# IONA: SCALING LAWS FOR MASS SPECTROMETRY FOUNDATION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Mass spectrometry is a core analytical technique for identifying chemical molecules, with applications ranging from cancer biomarker and metabolite detection to heavy isotope analysis, pathogen identification, and explosive screening at airports. Building machine learning models that genuinely understand mass spectra is therefore an important problem. Recent transformer models trained on tandem mass spectra span roughly 1M to 400M parameters, yet a basic question remains unanswered: *do larger models actually understand spectra better, and how large is large enough?* No published spectrum model reports a systematic size sweep or a fitted scaling law. In this work, we conduct the **first such study for self-supervised mass spectrometry foundation models**. We present **Iona**, a family of encoders pretrained at 25M, 50M, 100M, 200M, and 400M parameters on MSConsensus-100M, a corpus of 100 million consensus MS/MS spectra, using a masked intensity-reconstruction objective. Evaluation loss follows a power law in pretraining compute, with a compute-efficient frontier on which each larger model overtakes the previous one, and extrapolating the frontier places the compute-optimal model size well beyond the largest spectrum transformer published to date; we therefore aim to push Iona past 1 billion parameters next. Capability grows with scale downstream as well, from denoising and retrieval to chimeric-spectrum decomposition, and with simple fine-tuning Iona reaches or surpasses strong specialised models in real identification pipelines: it beats a supervised proteomics encoder on real chimeric spectra, adds 20,763 peptide–spectrum matches at 1% FDR to database search, and recovers 37–60% more peptides from the weakest single-cell acquisitions. Random-initialisation controls show that this performance comes from pretraining, not fine-tuning. We further test a previously unexamined hypothesis: *can a model succeed without encoding any absolute m/z?* In our design, all mass information enters through a learned Fourier-feature bias on pairwise mass differences, making the encoder translation-invariant in $m/z$. Contrary to the assumption that absolute mass positioning is necessary, this relative-only design solves every task above, and its learned bias recovers amino-acid residue, neutral-loss and isotope masses without supervision: the model learns chemical reality from data. Because the recipe uses no peptide labels, annotations or any other proteomics-specific ingredient, we expect the same scaling behaviour to hold for metabolomics and other mass spectrometry applications.

## 1 INTRODUCTION

Tandem mass spectrometry (MS/MS) is the workhorse technology for identifying proteins and small molecules. In tandem mass spectrometry, molecules are isolated and fragmented, and the instrument records the mass-to-charge ratio ($m/z$) and abundance of every fragment. The output, a *spectrum*, is a set of peaks $\{(m_i, I_i)\}$ with continuous coordinates and variable cardinality. Each molecule and fragment has a definite mass, and peaks from different molecules can overlap. The fragmentation chemistry is written in the *differences* between positions ($\Delta m/z$). Two peptide fragments that differ by one amino acid are separated by that residue's mass, an isotope peak sits 1.003 units above its parent, and a lost water molecule shifts a fragment by 18.011. Every task in the field reads these relationships, from denoising, retrieval, search and rescoring to intensity prediction and chimeric-



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.

spectrum detection, and each is currently solved by a separate, largely hand-engineered pipeline (Kong et al., 2017; Käll et al., 2007; Gessulat et al., 2019; Bittremieux et al., 2022).

Deep learning reached this field through supervised task models (Gessulat et al., 2019; Gabriels et al., 2019; Bouwmeester et al., 2021; Yilmaz et al., 2022; Eloff et al., 2025; Lapin et al., 2025) and, recently, self-supervised spectrum encoders: DreaMS for small molecules (Bushuiev et al., 2025), MSBERT, LSM1-MS2 and PRISM (Zhang et al., 2024; Asher et al., 2024; Enveda Biosciences, 2024), Maldi Transformer (De Waele et al., 2025), MS1-MPM (Zhou et al., 2026), Casanovo Foundation (Sanders et al., 2025) and, concurrently with this work, InstaNovo-FM for proteomics (Nieuwoudt et al., 2026).

Two questions remain open (Table 3, Appendix B). **Does scale help?** In language, vision and protein sequences, scaling laws relate size, data and loss and guide how compute is spent (Kaplan et al., 2020; Hoffmann et al., 2022; Cheng et al., 2024). Nothing comparable exists for spectra: across 33 published spectrum models with disclosed sizes the median is about 64M parameters, no model sweeps size above 15M, and no fitted compute–loss relation has been reported (Appendix B); the one multi-size study, Maldi Transformer at 1.65M–14.84M, found no benefit from its largest size (De Waele et al., 2025), and a recent survey lists no scaling result (Carnoli et al., 2026). **Is absolute mass necessary?** Every published spectrum transformer feeds each peak's absolute $m/z$ into its embedding as sinusoidal or Fourier features, tokens or bins (Yilmaz et al., 2022; Eloff et al., 2025; Bushuiev et al., 2025; Lapin et al., 2025; Zhang et al., 2024; Enveda Biosciences, 2024; Bittremieux et al., 2022). Two models additionally add *relative* mass, the pairwise difference $m_i - m_j$, as an attention bias and report that it helps (Lapin et al., 2025; Bushuiev et al., 2025). None has removed absolute mass, so whether the relational structure alone suffices is unknown.

We answer both questions with Iona, a family of five transformer encoders (25.6M to 395.5M parameters) pretrained with one recipe on 100 million consensus MS/MS spectra by masked intensity prediction, in which mass enters *only* through a learned Fourier-feature bias on pairwise mass differences. Our contributions are:

1. **The first scaling-law study of mass spectrometry foundation models.** Over five sizes and 40 checkpoints, held-out loss follows $L(N, D) = AN^{-\alpha} + BD^{-\beta}$ ($R^2 = 0.99$) and the compute-efficient frontier places the optimal model beyond our largest (Section 4.1).
2. **Real-world utility in identification pipelines.** Iona embedding features add 20,763 peptide–spectrum matches at 1% FDR over a shuffled-embedding null on 18 unseen runs, and Iona denoising recovers 37–60% more peptides and 20–43% more protein groups from the weakest single-cell acquisitions without error inflation (Section 4.5).
3. **Capability grows with scale, and simple fine-tuning reaches specialised models.** Denoising, retrieval, chimera decomposition and chimericity improve with size (Section 4.3). We do not beat every specialised system, but fine-tuned Iona-400M surpasses a supervised Casanovo encoder by 0.033 AUROC on real chimeric spectra and matches Prosit on intensity prediction and MS2Rescore on rescoring (Section 4.4).
4. **The relative-mass design learns chemistry from data.** With no absolute $m/z$ and no annotation, the learned attention bias is enriched 2–3$\times$ at residue, neutral-loss and isotope mass differences (Section 4.2).
5. **The performance comes from pretraining, and nothing in the recipe is proteomics-specific.** Random-initialisation controls fall far behind at every size and a 2% pretraining checkpoint recovers most of the downstream gain; the recipe uses no peptide labels, precursor information or ion annotation (Sections 4.3 and 5).

All five Iona models (checkpoints) and code are available anonymously at https://anonymous.4open.science/r/iona-7E82/

## 2 Related Work

**Spectral foundation models.** DreaMS (Bushuiev et al., 2025) pretrains a 116M encoder on small-molecule spectra by masked $m/z$ classification plus retention-order prediction and reports that data quality beats quantity. MSBERT and LSM1-MS2 add contrastive learning to masked peak modelling (Zhang et al., 2024; Asher et al., 2024); PRISM reports, qualitatively, no levelling off

with data (Enveda Biosciences, 2024); MS1-MPM reconstructs masked precursor intensities (Zhou et al., 2026). In proteomics, Casanovo Foundation reuses a supervised sequencing encoder for spectrum-quality, chimericity and modification classification (Sanders et al., 2025), and the concurrent InstaNovo-FM (Nieuwoudt et al., 2026) pretrains one 89.5M encoder on a 1.63-billion-scan corpus by physics-aware masked reconstruction, at a single size and without a scaling study. *Our delta:* a systematic size sweep with a fitted loss surface and a ladder of transfer tasks with matched controls; we claim neither corpus-size nor parameter leadership.

**Relative mass-difference attention.** Pairwise bias on attention logits comes from AlphaFold2's pair representation (Jumper et al., 2021) and Graphormer (Ying et al., 2021). Pairwise Attention (Lapin et al., 2025) applies it to spectra: each $\Delta m_{ij}$ is expanded in Fourier features (Tancik et al., 2020), mapped by a small network to one scalar per head and added to the logit; the 29k-parameter bias raises de novo precision by 6.7 points, although the gain narrows on larger corpora. DreaMS uses the same construction. Both keep absolute $m/z$ in the peak embedding and train one size. Iona adopts this mechanism, removes the absolute-mass channel and predicts masked *intensity* rather than masked *mass*.

## 3 IONA MODEL

### 3.1 A RELATIVE-MASS-ONLY SPECTRUM ENCODER

A spectrum is $S = \{(m_i, I_i)\}_{i=1}^{N}$ with intensities normalised to $\bar{I}_i = I_i / \sum_j I_j$, so that it is a probability mass function over its own peaks. Each peak is a token whose embedding is a learned projection of $\log \bar{I}_i$ plus a mask indicator; *the token contains no $m/z$ value*, and there is no positional encoding, precursor token or absolute-mass side channel. Mass enters only through pairwise differences $\Delta m_{ij} = m_i - m_j$, mapped to Fourier features

$$\phi(x) = \big[\sin(2\pi f_1 x), \ldots, \sin(2\pi f_K x), \cos(2\pi f_1 x), \ldots, \cos(2\pi f_K x)\big], \tag{1}$$

with $K = 256$ frequencies log-spaced from $10^{-3}$ to 190 cycles per Da, and then, per head $h$, through a two-layer network with 32 hidden units to a scalar added to the attention logit:

$$b_{ij}^{(h)} = \mathrm{MLP}_h\big(\phi(\Delta m_{ij})\big), \qquad A_{ij}^{(h)} = \frac{\mathbf{q}_i^{(h)} \cdot \mathbf{k}_j^{(h)}}{\sqrt{d_h}} + b_{ij}^{(h)}. \tag{2}$$

The bias network is shared across layers, so each head learns one readable function $b_h(\Delta m)$ (Section 4.2). Fourier features are needed to resolve a residue mass to about 0.01 Da over a 2,000 Da span (Tancik et al., 2020). Two properties follow: the encoder is permutation-equivariant over peaks, and it is exactly invariant to a common shift of all coordinates, so it cannot memorise absolute positions and must transfer through relationships. Five sizes share this design and differ only in depth and width: 25.6M (8 layers, width 512), 49.8M (10, 640), 100.8M (13, 800), 202.9M (16, 1,024) and 395.5M (20, 1,280); full configurations are in Appendix C. Inputs are capped at 512 peaks, which covers 92.2% of corpus spectra.

