2025). Casanovo 4.0 also provides a separately trained non-enzymatic model based on MassIVE-KB v1 and PROSPECT. By contrast, our matched super-

vised de novo experiments train on MSKB-final, the smaller mixed-enzyme corpus associated with Casanovo 4.2 and 5.0, and on DNL-v1.

A.6 AUXILIARY DOWNSTREAM DATASETS

We derive the spectrum-quality task from a 19-project human Orbitrap HCD collection, using a class-balanced charge-1–10 subset. Chimericity labels are derived from FragPipe/MSFragger wide-window searches at 1% PSM FDR: the primary mixture task uses the PXD024584 HYE mixture corpus, while a second task uses the PXD010613 bacterial corpus with species-disjoint splits. From the same bacterial source corpus, we construct an oxidized-methionine task on methionine-containing peptides.

B ARCHITECTURE AND TRAINING CONFIGURATION

B.1 MODEL AND SSL CONFIGURATION

The dIon encoder is a 15-block Transformer with model dimension $d = 1024$ and approximately 127.3M parameters. Layer normalization is applied after each residual branch, and dropout of 0.15 is applied to the attention and feed-forward residual-branch outputs, with no dropout on the attention weights themselves. Table 7 summarizes the configuration.

Table 7: dIon spectrum encoder configuration.

Component	Setting
Transformer blocks	15
Model dimension d	1024
Attention heads	8 (128 dimensions per head)
Feed-forward network	Linear(1024, 2048) $\rightarrow$ ReLU $\rightarrow$ Linear(2048, 1024)
Normalization	Post-residual LayerNorm (attention and FFN branches)
Dropout	0.15 (attention and FFN residual outputs)
Parameters	$\approx$127.3M
m/z encoding	Four components, 256-D sinusoidal each (wavelengths 1–500); 1024 total
Intensity encoding	256-D sinusoidal (wavelengths 10^{-6}–1)
Peak projection	Bias-free Linear(1280, 1024)
Precursor mass encoding	1024-D sinusoidal (wavelengths 10^{-3}–10^{4})
Charge embedding	Learned Embedding(11, 1024), charges 0–10
Precursor token	Sum of mass encoding and charge embedding, prepended
Collision energy	Not used
[CLS] token	None
Global pooling	One learned 1024-D query, 8-head cross-attention
Pooling scope	Non-padding encoder tokens

B.1.1 PEAK PREPROCESSING

Every spectrum supplied to the model is preprocessed identically: peaks are restricted to $0 \leq m/z \leq 2500$; if more than 200 peaks remain, the 200 most intense are retained and intensities are base-peak normalized,

$$I'_i = \frac{I_i}{\max_j I_j + 10^{-12}}, \tag{8}$$

where the maximum runs over the retained peaks. Spectra are then zero-padded within each batch and supplied with a padding mask. This preprocessing is shared by dIon, the scratch and randomly initialized controls, the metric-learning fine-tunes, the auxiliary-task models, and the in-house de novo models. External reference models use their own released preprocessing. The de novo models fine-tuned with a 1,000-peak cap retain the 1,000 most intense peaks instead.

B.1.2 PEAK AND PRECURSOR ENCODING

Peak features concatenate sinusoidal encodings of m/z and base-peak- normalized intensity. The m/z value m_i is decomposed into four numeric components: the 100-Da bin, the integer remainder within that bin, and the first and second pairs of fractional decimal digits,

$$d_1 = \left\lfloor \tfrac{m_i}{100} \right\rfloor, \quad d_2 = \lfloor m_i \rfloor - 100\, d_1, \quad d_3 = \left\lfloor \tfrac{r_i}{100} \right\rfloor, \quad d_4 = r_i - 100\, d_3, \tag{9}$$

with $r_i = \mathrm{round}\big(10^4\,(m_i - \lfloor m_i \rfloor)\big)$. Each component receives a 256-dimensional sinusoidal encoding ϕ_{256} over wavelengths 1–500, giving a 1024-dimensional m/z representation, and the base-peak-normalized intensity I_i' receives a separate 256-dimensional encoding over wavelengths 10^{-6}–1. The resulting 1280-dimensional peak feature is mapped to the model width by a bias-free linear projection,

$$p_i = W_{\mathrm{peak}}\big[\phi_{256}(d_1);\ \phi_{256}(d_2);\ \phi_{256}(d_3);\ \phi_{256}(d_4);\ \phi_{256}(I_i')\big], \qquad W_{\mathrm{peak}} \in \mathbb{R}^{1024\times 1280}. \tag{10}$$

The projected 1024-dimensional peak tokens enter the 15 Transformer blocks directly.

The precursor token is

$$c_{(M,c)} = \phi_{1024}(M) + \mathrm{Emb}(c) \in \mathbb{R}^{1024}, \tag{11}$$

where the neutral precursor mass M receives a 1024-dimensional sinusoidal encoding over wavelengths 10^{-3}–10^4 and Emb is a learned $\mathrm{Embedding}(11, 1024)$ over charges 0–10. No sequence-position encoding is added; the model relies on the physical m/z encoding rather than peak-list position. The null precursor embedding used by the precursor-withheld partial views is obtained by setting the neutral mass input to zero and the charge to the reserved index 0; charge 0 never occurs for a real spectrum, so its row of the charge embedding is learned exclusively for this state.

The encoder exports two representations. The dense representations $\{h_i\}_{i=1}^{N}$ are the contextualized peak-token outputs and form the encoder memory consumed by de novo decoders. The global representation is produced by learned-query pooling with a single learned query $u \in \mathbb{R}^{1024}$ and 8-head cross-attention over the non-padding encoder tokens,

$$z = \mathrm{MHA}\big(u,\ \{h_i\}_{i=0}^{N},\ \{h_i\}_{i=0}^{N}\big) \in \mathbb{R}^{1024}. \tag{12}$$

B.1.3 VIEW CONSTRUCTION

Table 8 gives the view multiplicities and mixture settings of the selected configuration. Each anchor spectrum yields two clean teacher global crops, two precursor-conditioned mixture views built from the same two crops, and two precursor-withheld partial views. The replicate count of two per kind was not swept or ablated: it is the largest multi-view configuration that fits in the memory of a single GH200 96 GB accelerator at this model scale, and DINO reports that additional views improve representations with diminishing returns (Caron et al., 2021).

Table 8: View construction and objective weighting for the selected dIon configuration. Peak fractions are uniform random subsets of the anchor spectrum.

View	Count	Construction
Clean teacher global	2	95% anchor subset, true precursor
Mixture view	2	95% anchor subset + equal-size competing subset, true precursor
Partial view	2	60% anchor subset, reserved null precursor (mass 0, charge 0)
Mixture settings		
Mixing strength	1.0 (competing subset matches anchor subset peak count)	
Competing spectrum	Independently sampled per view from the same batch	
Eligibility	>10 ppm separation under same-charge precursor m/z and neutral mass	
Peak merging	Within 5 ppm after mixing; summed intensity, intensity-weighted m/z	
Intensity renormalization	None after mixing	
Objective weighting		
Mixture task	0.5	
Partial-view task	0.5	
Prototypes K	16,384	

Mixture construction. For an anchor spectrum A, a clean crop A_i retaining 95% of its peaks is formed and an eligible in-batch competing spectrum B is selected. The competing contribution is explicitly size-matched to the anchor crop:

$$n_B = \min\big(n_A,\, N_B\big) \tag{13}$$

peaks are sampled uniformly without replacement from B, where $n_A = |A_i|$ and N_B is the number of peaks available in B. At the selected mixing strength of 1.0 this is exactly one competing peak per anchor-crop peak whenever B holds at least n_A peaks, which prevents unusually dense competing spectra from dominating a mixture through peak count alone. The mixed view is the ppm-aware union

$$M_i = \mathrm{merge}_{5\,\mathrm{ppm}}\big(A_i \cup B_i\big), \tag{14}$$

formed by concatenating the two peak sets, sorting by m/z, and merging connected adjacent peaks within 5 ppm. Each merged group takes summed intensity and intensity-weighted m/z,

$$I_{\mathrm{merged}} = \sum_{\ell} I_\ell, \qquad (m/z)_{\mathrm{merged}} = \frac{\sum_\ell I_\ell\,(m/z)_\ell}{\sum_\ell I_\ell}, \tag{15}$$

falling back to the arithmetic mean of m/z when the total intensity is zero. Peaks separated by more than 5 ppm remain distinct, so a mixed view contains at most $2n_A$ peaks before merging, and no intensity renormalization is applied afterwards. Competing spectra are drawn uniformly from the batch, excluding the anchor itself and any spectrum whose neutral precursor mass lies within 10 ppm of the anchor's. The implementation additionally excludes same-charge spectra whose precursor m/z lies within 10 ppm; because neutral mass is determined by precursor m/z and charge, this second criterion excludes almost nothing beyond the neutral-mass one at equal charge, and is retained as a conservative guard on the precursor representation the model actually receives rather than as an independent chemical constraint. A fresh competing spectrum and peak subset are drawn for each mixture view. If no eligible spectrum is available the view remains anchor-only, which does not occur in practice at the batch sizes used here.