### 3.2 MASKED INTENSITY PREDICTION AND PRETRAINING RECIPE

We mask the intensity embedding of about 50% of peaks $M$ with a learned token, keep $m/z$ , and predict the distribution of intensity over the masked peaks with a linear head giving logits $u_i$:

$$p_i = \frac{\exp u_i}{\sum_{j \in M} \exp u_j}, \qquad q_i = \frac{I_i}{\sum_{j \in M} I_j}, \qquad \mathcal{L}_{\mathrm{MIP}} = \sum_{i \in M} q_i \log \frac{q_i}{p_i}. \tag{3}$$

The model must therefore infer how much signal a peak carries from the visible intensities and from the mass relationships among all peaks. Peptide labels are never used. All sizes are trained for 540,423 steps (three epochs at an effective batch of 512) with AdamW (Loshchilov & Hutter, 2019), $\beta = (0.9, 0.95)$, weight decay 0.01, peak learning rate $1.3 \times 10^{-4}$, 2,000 warm-up steps then cosine decay, clipping at 1.0 and dropout 0.1; hyperparameters were fixed at 50M and reused unchanged. Training used 8–32 supercomputer nodes for two to three days per model; loss on 1M held-out spectra was logged at 40 post-warm-up checkpoints across the family.

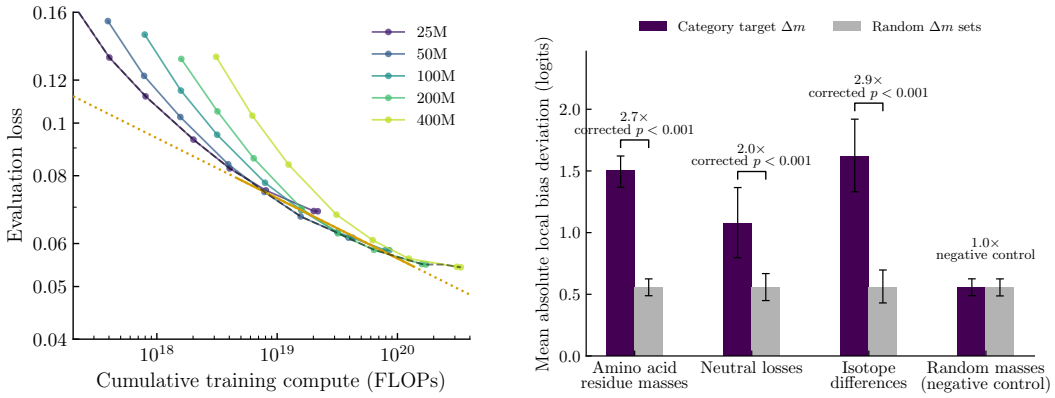



Figure 1: **Left:** held-out masked-intensity loss against cumulative training compute for the five sizes; each point is a post-warm-up checkpoint. Compute is the proxy $6ND$ ($D$ = nominal peak positions processed). Dashed black: empirical compute-efficient frontier. Gold: compute-optimal loss implied by the joint fit, solid where the optimal size lies within 25M–400M and dotted where it extrapolates beyond. **Right:** chemical enrichment of the learned relative-mass bias of Iona-50M: mean absolute local bias deviation at known mass differences versus equally sized random mass sets (bootstrap 95% intervals; Holm-corrected permutation $p$ over 10,000 random sets; negative control: two halves of the random null).

### 3.3 Corpus, downstream data and adaptation protocols

Iona is pretrained on MSConsensus-100M (Gao et al., 2026), a 100-million-spectrum release built from 1.01 PB of public raw data in 1,500 PRIDE repositories (Perez-Riverol et al., 2022). Identified spectra sharing a peptide, charge and fixed modification are clustered and a weighted-score binning algorithm emits several consensus spectra per cluster, suppressing instrument- and laboratory-specific batch effects; the release is 19–55× larger than existing consensus libraries (Wang et al., 2018). Downstream data (Appendix D) comprise a denoising set, two contrastive corpora, the nine-species yeast and mouse sets (Tran et al., 2017), a low-resolution HEK293 set, 8 + 18 HEK293/HCT116 runs with MSFragger candidates, seven Q Exactive Plus files, single K562 cells from an Orbitrap Astral (Guzman et al., 2024), and Orbitrap runs labelled for chimericity. We use four adaptation protocols throughout. *Frozen probe*: encoder fixed, head trained on per-peak outputs or a pooled embedding. *Fine-tuning*: encoder updated at a lower learning rate than the head (in chimera experiments the mass-bias network is frozen as a lookup table). *Contrastive specialisation*: supervised contrastive loss (Khosla et al., 2020) on spectra of the same modified peptide and charge, then a peptide *student* trained to reproduce the spectrum teacher's embedding. *Random-init control*: same architecture and recipe without pretraining. Task heads and recipes are in Appendix E.

## 4 Results

### 4.1 Pretraining scales as a power law, with the optimum beyond 400M

Held-out loss decreases monotonically with size at every checkpoint and at the end of training: 0.0688, 0.0615, 0.0583, 0.0549 and 0.0543 for 25M to 400M. The gain per doubling shrinks at a fixed three-epoch budget, from 0.0073 (25M→50M) to 0.0006 (200M→400M). Fitting all 40 post-warm-up checkpoints jointly with

$$L(N, D) = A\Big(\frac{N}{10^8}\Big)^{-\alpha} + B\Big(\frac{D}{10^{10}}\Big)^{-\beta} \tag{4}$$

gives $A = 0.053$, $\alpha = 0.14$, $B = 0.029$, $\beta = 0.57$, log-space $R^2 = 0.99$ (mean absolute error 1.1%). Both exponents are positive, so neither parameters nor exposure has saturated in the sampled regime, and the data exponent is four times the parameter exponent. The empirical frontier (Figure 1) tells the same story without the fit: each larger model overtakes the smaller ones as the budget grows,

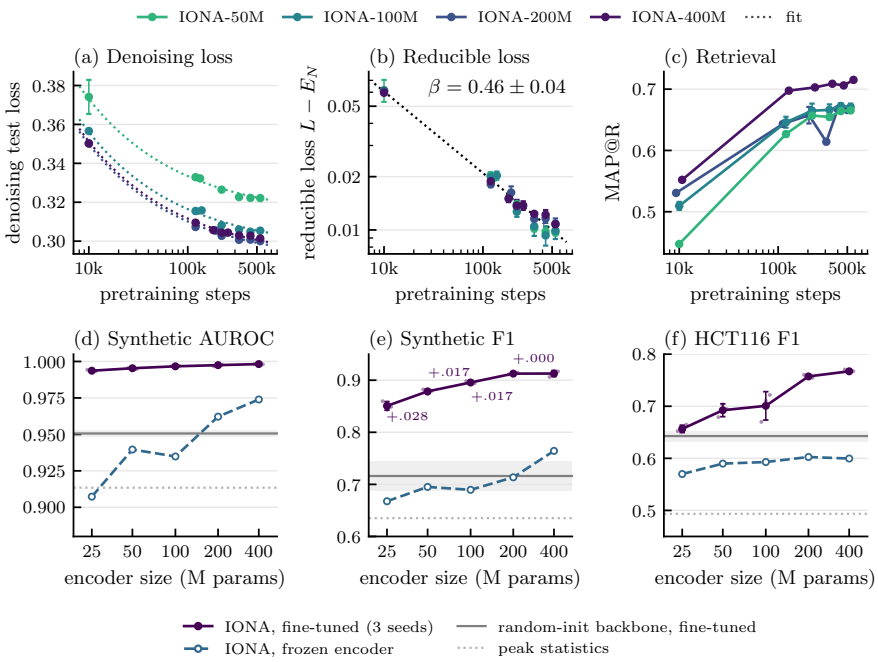



Figure 2: Downstream scaling ladder. **Top a,b:** fine-tuned denoising test loss against pretraining steps per size with the fit of Eq. 5 (dotted) and the reducible loss $L - E_N$ on log–log axes. **Top c:** contrastive retrieval MAP@R against pretraining checkpoint and size. **Bottom:** chimera decomposition against size: synthetic detection AUROC, synthetic assignment F1 with adjacent-size paired $\Delta$, and HCT116 F1; fine-tuned versus frozen with random-backbone and peak-statistics references.

and 400M defines the frontier at the largest budgets we trained. Solving the fit for the compute-optimal allocation puts the optimal size beyond 400M at our final budget (gold line dotted at the right edge), so we regard the 400M model as under-sized rather than saturated. Caveats: FLOPs are our best estimation, not actual hardware count, and the fit is in-sample over one run per size (Appendix G); the held-out prediction in Section 4.3 is our independent check that the behavior is predictive.

### 4.2 The learned bias is chemically structured

Iona receives neither chemical annotation nor absolute mass, so whatever it knows about chemistry must be expressed in the learned functions $b_h(\Delta m)$, which we read directly from the pretrained 50M model. For each target mass difference we take the absolute deviation of $b_h$ from its median at flanking offsets ($\pm$0.1–0.8 Da), average over heads and over the masses in a category, and compare with 10,000 random mass sets of equal size drawn from 0–200 Da away from known masses. The 19 amino-acid residues show a $2.72\times$ larger deviation than random masses, the 7 common neutral losses $1.95\times$ and the 5 isotope spacings $2.93\times$, each with Holm-adjusted $p = 3 \times 10^{-4}$, while the negative control gives $1.00\times$ (Figure 1, right). Without being told what an amino acid is, the model has placed sharp features in its attention bias exactly at the differences that connect fragments of one peptide, a fragment to its dehydrated form, and isotopes to each other: attention between two peaks is modulated by whether they are chemically related. Initialisation and 400M controls are in Appendix I.

### 4.3 Downstream scaling ladder

Each task below uses one recipe applied identically at every size (Figure 2, Table 1).

**Denoising.** Each peak is either a fragment of the analysed peptide or noise. Fine-tuning a per-peak head for four epochs from pretraining checkpoints between 10k and 540k steps, pretraining adds

Table 1: Scaling ladder table. Fine-tuned (FT) rows are seed means (3 seeds unless noted); frozen rows are single fits.

| Task | Metric | 25M | 50M | 100M | 200M | 400M | Control |
|---|---|---|---|---|---|---|---|
| Pretraining | held-out loss, 540k steps | 0.0688 | 0.0615 | 0.0583 | 0.0549 | **0.0543** | – |
| Denoising, FT | test AUROC, 540k encoder | – | 0.936 | 0.943 | **0.946** | 0.945 | from scratch 0.882–0.898 |
| Denoising, FT | loss floor $E_N$ of Eq. 5 | – | 0.3125 | 0.2955 | **0.2892** | 0.2906 | – |
| Retrieval, contrastive | MAP@R, 220k encoder | – | 0.657 | 0.665 | 0.658 | **0.703** | random init $\approx$ chance |
| Decomposition, FT | synthetic F1 | 0.851 | 0.879 | 0.896 | **0.913** | **0.913** | random backbone 0.716 |
| Decomposition, FT | HCT116 F1 | 0.657 | 0.692 | 0.701 | 0.757 | **0.767** | random backbone 0.643 |
| Chimericity, frozen MLP | AUROC, 60k labels | 0.714 | 0.770 | 0.733 | 0.764 | **0.811** | Casanovo frozen 0.802 |
| Reranking, 8 runs | net PSM gain over null | – | +1.3% | – | – | **+2.5%** | null $-0.1$ to $+0.3$% |
| Peak selection | peptides, % of original | – | 68.2 | 68.6 | 73.7 | **75.2** | matched top-$k$ 63.5–69.5 |

about $+0.05$ AUROC at every size over the same architecture from scratch; a 10k-step checkpoint (2% of pretraining) already recovers most of the gain, and gains flatten after about 330k steps (Appendix J). AUROC rises with size to 200M (0.946) and is marginally lower at 400M (0.945; seed SD 0.0005). The test *loss* gives a cleaner statement: across 29 (size, checkpoint) points,

$$L(N, S) = E_N + B\,(S/10^5)^{-\beta}, \qquad \beta = 0.46 \pm 0.04,\ R^2 = 0.997, \tag{5}$$

with one power law in pretraining steps $S$ shared by all sizes and one floor $E_N$ per size (Table 1). The pretraining benefit therefore transfers with a single exponent regardless of size, while size enters only through the floor, which improves to 200M and not beyond. Fitting without the 540k checkpoint predicts it to within 0.002 at every size, the held-out check promised above.

**Retrieval.** With one contrastive recipe (replicate corpus, 24 epochs, 3 seeds), MAP@R rises steeply with the encoder's pretraining checkpoint (50M: 0.447 at 10k to 0.627 at 120k steps), flattens after 220k, and at that checkpoint reads 0.657, 0.665, 0.658 and 0.703 for 50M to 400M. Identical training from random initialisation stays at chance at every size (MAP@100 about 0.01 versus 0.34–0.41; six seeds), so pretraining is necessary under this recipe (Appendix K).

**Chimera decomposition.** Co-isolated peptides produce chimeric spectra that are identified at half the rate of clean ones (Houel et al., 2010; Michalski et al., 2011); the useful output is a per-peak assignment of fragments to the two sources. With one recipe (60,000 exposures, 40k synthetic MSConsensus mixtures then 20k half-real, three seeds), fine-tuned Iona detects two-component mixtures with AUROC 0.994–0.998 at every size against 0.913 for an intensity-split heuristic, and assignment F1 scales cleanly: synthetic F1 rises from 0.851 at 25M to 0.913 at 200M with paired adjacent-size gains of $+0.028$, $+0.017$, $+0.017$, then saturates ($+0.000$ to 400M), while F1 on 1,376 search-corroborated HCT116 pairs rises through every size from 0.657 to 0.767. A random-initialised 200M backbone that keeps only the pretrained bias table reaches 0.716, and the fine-tuning gain over frozen encoders grows with size on real data ($+0.087$ at 25M to $+0.168$ at 400M). Limits: synthetic mixtures come from the pretraining source, only three HCT116 acquisitions contribute real pairs, and on an external HEK set (200M) the advantage over the random backbone vanished (Appendix Q).

**Chimericity classification.** With frozen encoders, an identical MLP head and 60k labels, primary AUROC reads 0.714, 0.770, 0.733, 0.764 and 0.811 for 25M to 400M. The trend is upward with the largest step from 200M to 400M ($+0.047$ [$+0.043$, $+0.051$]), but it is not monotone (100M below 50M) and pretraining budgets were not audited across sizes, so we read it as a family pattern rather than a clean parameter effect; only 400M reaches the supervised Casanovo reference.

**Reading the ladder.** Reconstruction, retrieval, real-data decomposition, chimericity and reranking gain all improve from 200M to 400M; the denoising floor and synthetic decomposition, the tasks closest to the pretraining objective on consensus-derived data, saturate at 200M. A further task, fragment-intensity prediction from a peptide's theoretical ion ladder, reaches a median spectral angle of 0.909 at 200M and 400M, matching the Prosit RNN (Appendix N).

**Performance comes from pretraining.** Every task carries a control with the same architecture and recipe but no pretraining, and every control falls far behind: denoising from scratch is about $0.05$ AUROC lower at every size, contrastive retrieval from random initialisation stays at chance, and a random backbone on decomposition reaches F1 0.716 against 0.913. The gain also arrives early and largely without adaptation: a 10k-step checkpoint (2% of pretraining) recovers most of

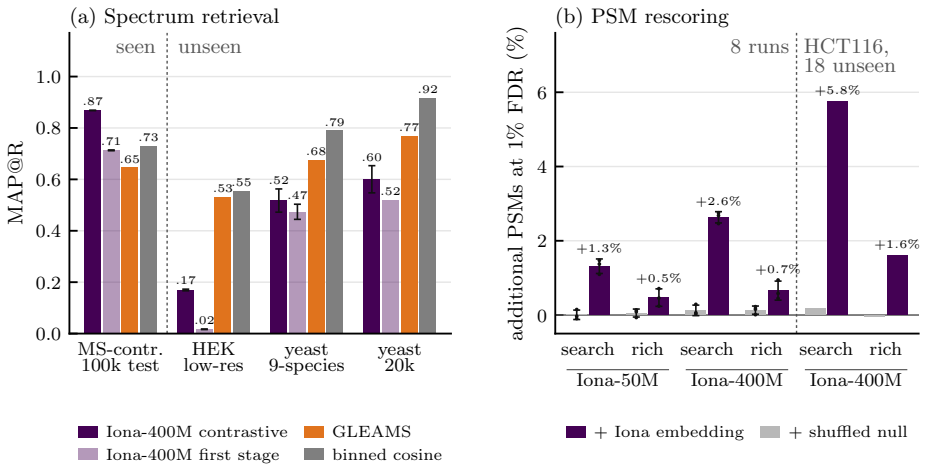



Figure 3: **Left:** spectrum retrieval MAP@R (same modified peptide and charge) on the in-distribution test set and four unseen benchmarks for contrastively fine-tuned Iona-400M, its first-stage model, GLEAMS and binned cosine at its better bin width; Iona values are 3-seed means except mouse and human (validation-selected seed). **Right:** additional PSMs at 1% FDR from Iona embedding features versus the shuffled-embedding null, by cohort, feature set and encoder size.

the denoising benefit, and the *frozen* 400M encoder with an MLP head already matches the frozen supervised Casanovo encoder on chimericity (Section 4.4). Fine-tuning refines what pretraining built; it does not create it.

### 4.4 Simple fine-tuning reaches specialised models

We do not claim to beat every specialised system; the question is whether one encoder with simple adaptation reaches them. **Retrieval.** Contrastive specialisation of Iona-400M reaches MAP@R 0.868 on its own distribution, far above GLEAMS (Bittremieux et al., 2022) (0.646) and cosine on 0.1 Da bins (Stein & Scott, 1994) (0.729). On unseen data the picture is mixed (Figure 3, left): Iona leads on nine-species human (0.876 versus 0.841 and 0.809), is level with GLEAMS on mouse (0.846 versus 0.834, both below binned cosine at 0.916), and loses on high-resolution yeast (0.600 versus 0.770 and 0.916) and, badly, on low-resolution ion-trap HEK spectra (0.170 versus 0.530 and 0.554). Contrastive fine-tuning on one acquisition distribution narrows the embedding, whereas the *frozen* encoder with ABTT post-processing (Mu & Viswanath, 2018) transfers to yeast better than any fine-tuned model (0.71, Appendix H). The peptide student of the same teacher identifies peptides with no mass filter for 92% of in-distribution and 93% of unseen mouse queries, against 20% and 24% for yHydra (Altenburg et al., 2026) (Appendix L).

**Chimericity against Casanovo.** We compare Iona-400M with the Casanovo v5.2.0 encoder (Yilmaz et al., 2024), the supervised proteomics encoder used as a foundation model by Sanders et al. (2025), under one protocol (same top-512 peaks, pooling, MLP head, three head seeds, validation-only selection, and search-derived labels: positive if two peptides are accepted at $q \leq 0.01$). The primary endpoint is the mean within-acquisition AUROC over four test acquisitions from a different day, column and a human–yeast mixture, with paired block-bootstrap intervals. Both pretrained encoders are far above hand-crafted baselines (0.660–0.688; Figure 4). Frozen, Iona-400M matches or slightly exceeds Casanovo (0.811 versus 0.802, $+0.009$). Under equal-data-budget adaptation the gap opens: fully fine-tuned Iona reaches 0.854 against 0.820, $\Delta = +0.033$ [$+0.031$, $+0.036$], on every test role. Disclosures: 400M was promoted after the family sweep on the same test acquisitions, while the pre-registered 200M analysis had frozen Casanovo ahead by 0.038 and adapted Iona ahead by 0.021, so the adapted direction is pre-registered and the 400M magnitudes are descriptive; Casanovo has 20$\times$ fewer trainable parameters; and with $\leq$10k labels Casanovo leads (Appendix Q).

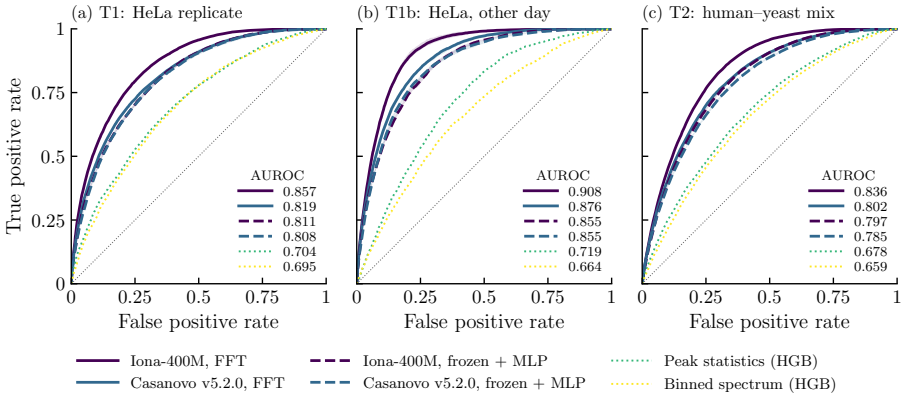



Figure 4: Real-spectrum chimericity, 60k labels: ROC curves of Iona-400M and the supervised Casanovo v5.2.0 encoder, frozen with an MLP head and fully fine-tuned (FFT) under equal budgets, with two hand-crafted baselines, on three test roles (T1: same-sample HeLa replicate; T1b: HeLa on a different day and column; T2: human–yeast mixtures). Curves are averaged over acquisitions and three head seeds; legend values are the mean within-acquisition AUROC. Paired $\Delta$ (Iona $-$ Casanovo) with 95% block-bootstrap intervals: frozen $+0.009$ [$+0.006$, $+0.013$], fine-tuned $+0.033$ [$+0.031$, $+0.036$] (Appendix Q).