B.2 Self-supervised objective and optimization

B.2.1 Prototype assignments and self-distillation targets

The pretraining projection head maps a pooled representation to prototype logits through

$$\mathrm{Linear}(1024, 2048) \to \mathrm{GELU} \to \mathrm{Linear}(2048, 256) \to \ell_2 \to \text{WN-Linear}(256, 16{,}384; \text{no bias}), \tag{16}$$

where ℓ_2 denotes ℓ_2 normalization of the bottleneck feature and WN denotes weight normalization. The final layer maps the 256-dimensional ℓ_2-normalized bottleneck to $K = 16{,}384$ outputs, and because it carries no bias its K learned output weight vectors are the prototypes of SwAV and DINOv2 (Caron et al., 2020; Oquab et al., 2024): output k is the inner product of the bottleneck feature with prototype k. The head's final weight scale remains trainable, and the head is discarded for all downstream evaluation. Writing g_θ and $g_{\bar\theta}$ for the student and teacher head applied to a pooled representation, student assignments are temperature-scaled softmax distributions,

$$p = \mathrm{softmax}\big(g_\theta(z)/\tau_s\big), \tag{17}$$

while teacher assignments are balanced across the batch by Sinkhorn–Knopp normalization of $\exp\big(g_{\bar\theta}(z)/\tau_t\big)$ with three iterations, following DINOv2 (Oquab et al., 2024),

$$q = \mathrm{SK}\big(g_{\bar\theta}(z)/\tau_t\big). \tag{18}$$

Teacher assignments are detached, so no gradient flows through the teacher branch, and after each optimizer step the teacher weights are updated as $\bar\theta \leftarrow m\,\bar\theta + (1-m)\,\theta$. The teacher temperature τ_t warms linearly from 0.03 to 0.07 over the first ten epochs and is then held constant, the student temperature is $\tau_s = 0.1$, and the momentum m is cosine-scheduled from 0.998 to 1.0 over training. The selected model sets both the masked-token (iBOT) and KoLeo loss weights to zero and applies no token masking, so Equation 6 is the complete pretraining objective, and collapse is avoided by the Sinkhorn–Knopp balancing and the teacher sharpening alone.

B.2.2 OPTIMIZATION

The learning rate is warmed up linearly over 40,000 optimizer steps and then cosine-decayed over a further 450,000 steps, with annealing applied per optimizer step. Batch-size-dependent learning-rate scaling is disabled, so the base learning rate is used directly. Weight decay is constant. Pretraining runs on 64 GH200 96 GB accelerators, 4 per node across 16 nodes, with a per-device batch of 128 and no gradient accumulation, giving an effective global batch of 8,192 spectra. Table 9 lists the settings.

Table 9: Self-supervised pretraining optimization settings.

Setting	Value
Base learning rate	8.0×10^{-5}
Learning-rate scaling	Disabled
Warmup	40,000 optimizer steps (linear, from 0)
Decay	Cosine over 450,000 steps to 1.0×10^{-6}
Annealing granularity	Per optimizer step
Weight decay	0.04 (constant)
Batch size per device	128
Devices	16 nodes $\times$ 4 GH200 96 GB (64 total)
Gradient accumulation	None
Effective global batch	8,192
Numerical precision	bf16 mixed
Training horizon	300 epochs nominal, stopped after 284
Checkpoint cadence	Every 10 epochs
Teacher EMA momentum	Cosine, $0.998 \rightarrow 1.0$
Teacher temperature τ_t	$0.03 \rightarrow 0.07$ over 10 epochs, then constant
Student temperature τ_s	0.1
Teacher centering	Sinkhorn–Knopp, 3 iterations

B.3 COMPUTE AND TRAINING EFFICIENCY

The canonical pretraining run used 64 GH200 96 GB accelerators, 4 per node across 16 nodes, in bf16 mixed precision with an effective global batch of 8,192 spectra (Table 9). Allocated device memory was stable at roughly 90% of the 96 GB available on each accelerator for the duration of training. Completing the 284 trained epochs took about 75 hours of wall-clock time, approximately 4,800 accelerator-hours. These figures describe the canonical run alone. The ablations run shorter horizons under separate configurations.

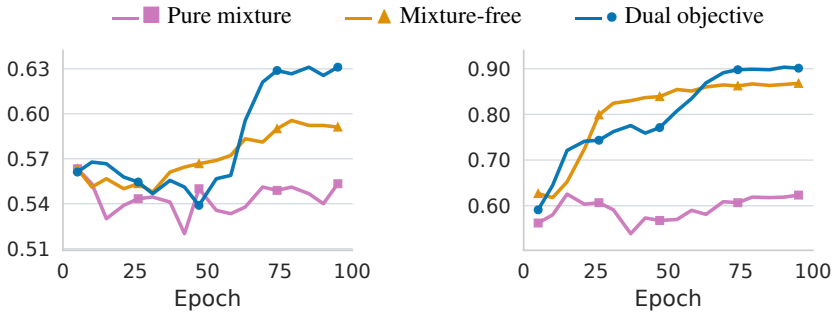
B.4 MODEL SELECTION

The released pretraining checkpoint, epoch 284 of the canonical run, was selected manually by inspecting the online validation probes of the objective ablations: strict pair discrimination and broad retrieval (Figure 2) and the end-AA probes (Appendix C.1). The teacher temperature and momentum schedules were chosen with the collapse diagnostics of Appendix F.1. Main-text similarity results use cosine similarity of the global representation. For the mechanistic probes, Appendix C.2 also reports Euclidean distance and the Jensen–Shannon divergence of the projection-head outputs. De novo checkpoints are selected by maximum peptide precision on the validation split of their training corpus: 10,004 spectra for MSKB-final and 170,140 for DNL-v1. De novo training and decoding hyperparameters, including greedy decoding, the 200-peak cap and the 50 ppm precursor tolerance, were fixed in advance and not tuned in this work (Appendix E.7). Test splits were used only for final evaluation.

C EMBEDDING EVALUATION AND ADDITIONAL ANALYSES

C.1 OBJECTIVE ABLATIONS

The end-AA probe is a small binary diagnostic tracked during development: tryptic peptides end in K or R, and the probe predicts which from the representation, by nearest neighbours or a linear classifier.



(a) End-AA kNN

(b) End-AA linear probe

Figure 3: End-AA probes during pretraining of the 100-epoch ablation checkpoints, complementing Figure 2. The dual objective leads both probes in the second half of pretraining.

The three runs are matched approximately 100-epoch runs on the same bacterial corpus, architecture, optimizer, schedule, preprocessing and 200-peak cap, evaluated with the same compact validation probe suite: strict pair AUROC and broad retrieval mAP (Section 4.3, Figure 2) and the end-AA probes (Figure 3). The probes read out the pooled representation of the EMA teacher on fixed, compact full-charge subsets rather than the full validation sets. The canonical run evaluates them every 12,000 optimizer steps. The ablation runs log them at differing cadences, so their curves differ in sampling density, but all are aligned by training epoch and span the same 100 epochs. Strict pair AUROC and broad retrieval mAP are each averaged over subsets of three validation sets: Nine-Species validation, the PXD010000 development subset, and Kingdoms validation. The end-AA probes use a dedicated set of 4,200 training, 900 validation and 900 test peptides. The kNN probe ($k = 64$, cosine similarity) is scored on its validation rows, and the linear probe is fitted on its training rows and scored on its test rows. The three objectives differ only in how precursor conditioning and mixture peaks are allocated across view scales. The pure mixture objective gives the precursor-conditioned global student views mixed fragment peaks and has no precursor-withheld partial-view objective. The mixture-free precursor-inference control introduces no distractor peaks at all: clean 95% global student views receive the null precursor and predict clean precursor-conditioned teacher targets, while clean 60% partial student views retain the true precursor, with the two losses weighted equally. The dual objective is the method of Section 3.