### 4.5 Iona in real identification pipelines

**Rescoring with a shuffled-embedding null.** Database search assigns each spectrum candidate peptides with engine scores, and Percolator-style rescoring (Käll et al., 2007) separates true from false matches. For MSFragger's top-10 candidates (Kong et al., 2017) we compute the cosine between the spectrum embedding and each candidate's peptide embedding and four within-spectrum relative features (rank, margin, $z$-score, top-2 gap), append them to either MSFragger's scores or a rich table from four engines, train a per-run linear rescorer with 3-fold cross-validation, and count PSMs at 1% FDR. The control is a *shuffled-embedding null*: the same five features computed from a random other spectrum's embedding, identical in dimension and marginal distribution but carrying no information about the match. The null adds nothing ($-0.1$ to $+0.3$%) while Iona features add identifications in every setting (Figure 3, right). On 18 HCT116 runs that none of our models saw, 400M features raise PSMs from 370,982 to 392,350 against 371,587 for the null: a net $+20{,}763$ (5.6% of base) on MSFragger scores and $+8{,}624$ on the rich table, with 50M giving about half the 400M gain. Shuffled target/decoy labels yield zero PSMs, so the gain is not label leakage. Our rescorer with rich features alone matches the full MS2Rescore pipeline with MS$^2$PIP and DeepLC predictions (Buur et al., 2024; Gabriels et al., 2019; Bouwmeester et al., 2021) (128.9k versus 128.2k PSMs), and adding Iona features to that pipeline gives a small net $+251$ over its null (Appendix M). The two are complementary: the shift-invariant encoder cannot verify absolute fragment positions, which the search engine supplies.

**Peak selection at a matched budget.** On seven Q Exactive Plus files searched with MSFragger, MSBooster and Percolator (Yang et al., 2023), Iona-denoised spectra yield more identified peptides than a control keeping the same number of peaks per spectrum chosen by intensity, with a margin that grows with size (75.2% versus 69.5% of the original count at 400M; paired $t$-tests across files, BH $q < 0.05$). Both filtered arms stay below the unfiltered originals, so this measures information per retained peak (Appendix O).

**Weak single-cell acquisitions.** Where filtering does pay is single-cell proteomics, where proteins cannot be amplified and some cells yield spectra with signal barely above the noise floor (Budnik et al., 2018; Slavov, 2021). A per-peak head on the frozen 400M encoder, its policy fixed before these cells were searched, cleaned four single K562 cells from an Orbitrap Astral (Guzman et al., 2024), re-searched with an otherwise unchanged pipeline (Table 2). In the two cells with the fewest original identifications the model recovers 37% and 60% more unique peptides and 20–43% more protein groups, leaves the two deeper cells within $\pm 2$%, and shows no error inflation under entrapment (Wen et al., 2025) (pooled FDP 0.88% against 1.10%). A 1% intensity floor recovers 30% and

Table 2: Single K562 cells (Orbitrap Astral, 8 Th DIA): unique peptides at 1% peptide-level $q$, protein groups and pooled entrapment FDP at $q = 0.01$; the denoising policy was fixed before these cells were searched (Appendix P).

| Unique peptides | blo5run12 | blo5run7 | blo5run3 | blo5run38 | Pooled FDP |
|---|---|---|---|---|---|
| Original search | 718 | 947 | 1,970 | 2,707 | 1.10% |
| Intensity floor 1% (baseline) | 935 (+30.2%) | 1,282 (+35.4%) | 1,957 (−0.7%) | 2,609 (−3.6%) | 0.99% |
| Iona-400M denoising | **984 (+37.0%)** | **1,518 (+60.3%)** | 2,009 (+2.0%) | 2,694 (−0.5%) | **0.88%** |
| Protein groups, original → Iona | 495 → 596 | 569 → 816 | 975 → 1,052 | 1,246 → 1,323 | |

35% in the same cells, so the learned margin over the strongest simple baseline is 7 and 25 points, and the gain is confined to cells below about 1,000 original peptides, exactly the acquisitions that current pipelines serve worst. With four cells this is a pilot (Appendix P).

## 5 Discussion and Limitations

**Pretraining loss is not a proxy for utility.** Loss scales regularly and the frontier says a larger model would spend our compute better, yet the downstream ladder is task dependent. Fixed adaptation budgets, the three-epoch schedule and the diversity of a consensus corpus are untested candidate explanations for the saturating tasks; scale must be evaluated per task.

**The model learns chemistry from data alone.** Every task above, including intensity prediction where the peptide reaches the model only as differences between its theoretical fragments, was solved by an encoder that never sees an absolute $m/z$, and the learned bias shows where the chemistry went. The stronger claim, that relative-only matches or beats an identical absolute+relative encoder at matched compute, needs an ablation we have not run; until then the architecture is a design choice whose sufficiency, not superiority, is shown.

**Nothing in the recipe is proteomics-specific.** The input is a set of ($m/z$, intensity) pairs, the objective is the intensity distribution over masked peaks, and pretraining uses no peptide labels, precursor information or annotation; the only proteomic ingredient is the corpus. The same recipe therefore applies unchanged to metabolomics, lipidomics or MALDI spectra given a comparable corpus; whether the scaling law itself transfers across domains is untested.

**Domain shift and concurrent work.** Contrastive fine-tuning produced an embedding that loses to binned cosine on external and low-resolution data while the frozen encoder transfers better, which points to the adaptation recipe. The concurrent InstaNovo-FM trains one 89.5M absolute-mass encoder on a larger annotated corpus; a same-protocol comparison is the obvious next experiment (Appendix J).

**Limitations.** One pretraining run per size; the scaling fit is in-sample with only the denoising held-out check; denoising scores reload about 0.014 lower than logged and the recipe was selected on test AUROC; the 400M chimericity comparison is post hoc relative to the pre-registered 200M one; decomposition uses synthetic mixtures from the pretraining source and partial search-derived labels; the single-cell pilot has four cells; the 18-run reranking result has one seed; pretraining overlap with downstream acquisitions is unaudited.

## 6 Conclusion

We trained the first size-swept family of self-supervised mass spectrometry foundation models and fitted the first scaling law for the modality; the compute-efficient frontier places the optimal model beyond our 400M encoder, so we aim next to train Iona beyond 1B parameters. Capability grows with size on most tasks, and although Iona does not beat every specialised system, simple fine-tuning of one encoder surpasses a supervised proteomics encoder on real chimeric spectra, matches Prosit and MS2Rescore, adds tens of thousands of confident identifications to database search and recovers 37–60% more peptides from the weakest single-cell acquisitions. Random-initialisation controls show that this capability comes from pretraining, not fine-tuning. The encoder never sees an absolute mass, yet it learns chemical reality from data; and because the recipe contains nothing specific to proteomics, we expect it to carry over to other kinds of mass spectra.

### ETHICS STATEMENT

All data are public cell-line proteomics datasets (HeLa, HEK293, HCT116, K562, yeast, mouse) cited by accession in Appendix D, plus local cell-line acquisitions; no patient-derived or personally identifiable data were used. The models predict properties of spectra and carry no foreseeable dual-use risk beyond existing proteomics software.

### REPRODUCIBILITY STATEMENT

All code used to generate Iona and the evaluations are available at https://anonymous.4open.science/r/iona-7E82/ Pretraining exports are in Anonymous HF, links provided in the supplementary; configurations in Appendix C; recipes in Appendix E; accessions and splits in Appendix D. The chimera experiments were frozen before test inference with integrity manifests. Checkpoints are released under Apache-2.0 and the corpus under CC BY 4.0.

### AI USE STATEMENT

Large language model tools were used to edit and polish the text of this manuscript. All experiments, analyses, results and claims were designed, executed and verified by the authors, who take full responsibility for the content.

## REFERENCES

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.

Gennady Asher et al. LSM1-MS2: A self-supervised foundation model for tandem mass spectrometry applications, encompassing extensive chemical property predictions and spectral matching. *ChemRxiv*, 2024. doi: 10.26434/chemrxiv-2024-k06gb. Preprint. VERIFY author list.

Wout Bittremieux, Damon H. May, Jeffrey Bilmes, and William Stafford Noble. A learned embedding for efficient joint analysis of millions of mass spectra. *Nature Machine Intelligence*, 4: 387–395, 2022. doi: 10.1038/s42256-022-00467-7. VERIFY DOI.

Robbin Bouwmeester, Ralf Gabriels, Niels Hulstaert, Lennart Martens, and Sven Degroeve. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods*, 18(11):1363–1369, 2021. doi: 10.1038/s41592-021-01301-5.

Bogdan Budnik, Ezra Levy, Guillaume Harmange, and Nikolai Slavov. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biology*, 19:161, 2018. doi: 10.1186/s13059-018-1547-5.

Roman Bushuiev, Anton Bushuiev, Niek F. de Jonge, Adamo Young, Fleming Kretschmer, Patroklos Samaras, Jiří Sedlacek, et al. Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. *Nature Biotechnology*, 2025. doi: 10.1038/s41587-025-02663-3. VERIFY author list.

Louise M. Buur, Arthur Declercq, Marina Strobl, Robbin Bouwmeester, Sven Degroeve, Lennart Martens, Viktoria Dorfer, and Ralf Gabriels. MS2Rescore 3.0 is a modular, flexible, and user-friendly platform to boost peptide identifications, as showcased with MS Amanda 3.0. *Journal of Proteome Research*, 23(8):3200–3207, 2024. doi: 10.1021/acs.jproteome.3c00785. VERIFY pages and DOI.

Andrea J. Carnoli, Francisco Padilla-Gonzalez, Leonie M. van den Bulk, Dimitri Korporaal, Martin Alewijn, Marco H. Blokland, and Bas H. M. van der Velden. Foundation models for liquid chromatography-high-resolution mass spectrometry: A new era beyond labeled datasets. *Analytical Chemistry*, 98(31):22586–22594, 2026. doi: 10.1021/acs.analchem.6c01825. VERIFY author first names.

Xingyi Cheng, Bo Chen, Pan Li, Jing Gong, Jie Tang, and Le Song. Training compute-optimal protein language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Gaetan De Waele, Gerben Menschaert, Peter Vandamme, and Willem Waegeman. Pre-trained Maldi Transformers improve MALDI-TOF MS-based prediction. *Computers in Biology and Medicine*, 186:109695, 2025. doi: 10.1016/j.compbiomed.2025.109695.

Markus Ekvall, Patrick Truong, Wassim Gabriel, Mathias Wilhelm, and Lukas Käll. Prosit Transformer: A transformer for Prediction of MS2 Spectrum Intensities. *Journal of Proteome Research*, 21(5):1359–1364, May 2022. ISSN 1535-3893, 1535-3907. doi: 10.1021/acs.jproteome.1c00870. URL https://pubs.acs.org/doi/10.1021/acs.jproteome.1c00870.

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.

Enveda Biosciences. PRISM: A foundation model for life's chemistry. https://enveda.com/prism-a-foundation-model-for-lifes-chemistry/, 2024. Company research report, 28 May 2024; not peer reviewed.

Ralf Gabriels, Lennart Martens, and Sven Degroeve. Updated MS$^2$PIP web server delivers fast and accurate MS$^2$ peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Research*, 47(W1):W295–W299, 2019. doi: 10.1093/nar/gkz299.

Yu Gao, Chris Grams, Chang Liu, and Michael Papka. MSBench. https://doi.org/10.7910/DVN/UXSC5H, 2026. Consensus MS/MS spectra of bottom-up proteomics for training foundation models.

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 (6):509–518, June 2019. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0426-7. URL https://www.nature.com/articles/s41592-019-0426-7.

Ulises H. Guzman, Ana Martínez-Val, Zilu Ye, Eugen Damoc, Tabiwang N. Arrey, Anna Pashkova, Santosh Renuse, Eduard Denisov, Johannes Petzoldt, Amelia C. Peterson, Florian Harking, Ole Ostergaard, Rasmus Rydbirk, Simon Bekker-Jensen, Anna Harder, Jesper Wieland, Eva Seizinger, Karsten Wetzel, Kevin Meyer, Hamish Stewart, Christian Hock, Alexander Makarov, Vlad Zabrouskov, and Jesper V. Olsen. Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nature Biotechnology*, 42:1855–1866, 2024. doi: 10.1038/s41587-023-02099-7. VERIFY author list and pages.

Adam Hájek, Helge Hecht, Elliott J. Price, and Aleš Křenek. SpecTUS: Spectral translator for unknown structures annotation from EI-MS spectra. *arXiv preprint arXiv:2502.05114*, 2025. VERIFY author list.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Stéphane 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.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Lukas Käll, Jesse D. Canterbury, Jason Weston, William Stafford Noble, and Michael J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–925, 2007. doi: 10.1038/nmeth1113.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

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*, volume 33, 2020.

Andy T. Kong, Felipe V. Leprevost, Dmitry M. Avtonomov, Dattatreya Mellacheruvu, and Alexey I. Nesvizhskii. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, 14(5):513–520, 2017. doi: 10.1038/nmeth.4256.

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 (8):3722–3730, 2025. doi: 10.1021/acs.jproteome.5c00063.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1702.01417.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check, 2020. URL https://arxiv.org/abs/2003.08505.

Mechiel Nieuwoudt, Marco Reverenna, Dhanush Patel, Rachel Catzel, Ivan H. J. Houngue, Jaeden 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. Preprint, posted 8 September 2026.

Yasset Perez-Riverol, Jingwen Bai, Chakradhar Bandla, David García-Seisdedos, Suresh Hewapathirana, Selvakumar Kamatchinathan, Deepti J. Kundu, Ananth Prakash, Anika Frericks-Zipper, Martin Eisenacher, Mathias Walzer, Shengbo Wang, Alvis Brazma, and Juan Antonio Vizcaíno. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research*, 50(D1):D543–D552, 2022. doi: 10.1093/nar/gkab1038.

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. URL https://arxiv.org/abs/2505.10848.

Aditya Divyakant Shrivastava, Neil Swainston, Soumitra Samanta, Ivayla Roberts, Marina Wright Muelas, and Douglas B. Kell. MassGenie: A transformer-based deep learning method for identifying small molecules from their mass spectra. *Biomolecules*, 11(12):1793, 2021. doi: 10.3390/biom11121793.

Nikolai Slavov. Single-cell protein analysis by mass spectrometry. *Current Opinion in Chemical Biology*, 60:1–9, 2021. doi: 10.1016/j.cbpa.2020.04.018.

Stephen E. Stein and Donald R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9):859–866, 1994. doi: 10.1016/1044-0305(94)87009-8.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains, 2020. URL `https://arxiv.org/abs/2006.10739`.

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.

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, 2018. doi: 10.1016/j.cels.2018.08.004.

Bo Wen, Jack Freestone, Michael Riffle, Michael J. MacCoss, William Stafford Noble, and Uri Keich. Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment. *Nature Methods*, 22:1454–1463, 2025. doi: 10.1038/s41592-025-02719-x.

Kevin L. Yang, Fengchao Yu, Guo Ci Teo, Kai Li, Vadim Demichev, Markus Ralser, and Alexey I. Nesvizhskii. MSBooster: improving peptide identification rates using deep learning-based features. *Nature Communications*, 14:4539, 2023. doi: 10.1038/s41467-023-40129-9.

Melih Yilmaz, William Fondrie, Wout Bittremieux, Sewoong Oh, and William Stafford Noble. De novo mass spectrometry peptide sequencing with a transformer model. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25514–25522, 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.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Hailiang Zhang et al. 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. VERIFY author list and pages.

Yu Zhou, Vladimir Vutov, Susmita Ghosh, Fabienne Meier-Abt, Sibylle Pfammatter, Sandra Kummer, Odit Gutwein, Rosary Yao, Thorsten Zenz, Matthias Gunzer, Junyan Lu, and Jianxu Chen. Beyond binning: Resolution-preserving MS1 pretraining for clinical proteomics classification. In *Proceedings of the 17th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB)*, 2026. doi: 10.1145/3807503.3819478.

# A Appendix

# B Model-size census of the field

The census was compiled on 18 September 2026 from 420 sources and covers neural models trained primarily on mass-spectrometry data (MS/MS, MS1, GC-MS, MALDI). Of 33 models with numeric sizes, the median is about 64M parameters and the range is 1.3M–400M: 6 models below 10M, 9 between 10M and 50M, 7 between 50M and 100M, 9 between 100M and 200M, none between 200M and 350M, and 2 above 350M. Both models above 200M were trained on simulated

spectra: MassGenie (Shrivastava et al., 2021) (about 400M, in-silico spectra from ZINC) and SpecTUS (Hájek et al., 2025) (354M, synthetic pretraining then NIST20 fine-tuning). The largest model trained end-to-end on experimental MS/MS is InstaNovo+ at 170M (Eloff et al., 2025); the largest self-supervised model on experimental spectra with a stated size is DreaMS at 116M (Bushuiev et al., 2025). Corpus size and model size are decoupled: InstaNovo-FM trains on 1.63B scans at 89.5M parameters, PRISM on 1.2B spectra at an undisclosed size. Iona-400M therefore ties the all-time field maximum and is 2.3× the experimental-data record; the 200M→400M step of our sweep crosses a regime no published spectrum model has occupied.

Table 3: Representative spectrum models (Yilmaz et al., 2024; Eloff et al., 2025; Lapin et al., 2025; Bushuiev et al., 2025; Enveda Biosciences, 2024; De Waele et al., 2025; Nieuwoudt et al., 2026); full census in Appendix B. Iona is the only model with relative mass alone and the only size sweep above 15M parameters.

| Model | Params | Objective (training data) | Mass encoding | Size sweep |
|---|---|---|---|---|
| Casanovo | 47M | supervised de novo (30M PSMs) | absolute | no |
| InstaNovo / + | 95M / 170M | supervised de novo (PSMs) | absolute | no |
| Pairwise Attention | 48M | supervised de novo (PSMs) | absolute + relative | no |
| DreaMS | 116M | masked $m/z$ + RT order (GNPS) | absolute + relative | no |
| PRISM | undisclosed | masked peak masses (1.2B spectra) | absolute (tokens) | no |
| Maldi Transformer | 1.7–14.8M | peak shuffling (MALDI-TOF) | absolute | 4 ($\leq$15M) |
| InstaNovo-FM | 89.5M | masked reconstr. (185M scans) | absolute | 2 (confounded) |
| **Iona (ours)** | **26M–396M** | masked intensity KL (100M cons.) | **relative only** | **5 sizes** |

Table 4: Spectrum models with disclosed or computable parameter counts (S = stated, C = computed from configuration, B = benchmark or secondary source). FM = self-supervised foundation model; DeNovo = supervised sequencing; Gen = small-molecule generative or similarity model.

| Model | Params | Class | Domain / data | Year | Conf. |
|---|---|---|---|---|---|
| MassGenie | ∼400M | FM | small molecules; 21M in-silico spectra | 2021 | S |
| SpecTUS | 354M | FM | GC-MS; synthetic pretraining + NIST20 | 2025 | S |
| FRIGID | 182M | Gen | small molecules (decoder over frozen encoder) | 2026 | S |
| InstaNovo+ | 170M | DeNovo | proteomics; experimental PSMs | 2025 | S |
| LSM1-MS2 | 137–205M | FM | small molecules; >100M spectra (commercial) | 2024 | C |
| MS-GPT | 150M | Gen | small molecules | 2026 | S |
| MS2Mol / MS-BART / TTT-BART | ∼139M | Gen | small molecules; BART-base | 2023–25 | B |
| DreaMS | 116M | FM | small molecules; GNPS | 2025 | S |
| MSBERT | 116M total (13M encoder) | FM | small molecules; GNPS/MoNA | 2024 | C |
| InstaNovo | 95M | DeNovo | proteomics; MassIVE-KB | 2025 | S |
| InstaNovo-FM | 89.5M | FM | proteomics; 1.63B scans | 2026 | S |
| IDSL_MINT | 80M | FM | small molecules | 2024 | S |
| AdaNovo | 66M | DeNovo | proteomics | 2024 | B |
| Pairwise Attention | 48M | DeNovo | proteomics | 2025 | S |
| Casanovo v2 | 47M | DeNovo | proteomics; MassIVE-KB | 2024 | S |
| $\pi$-HelixNovo | 47M | DeNovo | proteomics | 2024 | B |
| $\pi$-PrimeNovo | ∼32M | DeNovo | proteomics | 2025 | C |
| Maldi Transformer | 1.65 / 3.27 / 6.92 / 14.84M | FM | MALDI-TOF | 2025 | S |
| NovoB | ∼12M | DeNovo | proteomics | 2024 | S |
| CoSpred | ∼7M | FM | proteomics; 7M spectra | 2025 | S |
| PointNovo | 4.8M | DeNovo | proteomics | 2021 | B |
| GLEAMS | 1.7M | Gen | proteomics; siamese CNN | 2022 | C |
| MS2DeepScore | 1.3M | Gen | small molecules | 2021 | S |
| PRISM, pUniFind, LSM-MS2 (2025), MS1-MPM | undisclosed | FM | large corpora | 2024–26 | U |

## C Model family and corpus

**MSConsensus-100M.** The full MSConsensus corpus holds 110,209,043 consensus MS/MS spectra built from 1.01 PB of raw data in 1,500 PRIDE repositories. Identified spectra sharing a (peptide, charge, fixed modification) key are clustered and a weighted-score binning algorithm (WSBIN) emits 5–100 consensus spectra per cluster; each consensus peak's $m/z$ is the intensity-weighted mean of the contributing peaks in a bin and its intensity a trimmed mean scaled by the bin's weighted support. The machine-learning release used here holds 100M training, 1M validation and 1M test spectra (Parquet, 190 GB) with $m/z$ and intensity arrays, precursor charge (1–7, mean 2.66) and peptide string; spectra contain 1–5,505 peaks (median 207, mean 257). The full release on Harvard Dataverse additionally carries per-peak ion annotations with confidence tiers, provenance metadata (instrument, vendor, species, source accession) and the `msz` random-access container. Splits were

Table 5: The Iona family parameter counts. All sizes use 256 Fourier frequencies log-spaced from $10^{-3}$ to 190 cycles per Da, a 32-unit per-head bias network, dropout 0.1, feed-forward width four times the hidden size, and a 512-peak input cap.

| Model | Parameters | Layers | Hidden | Heads | FFN |
|---|---|---|---|---|---|
| Iona-25M | 25,616,393 | 8 | 512 | 8 | 2,048 |
| Iona-50M | 49,813,771 | 10 | 640 | 10 | 2,560 |
| Iona-100M | 100,785,291 | 13 | 800 | 10 | 3,200 |
| Iona-200M | 202,858,513 | 16 | 1,024 | 16 | 4,096 |
| Iona-400M | 395,525,141 | 20 | 1,280 | 20 | 5,120 |

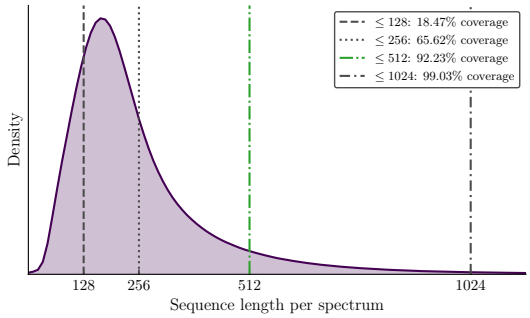



Figure 5: Distribution of spectrum lengths in MSConsensus-100M. Vertical lines show the fractions of spectra covered by cut-offs of 128, 256, 512 and 1,024 peaks; 512 covers 92.23%.

built with seed 42 and 1% validation and test fractions from the 90 source consensus files; the manifest records the SHA-256 of every shard.