For de novo transfer (Table 10), each pretrained encoder is paired with a fresh 94.6M-parameter decoder and trained for 80 epochs on MSKB-final training data at a 200-peak cap with precursor conditioning, either with the encoder frozen or end to end. The frozen control is a randomly initialized encoder, and the end-to-end control is trained from scratch. Each run reports its maximum validation peptide precision over the 80 epochs. The distributed validation sampler evaluates 10,016 items by repeating 12 of the 10,004 validation spectra, identically for every run, so the comparison is unaffected.

With the encoder frozen, the dual objective gives the highest peptide precision, followed by the pure mixture objective, the mixture-free control and a randomly initialized encoder (Table 10). End-to-end fine-tuning preserves this ordering, and every objective improves on training from scratch. Because fine-tuning adapts the encoder under full peptide supervision, it compresses the differences between objectives. The representations themselves differ substantially, as the frozen column and the probes of Figure 2 show, and the dual objective leads on every fragment-sensitive measure.

Representation geometry is additionally summarized by unsupervised embedding quality evaluation (Tsitsulin et al., 2023), a family of label-free diagnostics that includes rank-based measures such

Table 10: De novo transfer from the 100-epoch ablation checkpoints on the MSKB-final development-*validation* split. The last row is the control without pretraining: a randomly initialized frozen encoder, and a model trained from scratch.

	Peptide precision	
Pretraining objective	Frozen encoder	End-to-end
Dual objective	**0.675**	**0.782**
Pure mixture	0.638	0.778
Mixture-free control	0.610	0.764
No pretraining	0.582	0.744

as RankMe (Garrido et al., 2023), eigenspectrum-decay measures such as α-ReQ (Agrawal et al., 2022), and NESum (He & Ozay, 2022). We report normalized RankMe and self-cluster.

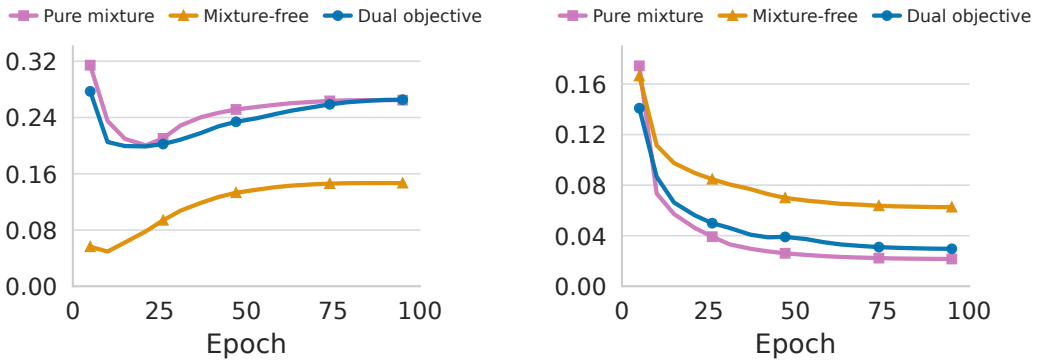


Figure 4: Unsupervised embedding quality evaluation for the objective ablation: normalized RankMe (Garrido et al., 2023) and self-cluster (Tsitsulin et al., 2023). These are coarse checks for pathological geometry and are not monotonic measures of representation quality, so they are not used to rank the objectives. In particular, the pure mixture objective shows no dimensional collapse: its effective rank is not low. Its failure is representational collapse onto precursor information, which these diagnostics cannot detect.

C.2 Mechanistic validation of the dual latent-prediction views

Both probes evaluate the same frozen checkpoint on the NineSpecies V2 validation retrieval artifact described in Appendix A: 6,119 spectra forming 900 repeated-peptide groups, 100 groups from each of the nine species, with at most ten spectra per group. Peptide backbones occurring in the reserved NineSpecies test benchmark were excluded under uppercase amino-acid normalization, so the probe population is disjoint from that benchmark. Spectra use the standard preprocessing of Appendix B.1 with the peak cap set to 200.

Pairing is deterministic. Up to 500 anchors are drawn per species, giving 4,488 directed (A, B) pairs. A candidate B must carry a different peptide identity from A, be a different spectrum, differ by more than 10 ppm in neutral precursor mass, and, when the charges match, differ by more than 10 ppm in precursor m/z. Anchors are stratified by species but B is drawn from the full artifact and need not share the anchor's species. All reported values are unweighted means over the nine anchor-species partitions.

For the precursor-swap probe the mixture is formed once per pair and embedded twice, so only the precursor differs between the two embeddings. The competing spectrum contributes at most as many peaks as the anchor retains (Equation 13), matching the size asymmetry of the training-time construction; peaks within 5 ppm are merged with summed intensity and intensity-weighted m/z, and no intensity renormalization follows. This bound, together with the one-sided objective in which the clean anchor is always the target, is the reason the competing-precursor condition is the harder of the two: the competing source is systematically the weaker contributor to the mixture and is

never itself a prediction target. Both are deliberate properties of a representation-learning objective. Removing the size bound and adding a symmetric term, so that either source must be recoverable under its own precursor, is the natural extension toward a decomposition model, and is not attempted here. For the partial-view probe, two 60% peak subsets of the anchor are sampled independently and embedded under the null precursor state, with no competing peaks mixed in.

Tables 11 and 12 report every readout. The margin is the mean gap between the correct and incorrect source in the units of each readout, so it is comparable within a row but not across rows.

Table 11: Precursor-conditioned source selection by dIon on the NineSpecies validation set. The mixture $A \cup B$ of an anchor A and a competing spectrum B is held fixed and only the precursor changes. A trial is correct when the representation lies closer to the clean representation of the conditioned source than to that of the other source.

	Correct-source accuracy			Separation
Readout	Precursor of A	Precursor of B	Both	Margin
Backbone cosine (selected)	0.997	0.929	0.926	0.767
Backbone Euclidean	0.999	0.847	0.846	18.860
Projection head, Jensen–Shannon	0.998	0.869	0.867	0.004

The margin is expressed in the units of each readout: cosine similarity, Euclidean distance, and Jensen–Shannon divergence.

Table 12: Source recovery by dIon from independent precursor-withheld partial views, NineSpecies validation set. Two independently sampled 60% peak subsets of each anchor are embedded under the null precursor state. A view is correct when it lies closer to the clean representation of its own anchor A than to that of a different-peptide spectrum B.

	Correct-source accuracy				Separation
Readout	View 1	View 2	Both	Mean	Margin
Backbone cosine (selected)	0.987	0.984	0.978	0.986	0.498

The two probes above use constructed mixtures. We also test precursor-conditioned selection on measured co-isolated spectra, in which the instrument isolated two peptides together (Houel et al., 2010; Michalski et al., 2011), and replace the two-point swap with a sweep of the conditioning mass.

A spectrum qualifies when two peptides were separately identified in the same run with the same precursor charge, eluted within 60 s of each other, and the second precursor lies inside the recorded isolation window of the first, targeted precursor. Each peptide must be identified in at least two other scans, which give its reference representation, embedded under its own precursor. We hold the peaks fixed, sweep the conditioning mass in 0.005 Da steps, and record the cosine margin: the cosine similarity to the reference of the targeted peptide A minus that to the co-isolated peptide B.