## D Downstream datasets and accessions

Table 6: Downstream datasets. "Unseen" means excluded from the task-specific training data; overlap with the MSConsensus source accessions has not been audited for any row.

| Use | Dataset | Size | Notes |
|---|---|---|---|
| Denoising | `ms-denoise-100k` | 8,567 test spectra | up to 512 peaks; noise ≈53% of test peaks |
| Contrastive stage 1 | `ms2-peptide-replicate-retrieval` | – | peptide-replicate groups |
| Contrastive stage 2, retrieval | `ms-contrastive-100k` | 25,137 test queries | in-distribution benchmark |
| Retrieval (unseen) | nine-species yeast (Tran et al., 2017) | 86,184 / 20,019 | high-resolution; groups $\geq$2, capped at 20 |
| Retrieval (unseen) | HEK293 low-resolution | 27,637 | 1% FDR MSFragger PSMs; ion-trap MS2 |
| Peptide student (unseen) | nine-species mouse | 19,710 queries | high-resolution |
| Reranking | `psm-rerank-hek-hct116`, 8 runs | 332,822 spectra | 6 HEK293 + 2 HCT116; MSFragger top-10 |
| Reranking (unseen) | HCT116, 18 runs | 1,954,190 spectra | none seen by any model |
| Peak selection | PXD010540 | 7 files | Q Exactive Plus |
| Single cell | PXD070201 | 4 + 4 confirmation cells | K562 and HeLa; Orbitrap Astral, 8 Th DIA |
| Chimericity | PXD045500, PXD037985, local HCT116 | 60k train; 4 primary test acq. | Exploris 480 and Fusion Lumos |
| Decomposition | MSConsensus mixtures; HCT116 pairs | 5,000 synthetic; 1,376 real | search-corroborated pairs |
| Intensity prediction | Prosit 2020 HCD (Gessulat et al., 2019) | 9.9M / 2.8M / 1.42M | published splits |

## E Adaptation recipes and task heads

**Denoising.** Per-peak MLP head (width 512), learning rate $2 \times 10^{-4}$, encoder learning rate $0.5\times$ the head's, 4 epochs, effective batch 12. This configuration won a 216-configuration search at 50M and was replicated at every size. From scratch: encoder learning rate $1.0\times$. Three seeds per point (five at 50M/540k); seeds change head initialisation, data order and dropout.

**Contrastive spectrum embedding.** Mean+max pooling over peaks and a projection; supervised contrastive loss with temperature 0.002, learning rate $10^{-4}$, 64 peptide groups $\times$ 4 replicates per batch. Stage 1 on the replicate corpus (12 epochs at 400M, 24 at 50M); stage 2 one epoch of

`ms-contrastive-100k`. The reported seed is the one with the best *validation* MAP@R (400M: seed 0, 0.8639 vs 0.8629 and 0.8634).

**Peptide student.** Transformer over the peptide sequence and charge, trained by mean-squared error to the teacher's spectrum embedding; three seeds.

**Reranking.** Five features per candidate (cosine, margin to best other, rank, within-spectrum $z$-score, top-2 gap). Per-run linear rescorer, 3-fold cross-validation by spectrum, iterative target/decoy labels; PSMs at 1% FDR counted over runs. MS2Rescore v4.0.2 with default MS$^2$PIP and DeepLC features for the external comparison. Shuffled-embedding null: the same features computed with a random other spectrum's embedding.

**Chimericity classification.** Both encoders receive the top 512 peaks by intensity sorted by $m/z$; embedding is the masked mean of per-peak outputs (Casanovo's global token excluded). Heads: MLP with hidden width 512, or a linear probe; three head seeds; validation-only selection on a HeLa replicate and a mouse IP run. Adaptation: 60,000 exposures, effective batch 64, learning rate from $\{10^{-5}, 5\times10^{-5}, 2\times10^{-4}\}$ selected on validation (Iona $5\times10^{-5}$, Casanovo $2\times10^{-4}$); Iona keeps its pretrained mass-bias network frozen as a lookup table. Labels: MSFragger 4.4.1 DDA+ search with Percolator, positive if $\geq 2$ distinct accepted peptide backbones at `selection_q` $\leq 0.01$, negative if exactly one. Uncertainty: paired moving-block bootstrap (100 retention-time-consecutive scans, 1,000 resamples, identical resampled scans for both models and all seeds).

**Chimera decomposition.** Detection head and a per-peak two-slot assignment head; 60,000 exposures (40k synthetic then 20k half-real), encoder learning rate $5\times10^{-5}$, heads $10^{-3}$, batch 32, seeds 17/29/43, checkpoint and decoding selected on validation, test cohort frozen before inference. The random-backbone reference is a 200M architecture with a random encoder that keeps the pretrained mass-bias table.

**Intensity prediction.** Each peptide is converted to its theoretical b/y-ion ladder with a peptide-agnostic prior intensity; a head scores each ion from its encoder representation plus embeddings of the ion slot, precursor charge and normalised collision energy; softmax over ions, reported in Prosit's 174-slot layout normalised to the base peak; Prosit's masked spectral-distance loss, three epochs.

**Single-cell denoiser.** Frozen Iona-400M, per-peak head trained with synthetic negatives (injected uniform and mass-randomised peaks drawn from the real peak $m/z$ distribution). Policies: IONA-soft, intensity attenuation by $(1-p_{\text{noise}})^{\lambda}$ with $\lambda = 0.1$; IONA-hard, removal at $p_{\text{noise}} > 0.95$; selected on two separate cells and frozen before the confirmation cells.

## F METRICS

**MAP@R** (Musgrave et al., 2020): for a query with $R$ relevant items in the gallery, average precision is computed over the top $R$ retrieved items and averaged over queries; relevant means the same modified peptide and charge. **Hit@1**: fraction of queries whose top-ranked candidate is the true peptide. **Denoising AUROC/AUPRC**: pooled over all peaks of the test split with noise as the positive class (AUPRC of a random classifier is about 0.53). **Chimericity AUROC**: computed within each acquisition and averaged with equal weight. **Decomposition F1**: macro F1 of the two-slot assignment over the better slot permutation; on HCT116, over exclusively supported fragments. **Spectral angle**: $1 - \frac{2}{\pi}\arccos(\hat{\mathbf{y}} \cdot \mathbf{y})$ on L2-normalised intensity vectors.

## G PRETRAINING AND SCALING ROBUSTNESS

The proxy counts padded peak positions, so it overstates non-padding exposure by a size-independent factor for the 25M–100M runs; the larger ratio at 400M is unexplained and is under investigation (resumed-run accounting, token definition and what the logger counts are candidates). The proxy also omits the $O(N^2K)$ cost of the pairwise Fourier bias, which Pairwise Attention

measured at about 31% of wall clock (Lapin et al., 2025). The fit of Eq. 4 has $A = 0.05294$, $B = 0.02878$, $\alpha = 0.1375$, $\beta = 0.5709$ and log-space $R^2 = 0.9886$ over 40 checkpoints; it has no irreducible-loss term, uses correlated checkpoints of one run per size under one cosine schedule, and a checkpoint under a long schedule is not equivalent to a separately trained short-budget run.

## H ZERO-SHOT BEHAVIOUR OF THE FROZEN ENCODERS

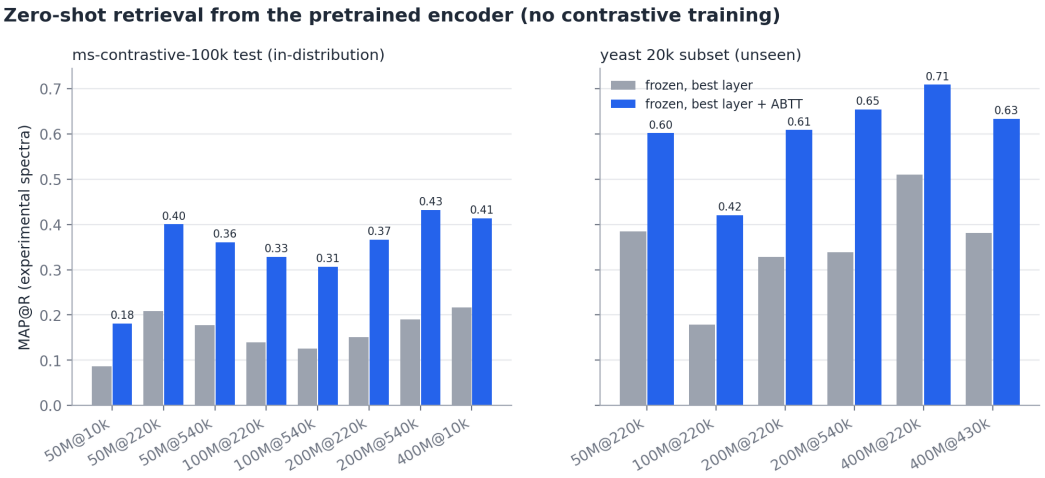



Figure 6: MAP@R of frozen encoders with no task training, best layer raw versus best layer after all-but-the-top (ABTT) post-processing, on the in-distribution test set and on unseen yeast.

Table 7: Zero-shot retrieval MAP@R of frozen encoders (size@pretraining steps). ABTT statistics (mean and top-$D$ principal components) are fitted on training spectra only: the `ms-contrastive-100k` train split for the in-distribution panel, and 25,000 spectra from the eight other species for yeast. The layer and $D$ are the best on each benchmark's test set, so ABTT values are upper bounds.

| Benchmark | Encoder | Raw, final layer | Raw, best layer | ABTT, best layer | Layer, $D$ |
|---|---|---|---|---|---|
| in-distribution | 50M@220k | 0.112 | 0.208 | 0.401 | block 8, 64 |
| in-distribution | 100M@220k | 0.093 | 0.140 | 0.328 | block 10, 64 |
| in-distribution | 200M@540k | 0.080 | 0.191 | **0.432** | block 12, 128 |
| in-distribution | 400M@10k | 0.168 | 0.216 | 0.414 | block 16, 128 |
| yeast 20k | 50M@220k | 0.316 | 0.385 | 0.602 | block 7, 128 |
| yeast 20k | 200M@540k | 0.221 | 0.339 | 0.654 | block 12, 128 |
| yeast 20k | 400M@220k | 0.355 | 0.510 | **0.709** | block 14, 128 |

The raw pooled embedding is dominated by a few directions that encode spectrum-level nuisance factors; projecting them out roughly doubles MAP@R for every encoder. On unseen yeast the frozen 400M encoder with ABTT (0.709) beats every contrastively fine-tuned model (at most 0.656 on any seed), though not GLEAMS (0.770).

## I LEARNED BIAS CONTROLS

The enrichment analysis of Section 4.2 uses the final checkpoint of Iona-50M (`data/delta_bias.csv`): residues 2.72× (19 targets, 95% CI of the known mean 1.37–1.62 versus random 0.49–0.62), neutral losses 1.95× (7 targets), isotopes 2.93× (5 targets), random-vs-random 1.00×.

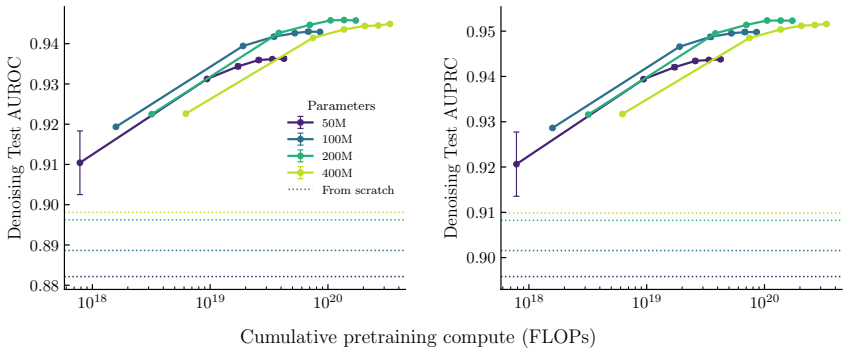



Figure 7: Test AUPRC and AUROC of the fine-tuned denoiser against cumulative pretraining compute of its encoder, one line per size.

Table 8: Peak denoising test AUROC (mean over 3 seeds; 5 seeds at 50M/540k; 50M scratch is one run scored on 8,584 rather than 8,567 spectra) against the pretraining checkpoint of the encoder, and AUPRC at the final checkpoint. Seed SD is about 0.0005 except 50M/10k (0.008, real seed variance).

| Size | scratch | 10k | 120k | 220k | 330k | 430k | 540k | AUPRC (540k) |
|---|---|---|---|---|---|---|---|---|
| 50M | 0.882 | 0.910 | 0.931 | 0.934 | 0.936 | 0.936 | 0.936 | 0.944 |
| 100M | 0.889 | 0.919 | 0.939 | 0.942 | 0.943 | 0.943 | 0.943 | 0.950 |
| 200M | 0.896 | 0.923 | 0.943 | 0.945 | 0.946 | 0.946 | **0.946** | **0.952** |
| 400M | 0.898 | 0.923 | 0.941 | 0.944 | 0.944 | 0.945 | 0.945 | 0.952 |

## J Denoising details

**Caveats.** All values are logged at the end of fine-tuning; models reloaded from saved weights currently re-evaluate about 0.014 AUROC lower (50M: 0.9227 vs 0.9366), which is under investigation; every point in Table 8 uses the same logged evaluation and is comparable within the table. The recipe was selected on test AUROC from a 216-configuration grid at 50M (`denoise/HP`).

## K Retrieval details

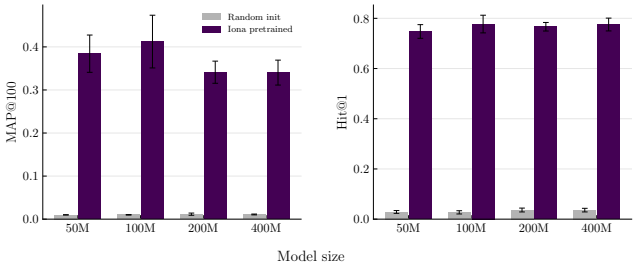



Figure 8: Identical contrastive training from the pretrained encoder and from random initialisation, six seeds per size, each size at its original pretraining checkpoint (50M 133k, 100M 138k, 200M 193k, 400M 181k). This uses an earlier short recipe (temperature 0.07) scored on the replicate-corpus evaluation set with MAP@100 and Hit@1, so the values are not comparable with the MAP@R tables. From random initialisation MAP@100 stays at about 0.01 and Hit@1 at about 0.03 at every size, against 0.34–0.41 and 0.75–0.78 from the pretrained encoder.

**Fine-tuned models and validation selection.** The fine-tuned 400M model (`iona-contrastive-400m`) scores 0.868 in distribution and 0.596 on yeast 20k for its

Table 9: Denoising loss-scaling fit, Eq. 5: weighted least squares over 29 points (4 sizes $\times$ 7–8 checkpoints, 3–13 seeds each; weights from the standard error over seeds floored at $5 \times 10^{-4}$).

| Quantity | Value |
|---|---|
| $\beta$ (shared) | $0.46 \pm 0.04$ |
| $B$ | 0.021 |
| $E_N$, 50M / 100M / 200M / 400M | 0.3125 / 0.2955 / 0.2892 / 0.2906 (each $\pm$0.002) |
| $R^2$ | 0.997 |
| Alternative with $A\,N^{-\alpha}$ term | $\chi^2$/dof 5.5 vs 3.2; needs $\alpha \approx 2$ |
| Single power law in $N \times S$ | $R^2$ 0.92, systematic residuals by size |
| Held out: fit without 540k, predict 540k | within 0.002 at every size |

Table 10: Frozen-probe denoising against other encoders (AUPRC / AUROC), same head.

| | Frozen encoder | | Fine-tuned (pretrained) | | Fine-tuned (from scratch) | |
|---|---|---|---|---|---|---|
| Model | AUPRC | AUROC | AUPRC | AUROC | AUPRC | AUROC |
| Casanovo 5.2.0 Orbitrap | 0.874 | 0.795 | 0.877 | 0.803 | 0.867 | 0.778 |
| InstaNovo-FM 0.1.0 | 0.890 | 0.840 | 0.897 | 0.856 | 0.840 | 0.683 |
| Iona-50M | 0.945 | 0.907 | – | – | – | – |
| Iona-100M | 0.952 | 0.918 | – | – | – | – |

validation-selected seed; another seed scores 0.656 on yeast and was not selected because that would use the test set. The 3-seed mean on yeast 20k is 0.600.

## L PEPTIDE STUDENT AND CROSS-MODAL IDENTIFICATION

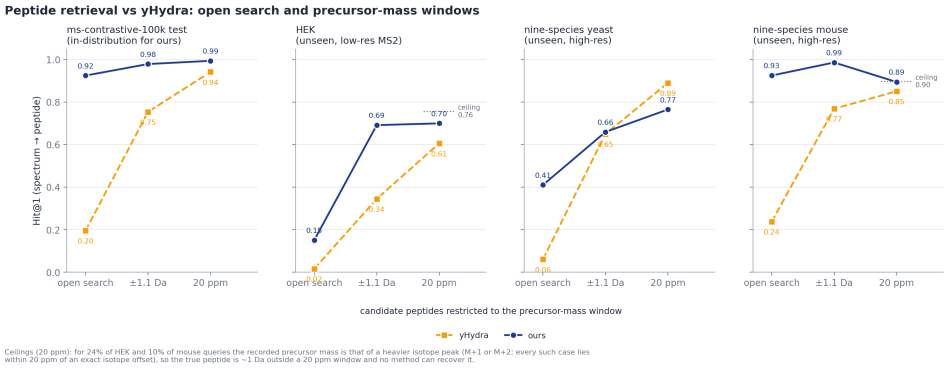



Figure 9: Hit@1 of Iona peptide embeddings against yHydra under three candidate windows on (a) ms-contrastive-100k test (in distribution for Iona), (b) HEK293 low-resolution, (c) nine-species yeast and (d) nine-species mouse.

## M RERANKING DETAILS

**Controls.** With shuffled target/decoy labels the rescorer finds 0 PSMs at 1% FDR. The embedding cosine alone gives AUROC 0.52–0.53 for target versus decoy (null 0.50). MS2Rescore with the embedding added: 128,909 against 128,658 with the null, a net $+251$; on the 6 HEK runs the gain is clean ($+257$ vs $+6$ for the null) while on the 2 small HCT116 runs the null gained more than the embedding ($+452$ vs $+343$), within MS2Rescore's sensitivity to added features on those runs. The 18-run MS2Rescore comparison is pending.

Table 11: Spectrum retrieval MAP@R and Hit@1 for every method and benchmark, both bin widths for cosine. Iona rows are 3-seed means (yeast 20k stage-1 model: 1 seed).

| Benchmark | Method | MAP@R | Hit@1 |
|---|---|---|---|
| ms-contrastive-100k test | Iona-400M fine-tuned | **0.868** | 0.914 |
| | Iona-400M stage 1 only | 0.713 | 0.809 |
| | GLEAMS | 0.646 | 0.746 |
| | binned cosine, 1 Da | 0.671 | 0.771 |
| | binned cosine, 0.1 Da | 0.729 | 0.819 |
| HEK293 low-resolution | Iona-400M fine-tuned | 0.170 | 0.285 |
| | Iona-400M stage 1 only | 0.017 | 0.044 |
| | GLEAMS | 0.530 | 0.684 |
| | binned cosine, 1 Da | **0.554** | 0.746 |
| | binned cosine, 0.1 Da | 0.247 | 0.463 |
| nine-species yeast (86k) | Iona-400M fine-tuned | 0.518 | 0.709 |
| | Iona-400M stage 1 only | 0.474 | 0.685 |
| | GLEAMS | 0.676 | 0.820 |
| | binned cosine, 1 Da | **0.790** | 0.892 |
| | binned cosine, 0.1 Da | 0.789 | 0.891 |
| yeast 20k subset | Iona-400M fine-tuned | 0.600 | 0.784 |
| | Iona-400M stage 1 only | 0.520 | 0.741 |
| | GLEAMS | 0.770 | 0.886 |
| | binned cosine, 1 Da | 0.916 | 0.967 |
| | binned cosine, 0.1 Da | **0.916** | 0.965 |

Table 12: Contrastive retrieval MAP@R against pretraining checkpoint and size (replicate-corpus recipe, 24 epochs, 3-seed means, `ms-contrastive-100k` test).

| Size | 10k | 120k | 220k | 330k | 430k / 540k |
|---|---|---|---|---|---|
| 50M | 0.447 | 0.627 | 0.657 | 0.655 | 0.664 / 0.665 |
| 100M | 0.510 | 0.646 | 0.665 | 0.666 | 0.672 / 0.671 |
| 200M | – | – | 0.658 | – | – |
| 400M | – | – | **0.703** | – | – |

## N Intensity prediction

Given a peptide sequence, precursor charge and normalised collision energy, the model predicts the relative intensities of all b- and y-ions. On the Prosit 2020 HCD holdout, Iona-200M and 400M fine-tuned for three epochs reach a median normalised spectral angle of 0.909 (mean 0.900 and 0.902), matching the Prosit RNN (0.908 / 0.886) (Gessulat et al., 2019) and approaching Prosit Transformer (0.927 / 0.906) (Ekvall et al., 2022), with a smaller training budget than either (Prosit uses early stopping with a patience of 10 epochs). Since the encoder never sees a residue label, the peptide's identity reaches it only as the set of mass differences between its theoretical fragments.