In the spectrum of Figure 5, conditioning on each true precursor mass selects the corresponding peptide, and the two references are nearly orthogonal (cosine 0.009). Selection also repeats at the ^{13}C spacing: the margin favours A again near A + 1.003 Da and B near B − 1.003 Da, consistent with the one-isotope errors common in precursor assignment. Across 45 co-isolated pairs from one run, the margin changes sign between the two true masses in 6, three of which rest on a single reference scan for one peptide, and in no pair closer than 0.2 Da. In the other pairs, conditioning on B moves the representation toward B without making it closer to B than to A. B is the harder direction here for the same reason as in the constructed probes (Table 11): it is the weaker contributor to the spectrum, and training never makes the competing source a target.

C.3 Complete representation-geometry diagnostics

The main tables emphasize precursor-controlled retrieval AP and static-pair AUROC. For completeness, the tables below additionally report unfiltered global retrieval and static-pair AP from the same

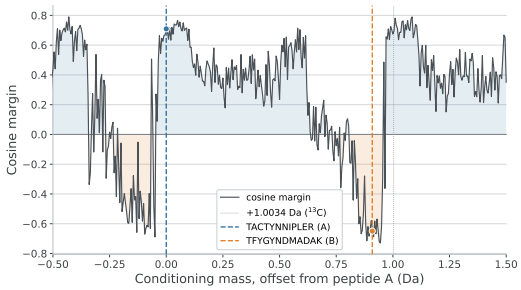


Figure 5: Conditioning-mass sweep on one measured spectrum containing the co-isolated peptides TACTYNNIPLER (A, targeted) and TFYGYNDMADAK (B). The peaks are fixed and only the conditioning mass varies. A positive margin means the representation is closer to A's reference than to B's. Dashed lines mark the true masses of A and B, and the dotted line the first ^{13}C isotope of A.

evaluation runs. Global retrieval admits precursor-mass and charge shortcuts, so we treat it as a diagnostic rather than a headline comparison.

Table 13: Native representation geometry across test benchmarks over all precursor charges. Unlike the charge-2–4 subsets of Table 2, these populations retain slightly more spectra and exclude GLEAMS, which embeds only charges 2–4. All metrics compare only spectra with the same precursor charge and precursor mass within 10 ppm.

	Bacterial		NineSpecies		Kingdoms	
Method	Retrieval mAP	Pairs AUROC	Retrieval mAP	Pairs AUROC	Retrieval mAP	Pairs AUROC
Binned spectral angle	0.913	0.941	0.972	0.954	0.908	0.871
Random Transformer	0.714	0.791	0.823	0.704	0.710	0.637
Peptide-supervised						
Casanovo 4.0	0.880	0.896	0.933	0.858	0.838	0.746
Self-supervised						
InstaNovo-FM	0.847	0.862	0.877	0.762	0.806	0.708
dIon	0.893	0.927	0.922	0.864	0.824	0.748
Peptide-supervised contrastive						
Scratch init. (1%)	0.907	0.932	0.970	0.946	0.909	0.873
dIon init. (1%)	0.920	0.952	0.971	0.952	0.914	0.887
Scratch init. (10%)	0.926	0.958	**0.977**	0.956	0.933	0.901
dIon init. (10%)	**0.930**	**0.966**	**0.977**	**0.965**	**0.942**	**0.926**

Table 14: Complete native-geometry diagnostics on the bacterial test set.

	Retrieval		Static pairs	
Method	Broad mAP	AP	AP	AUROC
Binned spectral angle	0.527	0.913	0.934	0.941
Random Transformer	0.126	0.714	0.804	0.791
Casanovo 4.0	0.416	0.880	0.900	0.896
InstaNovo-FM	0.377	0.847	0.876	0.862
dIon	0.492	0.893	0.926	0.927
Supervised metric learning				
Scratch 10%	0.544	0.926	0.947	0.958
dIon 10% f.t.	0.577	0.930	0.960	0.966

Table 15: Complete native-geometry diagnostics on the NineSpecies V2 test set.

	Retrieval		Static pairs	
Method	Broad mAP	AP	AP	AUROC
Binned spectral angle	0.784	0.972	0.939	0.954
Random Transformer	0.157	0.823	0.726	0.704
Casanovo 4.0	0.524	0.933	0.862	0.858
InstaNovo-FM	0.417	0.877	0.784	0.762
dIon	0.611	0.922	0.868	0.864
Supervised metric learning				
Scratch 10%	0.796	0.977	0.937	0.956
dIon 10% f.t.	0.807	0.977	0.952	0.965

Table 16: Complete native-geometry diagnostics on the Kingdoms test set.

	Retrieval		Static pairs	
Method	Broad mAP	AP	AP	AUROC
Binned spectral angle	0.499	0.908	0.879	0.871
Random Transformer	0.083	0.710	0.661	0.637
Casanovo 4.0	0.324	0.838	0.780	0.746
InstaNovo-FM	0.323	0.806	0.753	0.708
dIon	0.442	0.824	0.767	0.748
Supervised metric learning				
Scratch 10%	0.600	0.933	0.913	0.901
dIon 10% f.t.	0.655	0.942	0.934	0.926

Table 17: Complete charge-2–4 matched diagnostics on the bacterial test set.

	Retrieval		Static pairs	
Method	Broad mAP	AP	AP	AUROC
Binned spectral angle	0.537	0.915	0.938	0.943
Casanovo 4.0	0.424	0.881	0.903	0.898
InstaNovo-FM	0.384	0.847	0.880	0.864
GLEAMS	0.475	0.916	0.946	0.948
dIon	0.506	0.895	0.929	0.928
Supervised metric learning				
Scratch 10%	0.559	0.928	0.950	0.960
dIon 10% f.t.	0.589	0.933	0.962	0.968

Table 18: Complete charge-2–4 matched diagnostics on the NineSpecies V2 test set.

	Retrieval		Static pairs	
Method	Broad mAP	AP	AP	AUROC
Binned spectral angle	0.786	0.975	0.939	0.954
Casanovo 4.0	0.526	0.932	0.863	0.859
InstaNovo-FM	0.419	0.874	0.785	0.763
GLEAMS	0.746	0.976	0.951	0.962
dIon	0.612	0.926	0.869	0.865
Supervised metric learning				
Scratch 10%	0.797	0.978	0.938	0.957
dIon 10% f.t.	0.808	0.978	0.953	0.965

Table 19: Complete charge-2–4 matched diagnostics on the Kingdoms test set.

	Retrieval		Static pairs	
Method	Broad mAP	AP	AP	AUROC
Binned spectral angle	0.503	0.905	0.879	0.871
Casanovo 4.0	0.327	0.838	0.780	0.746
InstaNovo-FM	0.326	0.805	0.753	0.708
GLEAMS	0.440	0.903	0.895	0.876
dIon	0.445	0.821	0.767	0.748
Supervised metric learning				
Scratch 10%	0.605	0.930	0.914	0.902
dIon 10% f.t.	0.659	0.940	0.934	0.926

D Downstream transfer setup

D.1 Auxiliary spectrum-level task definitions

Spectral quality assessment, chimericity, and oxidized-methionine prediction share a single binary prediction head applied to the pooled global representation $z \in \mathbb{R}^{1024}$,

$$\begin{aligned} z \;&\to \mathrm{Linear}(1024, 128) \to \mathrm{GELU} \to \mathrm{LayerNorm}(128) \\ &\to \mathrm{Linear}(128, 64) \to \mathrm{GELU} \to \mathrm{LayerNorm}(64) \to \mathrm{Linear}(64, 1), \end{aligned} \tag{19}$$

whose scalar output is a binary logit. Dropout is disabled throughout this head.

Table 20: Auxiliary binary spectrum-level tasks on held-out test sets (AUROC). Frozen representations and end-to-end adaptation are reported as separate regimes.

	Task AUROC		
Representation / initialization	SQA	Chimericity	Oxidized Met
Frozen transfer			
Binned + precursor metadata	0.737	0.708	0.654
Random Transformer	0.753	0.760	0.642
Casanovo 4.0	0.807	0.807	0.795
InstaNovo-FM	0.810	0.827	0.835
dIon (conditioned)	0.805	0.704	0.730
Fine-tuned transfer			
Scratch, end-to-end	0.783	0.762	0.749
dIon, end-to-end	0.860	0.799	0.855

D.2 Supervised peptide metric learning

A freshly initialized projection head is applied to the pooled global representation, upstream of the discarded pretraining prototype head,

$$z \;\to\; \mathrm{Linear}(1024, 1024) \to \mathrm{GELU} \to \mathrm{Linear}(1024, 128) \to \ell_2\text{-normalization}, \tag{20}$$

producing unit-norm 128-dimensional embeddings.

Writing v_i for the normalized embedding of example i, y_i for its peptide label, and $P(i) = \{p \neq i : y_p = y_i\}$ for its set of in-batch positives, the supervised contrastive objective (Khosla et al., 2020) is

$$\mathcal{L}_i = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(v_i^\top v_p / \tau\right)}{\sum_{a \neq i} \exp\left(v_i^\top v_a / \tau\right)}, \qquad \tau = 0.07, \tag{21}$$

averaged over anchors with at least one positive. Different-peptide examples enter as negatives through the denominator, so no separate negative-loss term is required.

E De novo sequencing and dense transfer

E.1 De novo encoder–decoder architecture

The peptide decoder is a freshly initialized Casanovo-style autoregressive Transformer that performs causal self-attention over the peptide prefix and cross-attention over the encoder peak-token memory $\{h_i\}$, whose dimension matches the encoder model width. Target sequences use a 45-token vocabulary: 43 residue and N-terminal-modification tokens, one padding token, and an end-of-sequence token. A precursor token is prepended to the decoder target stream, formed by summing a sinusoidal encoding of the neutral precursor mass with a learned charge embedding over charges 1–10, and sinusoidal positional encoding is applied over the precursor token together with the peptide target sequence. Table 21 lists the configuration.

Table 21: De novo peptide decoder configuration.

Component	Setting
Decoder blocks	9
Model dimension	1024
Attention heads	8
Feed-forward width	1024
Dropout	0.25
Encoder memory dimension	1024
Self-attention	Causal, over peptide tokens
Cross-attention	Encoder peak-token memory $\{h_i\}$
Output projection	Linear(1024, 45) to token logits
Target vocabulary	45 tokens (43 residue/N-terminal modification, padding, [EOS])
Decoder precursor token	Sinusoidal precursor-mass encoding + learned charge embedding (1–10)
Positional encoding	Sinusoidal, over precursor token and peptide sequence

E.2 Autoregressive training and fine-tuning

Supervised de novo training minimizes teacher-forced categorical cross-entropy over peptide tokens without label smoothing, using the Adam optimizer with a cosine warmup and decay schedule. Table 22 lists the settings, which are held fixed across scratch and dIon-initialized conditions.

Table 22: De novo supervised training settings, held fixed across initialization conditions.

Setting	Value
Objective	Teacher-forced token cross-entropy
Label smoothing	None
Optimizer	Adam
Learning rate	1×10^{-4}
Weight decay	None
Gradient clipping	2.0
Schedule	Cosine warmup and decay

E.3 Unlabeled yeast evaluation

The unlabeled benchmark uses one public yeast run from PRIDE accession PXD026806, file `MDR27022021_S01_2_DDA01`, converted to peak lists with ThermoRawFileParser 1.4.5, giving 39,460 spectra. No ground-truth peptide enters the pipeline at any stage. All three models receive the identical 39,460-scan roster under the same identifiers and decode greedily with a single beam.

A prediction is scored by whether it occurs as a contiguous substring of the yeast proteome (UniProt UP000002311, 6,067 sequences), concatenated with separators so that no match may span two proteins. Modifications are stripped before matching, and isoleucine and leucine are collapsed to a single symbol on both sides, since the two are isobaric and therefore indistinguishable in principle. Because short strings occur by chance in a reference of this size, the same procedure is repeated

against two decoy references built from the same sequences, one with each protein reversed and one with each protein shuffled while preserving its amino-acid composition. These bound the match rate attainable without real signal.

dIon and InstaNovo return a prediction for every spectrum. Casanovo declines 3,934, of which 3,188 carry too few peaks and 746 exceed its supported charge range. These are retained as explicit declines, so the match fraction is reported both over emitted predictions and over all submitted spectra. Table 23 also gives coverage, the tryptic-restricted match fraction, the composition-preserving decoy, and the fraction of calls carrying a modification.

Table 23: Complete unlabeled yeast results. Coverage is the fraction of submitted spectra receiving a prediction, tryptic restricts matches to tryptic peptides, the shuffled decoy preserves each protein's amino-acid composition, and modified calls are predictions carrying at least one modification.

		Matched			Decoy		
Model	Coverage	Emitted	All spectra	Tryptic	Reversed	Shuffled	Modified calls
This work							
dIon init., DNL-v1 training	**1.000**	**0.332**	**0.332**	**0.319**	0.0025	0.0021	0.239
External reference							
Casanovo 5.2.1	0.900	0.324	0.292	0.311	0.0020	0.0024	**0.326**
InstaNovo 1.2.2	**1.000**	0.320	0.320	0.305	0.0020	0.0017	0.291

E.4 KINGDOMS EVALUATION DETAILS

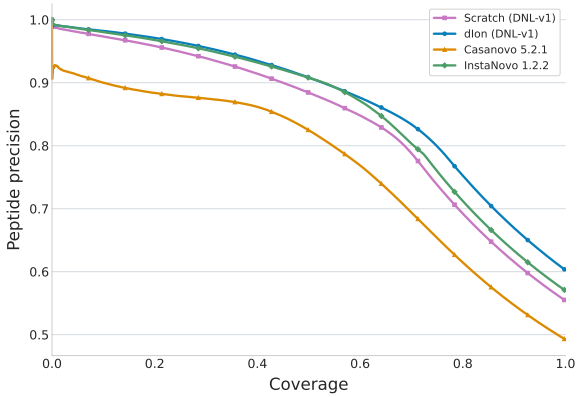


Figure 6: Peptide precision against coverage on the Kingdoms de novo test set. Each model ranks its predictions by its own confidence score, and coverage is the fraction of all test spectra accepted, so the right endpoint is the full-coverage precision of Table 1. dIon and scratch are fine-tuned end to end on DNL-v1 with 200 peaks.

Table 24 reports the complete Kingdoms results, including amino-acid metrics and, for both initializations, models fine-tuned with a 1,000-peak cap from the same 200-peak pretrained encoder. With the larger cap, dIon initialization gains in both training regimes. Scratch training gains when trained on MSKB-final and not when trained on DNL-v1.

The canonical Kingdoms test set contains 4,926,232 unique spectra. For the two DNL-v1-trained 200-peak models, peptide precision is rescored from their complete prediction tables over exactly these spectra, overall and per species, with the same matcher used for Casanovo and InstaNovo. The other six in-house configurations, and the amino-acid metrics of all eight, come from a distributed fast-evaluation mode. That mode padded the final batch across sixteen workers by repeating the first eight records, all belonging to *Akkermansia muciniphila*: logged per-species counts sum to 4,926,240, and that species shows 27,286 logged evaluations against 27,278 unique spectra. The external Casanovo and InstaNovo results come from separate sharded inference pipelines and do not share this behaviour. Nor does anything on the MSKB-final test set, where every model wrote

individual predictions to disk instead. The logged scores are reported unchanged. The effect is bounded and negligible: for a binary per-spectrum metric such as peptide precision the difference between the logged and unique denominators cannot exceed $8/4{,}926{,}240 \approx 1.6 \times 10^{-4}$ percentage points overall, or $8/27{,}286 \approx 0.03$ percentage points for *Akkermansia muciniphila* alone. No other species is affected.

Casanovo 5.2.1 supports precursor charges 1–4 and emitted predictions for 4,879,442 spectra. The 13,504 charge-> 4 spectra and 33,286 supported spectra without an emitted PSM are retained as incorrect, zero-confidence predictions, so its full-coverage precision and precision–coverage curve use the complete 4,926,232-spectrum denominator. InstaNovo 1.2.2 emitted a prediction for every spectrum.

Table 24: Complete Kingdoms de novo results, including amino-acid metrics and the models fine-tuned with a 1,000-peak cap for both initializations. Within each training regime, scratch and dIon initialization use an identical end-to-end training protocol.

	Amino acids		
Training / initialization	Precision	Recall	Precision @ 100% coverage
MSKB-final training			
Scratch, end-to-end	0.610	0.612	0.449
Scratch, end-to-end (1,000 peaks)	0.615	0.617	0.454
dIon init., end-to-end	0.702	0.702	0.533
dIon init., end-to-end (1,000 peaks)	0.709	0.708	0.538
DNL-v1 training			
Scratch, end-to-end	0.703	0.704	0.554
Scratch, end-to-end (1,000 peaks)	0.701	0.704	0.553
dIon init., end-to-end	0.755	0.755	0.602
dIon init., end-to-end (1,000 peaks)	**0.757**	**0.757**	**0.607**
External reference			
Casanovo 5.2.1	—	—	0.492
InstaNovo 1.2.2	0.696	0.698	0.570

Table 25 breaks the full-coverage peptide precision down by species for the two DNL-v1-trained 200-peak models, Casanovo 5.2.1 and InstaNovo 1.2.2. dIon initialization has the highest precision of the four models on every species, by a margin that rounds to zero only on *Rattus norvegicus*.

E.5 MSKB-final evaluation details

Table 26 reports the complete MSKB-final results, including amino-acid metrics, the models fine-tuned with a 1,000-peak cap, and InstaNovo 1.2.2, which the main text omits on this benchmark.

E.6 Oracle-precursor decoding of DIA spectra

In data-independent acquisition (DIA), wide isolation windows co-fragment several peptides, so each MS2 spectrum is a mixture. As a pilot, we test whether precursor conditioning can select one component of such a mixture. Each query pairs a real wide-window DIA MS2 scan, with isolation windows at least 26 m/z wide and its 100 most intense peaks retained, with the precursor m/z and charge of one peptide that DIA-NN identified in it (Demichev et al., 2020), and the model decodes that peptide autoregressively. A scan recurs in several queries with different precursors, giving 17,781 queries over 9,973 scans. The DIA-NN precursor acts as an oracle selection query, so the pilot tests decoding of a supplied component rather than MS1 precursor detection or full deconvolution.

The encoder is the Gram-refined dIon encoder (Appendix E.9), held frozen. The de novo decoder is trained on synthetic two-spectrum mixtures of DNL-v1, built with the pretraining mixture construction: each mixture is decoded twice, conditioned on either precursor, with that precursor's peptide as the target. Neither pretraining nor decoder training uses DIA data. Inference uses precursor conditioning and greedy decoding, and every query receives a prediction. Table 27 reports the results. No baseline has been evaluated on these queries. The absolute precision is low, but it is exact recovery

Table 25: Per-species peptide precision at full coverage on the Kingdoms de novo test set (deterministic run-disjoint split, at most 100k spectra per species).

		DNL-v1, 200 peaks		External reference	
Species	Spectra	dIon init.	Scratch	Casanovo 5.2.1	InstaNovo 1.2.2
Akkermansia muciniphila	27,278	**0.584**	0.535	0.487	0.546
Arabidopsis thaliana (callus)	100,000	**0.617**	0.573	0.528	0.586
Arabidopsis thaliana (root)	100,000	**0.620**	0.574	0.510	0.581
Arabidopsis thaliana (sprout)	100,000	**0.652**	0.605	0.537	0.620
Bacillus subtilis	17,126	**0.559**	0.527	0.506	0.546
Bacteroides fragilis	76,935	**0.631**	0.583	0.505	0.574
Bacteroides thetaiotaomicron	75,950	**0.607**	0.561	0.493	0.554
Bacteroides uniformis	77,854	**0.632**	0.585	0.512	0.576
Bacteroides vulgatus	69,903	**0.612**	0.562	0.491	0.554
Bifidobacterium adolescentis	50,338	**0.563**	0.519	0.471	0.537
Bifidobacterium longum subsp. *longum*	40,502	**0.525**	0.487	0.451	0.503
Blautia obeum	46,515	**0.606**	0.554	0.502	0.572
Bos taurus	100,000	**0.672**	0.617	0.540	0.649
Caenorhabditis elegans	100,000	**0.633**	0.584	0.519	0.600
Canis lupus	100,000	**0.631**	0.576	0.504	0.607
Chlamydomonas reinhardtii	45,545	**0.537**	0.504	0.476	0.520
Chromera velia	100,000	**0.535**	0.492	0.450	0.519
Clostridium bolteae	56,691	**0.587**	0.534	0.485	0.553
Clostridium perfringens	59,588	**0.642**	0.594	0.538	0.600
Clostridium saccharolyticum	61,717	**0.608**	0.556	0.502	0.570
Collinsella aerofaciens	45,181	**0.570**	0.525	0.475	0.538
Coprococcus comes	55,776	**0.612**	0.560	0.500	0.569
Cricetulus griseus	100,000	**0.551**	0.500	0.451	0.545
Danio rerio	100,000	**0.603**	0.554	0.485	0.575
Dictyostelium discoideum	100,000	**0.622**	0.566	0.476	0.569
Didelphis didelphinae	100,000	**0.645**	0.589	0.508	0.611
Dorea formicigenerans	40,528	**0.555**	0.514	0.485	0.530
Drosophila melanogaster	100,000	**0.640**	0.586	0.501	0.596
Eggerthella lenta	25,395	**0.519**	0.462	0.408	0.482
Emiliania huxleyi	100,000	**0.560**	0.515	0.449	0.530
Escherichia coli	61,880	**0.619**	0.567	0.495	0.573
Eubacterium rectale	54,888	**0.564**	0.516	0.467	0.521
Fusarium oxysporum	100,000	**0.494**	0.457	0.438	0.476
Fusobacterium nucleatum subsp. *nucleatum*	60,505	**0.641**	0.594	0.538	0.599
Gallus gallus	100,000	**0.659**	0.607	0.528	0.622
Glycine max	100,000	**0.608**	0.561	0.494	0.573
Gossypium hirsutum	100,000	**0.607**	0.560	0.504	0.575
Guillardia theta	100,000	**0.579**	0.533	0.472	0.550
Haloarcula marismortui	1,967	**0.652**	0.616	0.581	0.589
Halobacterium sp.	10,815	**0.556**	0.522	0.489	0.518
Homo sapiens	100,000	**0.679**	0.624	0.543	0.662
Lactobacillus paracasei	51,446	**0.586**	0.541	0.489	0.554
Mus musculus	100,000	**0.611**	0.555	0.479	0.596
Mycoplasma intracellular	5,325	**0.761**	0.698	0.636	0.743
Neurospora crassa	100,000	**0.553**	0.511	0.461	0.522
Odoribacter splanchnicus	71,173	**0.598**	0.552	0.500	0.551
Oryctolagus cuniculus	100,000	**0.648**	0.582	0.497	0.611
Oryzias melastigma	100,000	**0.593**	0.539	0.471	0.557
Parabacteroides distasonis	70,978	**0.637**	0.585	0.523	0.588
Parabacteroides merdae	68,437	**0.627**	0.577	0.520	0.577
Phaeodactylum tricornutum	100,000	**0.555**	0.511	0.457	0.532
Porphyra umbilicalis	100,000	**0.617**	0.574	0.497	0.583
Prevotella copri	10,287	**0.646**	0.598	0.537	0.599
Rattus norvegicus	100,000	**0.580**	0.531	0.482	**0.580**
Roseburia intestinalis	34,833	**0.522**	0.476	0.445	0.497
Ruminococcus bromii	50,037	**0.595**	0.549	0.505	0.562
Ruminococcus gnavus	49,639	**0.623**	0.573	0.505	0.578
Ruminococcus torques	14,859	**0.649**	0.600	0.530	0.604
Saccharomyces cerevisiae	100,000	**0.650**	0.602	0.517	0.602
Stentor coeruleus	100,000	**0.615**	0.562	0.486	0.572
Streptococcus parasanguinis	50,003	**0.627**	0.577	0.518	0.583
Sulfolobus solfataricus	6,916	**0.567**	0.514	0.473	0.543
Sus scrofa	100,000	**0.577**	0.525	0.462	0.572
Symbiodinium microadriaticum	100,000	**0.551**	0.510	0.477	0.546
Synechocystis nigrescens	2,324	**0.652**	0.615	0.577	0.633
Tardigrade	25,726	**0.618**	0.564	0.493	0.578
Thalassiosira pseudonana	100,000	**0.562**	0.515	0.454	0.531
Triticum aestivum	100,000	**0.616**	0.569	0.488	0.577
Veillonella parvula	51,372	**0.554**	0.509	0.454	0.511
Vitis vinifera	100,000	**0.581**	0.543	0.510	0.559

Table 26: Complete MSKB-final de novo results, including amino-acid metrics and the dIon-initialized models fine-tuned with a 1,000-peak cap. Within each training regime, scratch and dIon initialization use an identical end-to-end training protocol.

	Amino acids		
Training / initialization	Precision	Recall	Precision @ 100% coverage
MSKB-final training			
Scratch, end-to-end	0.895	0.895	0.747
dIon init., end-to-end	0.927	0.926	0.802
dIon init., end-to-end (1,000 peaks)	0.935	0.934	0.822
DNL-v1 training			
Scratch, end-to-end	0.933	0.934	0.819
dIon init., end-to-end	0.945	0.945	0.842
dIon init., end-to-end (1,000 peaks)	**0.949**	**0.949**	**0.856**
External reference			
Casanovo 5.2.1	0.911	0.912	0.783
InstaNovo 1.2.2*	0.961	0.961	0.916

InstaNovo 1.2.2 (grey) is documented as training on MassIVE-KB v1, the source corpus of this test set. Its exact training subset and split assignments are unpublished, so overlap with the held-out partition cannot be excluded.*

of target peptides from real DIA mixtures by a model trained only on synthetic two-source mixtures of DDA spectra. A decisive test would replace the oracle precursor with observed MS1 precursor features and use narrower isolation windows.

Table 27: Oracle-precursor decoding of real wide-window DIA MS2 scans: 17,781 queries, each pairing one of 9,973 scans with the precursor of a DIA-NN-identified peptide in it. Metrics use the matcher of Appendix E.7.

	Amino acids			
Model	Precision	Recall	Peptide precision	Precision–coverage AUC
dIon, frozen encoder	0.225	0.227	0.106	0.328

E.7 Decoding and sequence evaluation

All reported models use one frozen decoding protocol, summarized in Table 28. Peptides are decoded in reverse order with single-beam (greedy) decoding under a precursor-mass constraint.

Table 28: De novo decoding protocol, identical across all reported models.

Setting	Value
Decoding order	Reverse peptide order
Search	Single beam (greedy)
Maximum peptide length	100
Minimum peptide length	6
Precursor mass tolerance	50 ppm
Isotope-error range	0–1

Sequence matching. Every model, including Casanovo and InstaNovo, is scored with one implementation: a vendored copy of the Casanovo evaluator (Yilmaz et al., 2022), which applies the matching criterion introduced by DeepNovo (Tran et al., 2017). Predicted and reference peptides are first mapped to one token vocabulary and one table of residue masses, with isoleucine and leucine treated as identical, so that each model's modification notation resolves to the same tokens. Residues are aligned from the N terminus and, if the peptides do not fully match, again from the C terminus.

A predicted residue matches when its mass differs from the aligned reference residue by less than 0.1 Da and the cumulative mass up to and including it differs by less than 0.5 Da.

Amino-acid precision and recall divide the number of matched residues by the number of predicted and reference residues, respectively. A peptide is correct only when every aligned position matches, so a missing or extra residue makes it incorrect. Peptide precision is the number of correct peptides divided by the number of test spectra. Spectra without a prediction, including those an external model does not support, count as incorrect. On MSKB-final, Casanovo 5.2.1 with its stock preprocessing receives 196,979 charge-1–4 spectra. Of these, 20 fall below its 20-peak minimum after filtering and one yields no PSM, and all 21 count as incorrect.

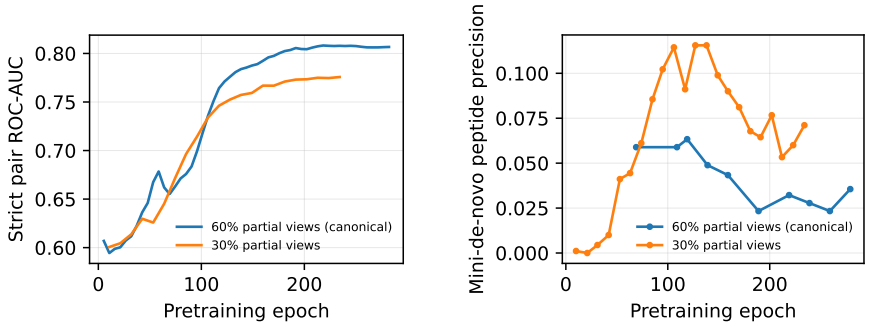
Peptide confidence. Following Casanovo (Yilmaz et al., 2024), the confidence of a predicted peptide is the mean softmax probability of its emitted tokens, including the stop token. When decoding ends without a stop token, a probability of zero takes its place. If the peptide does not match the precursor mass within 50 ppm, allowing isotope errors of 0–1, the confidence is reduced by 1, so every precursor-compatible prediction ranks above every incompatible one. Casanovo and InstaNovo keep their native confidence scores. Scores only rank the predictions of one model and are not comparable across models.

A precision–coverage curve sorts one model's predictions by decreasing confidence. At rank k, coverage is k divided by the number of test spectra and precision is the fraction of the top k predictions that are correct. Spectra without a prediction rank last as errors, so the curve ends at the full-coverage peptide precision.

E.8 Online mini-de-novo probe

The online mini-de-novo probe tracks how readily the dense peak-token representations support sequence prediction during pretraining. At each evaluation the EMA-teacher encoder is frozen and its per-peak outputs are cached, and a fresh two-layer Transformer decoder (two attention heads, feed-forward width 128) is trained on them for four epochs with AdamW (learning rate 3×10^{-4}, batch size 32). The probe data are drawn from the MSKB-final training set, restricted to unmodified peptides of 6 to 10 residues over 16 amino acids, and split by run and peptide into 7,000 training, 900 validation and 900 test spectra. Decoding and scoring follow Appendix E.7, and we report validation peptide precision. The probe is a development diagnostic, separate from the de novo evaluation of Section 4.2.

Partial views retaining 30% instead of 60% of peaks give a much stronger mini-de-novo probe, which peaks mid-training and then declines, but lower strict pair ROC-AUC from mid-training onward (Figure 7). The two runs differ only in this fraction, and the canonical recipe keeps 60%. Its weaker frozen dense probe does not carry over to fine-tuning: anchoring the canonical encoder to the 30% run's token features improves frozen de novo transfer but not end-to-end fine-tuning, where the canonical encoder is already as good (Appendix E.9).



(a) Strict pair ROC-AUC

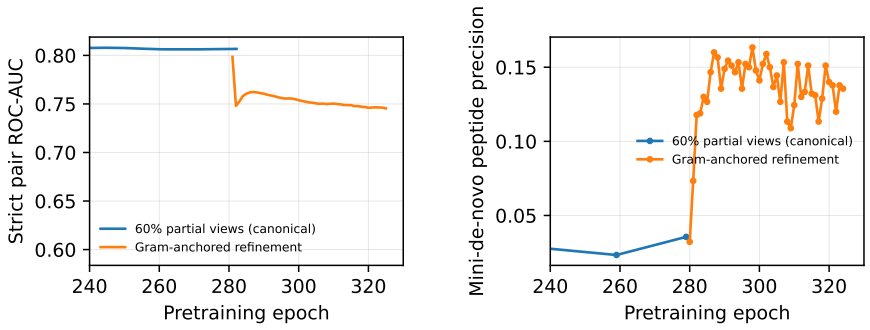
(b) Mini-de-novo peptide precision

Figure 7: Partial views retaining 30% or 60% of peaks, *validation* probes. The canonical run's mini-de-novo points are an offline reconstruction from its saved checkpoints.

E.9 GRAM-ANCHORED REFINEMENT

The mini-de-novo probe peaks mid-training and then declines, more steeply with 30% partial views (Appendix E.8). To recover the dense information, we continued the canonical run from epoch 279 with Gram anchoring (Siméoni et al., 2025). Gram anchoring constrains the similarity structure of the dense features rather than the features themselves. For a spectrum with N peaks, let $F \in \mathbb{R}^{N\times d}$ hold the ℓ_2-normalized peak-token outputs of the student and F_a those of a frozen anchor encoder. The Gram matrices $FF^\top$ and $F_aF_a^\top$ hold the cosine similarity of every pair of peak tokens, and the loss is their mean squared difference, $\|FF^\top - F_aF_a^\top\|_F^2/N^2$. Because a Gram matrix is unchanged by rotations of the feature space, the student remains free to move its features under the DINO objective as long as it keeps the anchor's pattern of which peaks resemble which. The anchor is the teacher of the 30% run at epoch 109, near the peak of its probe. The loss is added to the DINO objective at weight 2.0, ramped in over 1,000 steps, with a fixed learning rate of 4.5×10^{-5} and a fixed EMA momentum of 0.9986. Refinement ran for 45 epochs, and the epoch-298 checkpoint, where the probe peaked, is used downstream.

Refinement raises the mini-de-novo probe well above the unrefined run but lowers strict pair ROC-AUC (Figure 8). In de novo transfer it substantially improves the frozen encoder, but end-to-end fine-tuning of the canonical encoder is already as good (Table 29), so the main results use the unrefined checkpoint. The refined encoder was fine-tuned end to end only on DNL-v1, which leaves the MSKB-final end-to-end entry of Table 29 empty. The frozen DIA pilot of Appendix E.6 uses the Gram-refined epoch-298 encoder.



(a) Strict pair ROC-AUC

(b) Mini-de-novo peptide precision

Figure 8: Gram-anchored refinement continued from epoch 279 of the canonical run, *validation* probes. The canonical run's mini-de-novo points are an offline reconstruction from its saved checkpoints.

Table 29: De novo peptide precision on the MSKB-final test set with and without Gram-anchored refinement of the pretrained encoder.

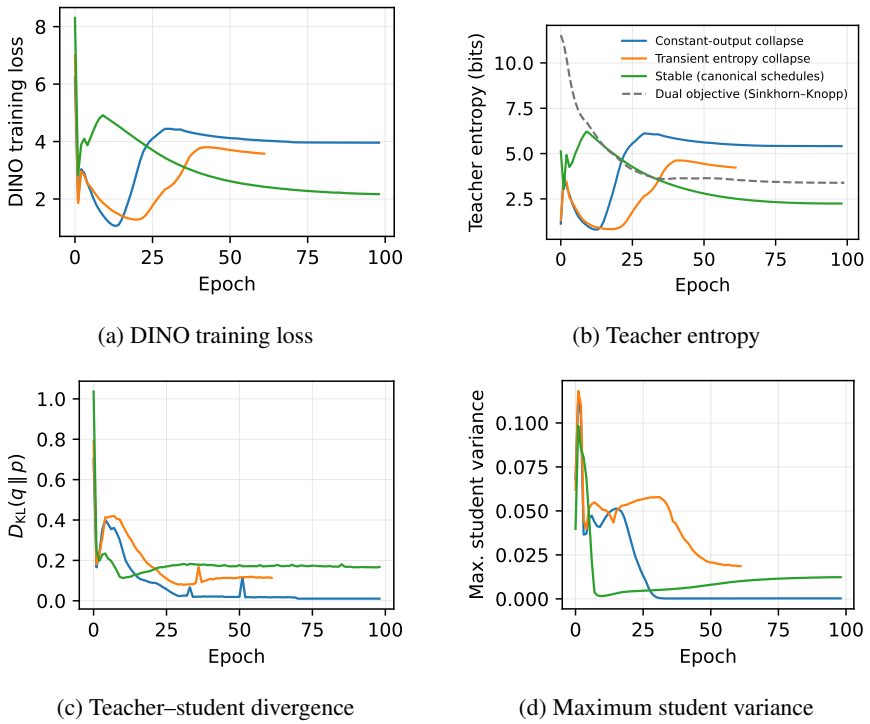
	Peptide precision	
Training corpus / encoder	Frozen encoder	End-to-end
DNL-v1 training		
dIon	0.555	0.842
dIon, Gram-refined	0.712	0.842
MSKB-final training		
dIon	0.487	0.802
dIon, Gram-refined	0.658	—

F Additional training diagnostics

F.1 Collapse monitoring

Pretraining logs three diagnostics from the collapse analysis of DINO (Caron et al., 2021): the mean entropy $H(q)$ of the teacher assignments in bits, the Kullback–Leibler divergence $D_{\mathrm{KL}}(q \,\|\, p)$ between teacher and student assignments, and the largest across-batch variance of any student output probability. A collapsed model gives the same output for every input, which drives the divergence and the variance to zero. The loss alone does not reveal this.

We chose the teacher temperature and momentum schedules on three early development runs (Figure 9). These runs used the teacher centering of the original DINO (Caron et al., 2021) instead of the Sinkhorn–Knopp balancing of DINOv2 (Oquab et al., 2024), and a naive early objective with random peak subsets as views instead of the mixture and partial-view tasks. The schedules they selected nonetheless carried over unchanged to the dual objective. Their two global views kept 95% of the peaks and their five local views 20%. Every run warms the teacher temperature up from 0.03. With a final temperature of 0.09, a 30-epoch warmup and a teacher momentum rising from 0.996 to 0.9995, student and teacher converge to one output distribution. The divergence and the variance fall to near zero while $H(q)$ stays moderate, so the entropy alone does not detect this collapse. With a final temperature of 0.07, the same slow warmup and a momentum rising from 0.997, the teacher instead over-sharpens, and $H(q)$ dips toward zero for several epochs before partially recovering. A 10-epoch warmup to 0.07 with a momentum rising from 0.998 to 1.0 keeps all three diagnostics away from their degenerate values. These are the schedules of the canonical recipe (Table 9), and the switch to Sinkhorn–Knopp balancing, with the settings of DINOv2, required no further tuning. With these schedules and Sinkhorn–Knopp balancing, the teacher entropy of the 100-epoch dual-objective ablation run (Figure 2) falls from near-uniform at initialization to an intermediate value and stays well away from both collapsed extremes (Figure 9, panel b).



(a) DINO training loss

(b) Teacher entropy

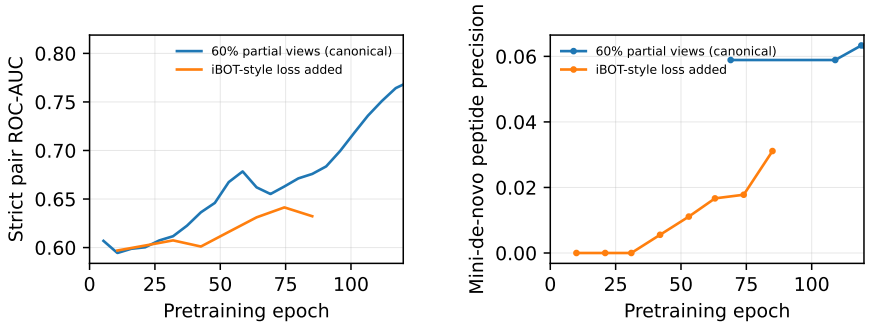
(c) Teacher–student divergence

(d) Maximum student variance

Figure 9: Collapse diagnostics of the three development runs and, in panel (b), the 100-epoch dual-objective ablation run, as per-epoch means. For the dual-objective run the entropy is that of the sharpened teacher distribution before Sinkhorn–Knopp balancing. The transient-collapse run was stopped early.

F.2 MASKED-TOKEN OBJECTIVE SCREEN

We screened adding an iBOT-style masked-token loss (Zhou et al., 2022) to the dual objective. The screen adds two precursor-conditioned global student views in which 10–50% of peak slots are replaced by a learned mask token, and aligns the masked tokens with the teacher's clean token outputs at weight 0.5, with all other settings as in the canonical recipe. The run stopped at epoch 85 of a nominal 300. Compared with the canonical run at matched epochs, it reached lower strict pair ROC-AUC and lower mini-de-novo precision (Appendix E.8, Figure 10), so the loss was not carried forward.



(a) Strict pair ROC-AUC

(b) Mini-de-novo peptide precision

Figure 10: iBOT-style masked-token screen against the canonical run, *validation* probes. The canonical run's mini-de-novo points are an offline reconstruction from its saved checkpoints.