## O Matched-budget peak selection

The gain is concentrated in short peptides ($+14.5$ to $+18.4\%$ of the original short peptides); medium ($+2.8$ to $+3.9\%$) and long ($+0.9$ to $+1.2\%$) gains are small and mostly not significant. The largest gains are on the sparse, low-peak files. Caveats: for file 01277_E01 at 200M the top-$k$ was matched to a re-denoised revision that differs from the searched file (21.7% of spectra have identical peak counts; $\geq 99.8\%$ at other sizes); two files used msconvert rather than ThermoRawFileParser conversions for the top-$k$ arm (search counts differ by $\leq 0.2\%$).

## P Single-cell pilot

**Mechanism check and diagnostics.** On a quarantined pilot-development K562 run, repeat search, identity re-encode and exact injection removal all reproduce the original 2,023 peptides with iden-

Table 13: Hit@1 by candidate window (open search / $\pm 1.1$ Da / 20 ppm). Queries are the spectra whose peptide yHydra can represent, identical for both methods; yHydra uses L2 distance (better than cosine for it on every dataset). Our candidates are (peptide, charge) pairs, yHydra's are sequences, so our lists are slightly larger. The yeast model was selected on the eight other species; the others are the released 400M student. At 20 ppm, 24.9% of HEK and 11.9% of mouse queries have a recorded precursor mass belonging to a heavier isotope, which caps Hit@1 at 0.76 and 0.90.

| Dataset (queries) | yHydra | Iona student |
|---|---|---|
| ms-contrastive-100k test (22,869; in distribution) | 0.20 / 0.75 / 0.94 | 0.92 / 0.98 / 0.99 |
| HEK293 low-resolution (26,628) | 0.02 / 0.34 / 0.61 | 0.15 / 0.69 / 0.70 |
| Nine-species yeast (75,476) | 0.06 / 0.65 / 0.89 | 0.41 / 0.66 / 0.77 |
| Nine-species mouse (19,710) | 0.24 / 0.77 / 0.85 | 0.93 / 0.99 / 0.89 |

Table 14: A better teacher gives a better student. Students trained identically against three teachers; test split of `ms-contrastive-100k`, 25,848 spectra against all 9,771 candidate peptides (open search), 3 seeds.

| Teacher | Teacher MAP@R | Student Hit@1 | Hit@5 | MRR |
|---|---|---|---|---|
| fine-tuned 50M, step 600 | 0.831 | $0.898 \pm 0.0007$ | 0.932 | 0.914 |
| fine-tuned 400M seed 1, step 600 | 0.859 | $0.918 \pm 0.0003$ | 0.946 | 0.931 |
| fine-tuned 400M seed 0, end of epoch | 0.868 | $0.923 \pm 0.0001$ | 0.949 | 0.935 |

tical PSM sets, and injecting about 50% noise cuts peptides by 36.8%, so the search pipeline is sensitive to the corruption the head is trained to find. The noise floor (median peak intensity 117–122) and fragment mass error ($+4.5$ to $+5.1$ ppm) are identical across the four cells; the cells differ in how much signal rises above the floor. The rescued cells are those in which MSFragger-DIA underperformed Spectronaut most (2nd and 7th percentile of the 100-cell plate), and a plain intensity floor recovers most of the same gain, so the hypothesis (untested) is that peak cleanup compensates for engine underperformance on sparse spectra rather than recovering information the spectra lack. Four HeLa confirmation cells, all with more than 2,400 original peptides, changed by $-1.9\%$ to $+2.8\%$. The pilot's decision rule returned "inconclusive" and an extension with four more K562 cells is running.

## Q CHIMERICITY FULL REPORT

ROC curves for every role and arm are in Figure 4 of the main text.

**Pre-registered 200M analysis.** Frozen + MLP: $0.764 \pm 0.002$, $\Delta$ vs Casanovo $-0.038$ [$-0.042$, $-0.034$]. Adapted (fixed bias, learning rate $5 \times 10^{-5}$): $0.841 \pm 0.013$, $\Delta$ $+0.021$ [$+0.019$, $+0.023$]. The 400M analysis (R1) was run under its own evaluation freeze written before any 400M adapted test inference, with every other protocol element unchanged.

**Label efficiency (400M vs Casanovo, MLP).** 1k: 0.728 vs 0.747 ($-0.019$ [$-0.023$, $-0.014$]); 5k: 0.780 vs 0.786 ($-0.006$); 10k: 0.786 vs 0.794 ($-0.009$); 60k: 0.811 vs 0.802 ($+0.009$). With the linear probe, Iona-400M overtakes at 10k labels ($+0.003$ [$-0.001$, $+0.007$]). The 200M→400M step adds $+0.066$ at 1k labels.

**T3 (local HCT116 runs).** Frozen: Casanovo better by 0.056 [0.023, 0.095]. Adapted: Iona-400M 0.888 vs 0.877, $\Delta$ $+0.011$ [$-0.000$, $+0.025$]. T3 includes a run with 13 positives that receives equal weight and a run with 0 positives that is excluded.

**Label-rule robustness.** $\Delta$ (frozen MLP) is $+0.009$ under all three label rules: `selection_q` $\leq 0.01$ (primary), $\leq 0.001$ ([0.004, 0.015], three acquisitions because T1b has no positives), and plain Percolator $q \leq 0.01$ ([0.006, 0.013]). The primary rule was amended before the data freeze from $\leq 0.001$ to $\leq 0.01$ because the stricter threshold is unreachable for secondary ranks in single-run Percolator groups.

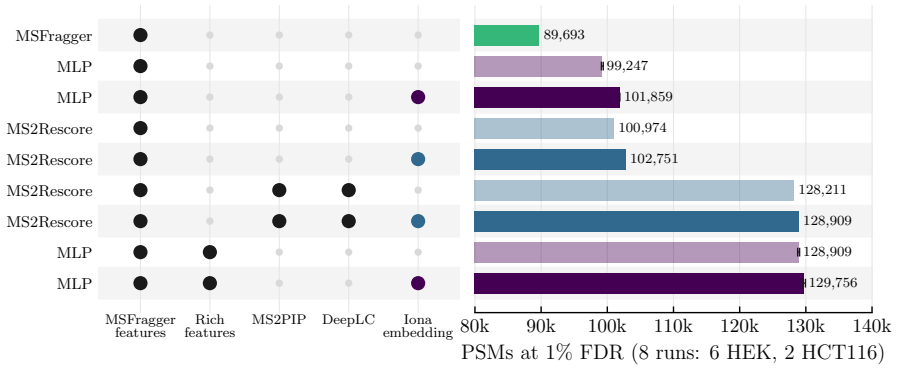



Figure 10: PSMs at 1% FDR on the 8-run set for MSFragger alone, MS2Rescore with and without MS$^2$PIP and DeepLC, and our per-run rescorer, each with and without Iona embedding features; the grid shows which feature groups each method uses.

Table 15: Net gain (embedding minus shuffled-embedding null) in PSMs at 1% FDR, per rescoring seed on the 8-run set and one seed on the 18 unseen HCT116 runs.

| Cohort | Base features | Iona 50M | Iona 400M |
|---|---|---|---|
| 8 runs | MSFragger | +1,495 / +1,215 / +1,186 (+1.3%) | +2,705 / +2,373 / +2,386 (+2.5%) |
| 8 runs | rich | +398 / +738 / +473 (+0.4%) | +564 / +924 / +544 (+0.5%) |
| HCT116, 18 runs | MSFragger | – | +20,763 (+5.6%) |
| HCT116, 18 runs | rich | – | +8,624 (+1.65%) |

**Disclosures.** 400M was chosen as the main encoder after seeing the family-sweep test results, and the same test acquisitions were reused across the initial study, the sweep and R1. Casanovo's selected adaptation learning rate ($2 \times 10^{-4}$) is the upper edge of the grid ($10^{-5}$: 0.783, $5 \times 10^{-5}$: 0.802, $2 \times 10^{-4}$: 0.808 on validation); Iona-400M degrades at $2 \times 10^{-4}$ (validation 0.697). Trainable parameters differ about $20\times$ (395M vs 19.4M). The Casanovo checkpoint is v5.2.0 Orbitrap (software 5.2.1), not the v4.0.0 checkpoint of the Casanovo Foundation paper; the first head layer is $1280 \times 512$ for Iona and $512 \times 512$ for Casanovo. Labels are search-derived; single-peptide support does not prove a clean spectrum. Pretraining overlap with these acquisitions is unknown for both model families, and the family sweep did not audit pretraining data or token budgets across sizes. Intervals are conditional within-run block bootstraps over four acquisitions; no acquisition-cluster interval is possible.

## R CHIMERA DECOMPOSITION FULL REPORT

The 200M rows are the original six-hour frozen runs; 25M, 50M, 100M and 400M were added later against the same frozen test cohort. On an external HEK set of 2,071 pairs evaluated at 200M only, Iona's F1 difference from the random-backbone control is $-0.0072$ [$-0.0173$, $+0.0022$], a transfer failure we keep visible. No Casanovo comparison exists for this task; the per-size runs skipped HEK and the random-backbone control; the random backbone keeps Iona's pretrained mass-bias table and is not fully from scratch; the synthetic mixtures come from MSConsensus, the pretraining source; HCT116 labels are corroborated search-derived pseudo-labels, not pair-level FDR. All runs use a fixed 60k-exposure budget, so these are fixed-budget comparisons, not convergence claims.

Table 16: Method comparison on the 8-run set, PSMs at 1% FDR (our rescorer: 3-seed mean).

| Method | PSMs |
|---|---|
| MSFragger e-value | 89,693 |
| MS2Rescore, MSFragger features | 100,974 |
| MS2Rescore, MSFragger features + Iona embedding | 102,751 |
| MS2Rescore, MSFragger features + MS$^2$PIP + DeepLC | 128,211 |
| MS2Rescore, MSFragger features + MS$^2$PIP + DeepLC + Iona embedding | 128,909 |
| Ours, MSFragger features | 99,247 |
| Ours, MSFragger features + Iona embedding | 101,859 |
| Ours, rich features | 128,909 |
| Ours, rich features + Iona embedding | 129,756 |

Table 17: Fragment-intensity prediction on the Prosit 2020 HCD holdout (1.42M spectra); normalised spectral angle, higher is better.

| Method | Median | Mean |
|---|---|---|
| Prosit Transformer | 0.927 | 0.906 |
| Prosit RNN | 0.908 | 0.886 |
| Iona-200M, fine-tuned | 0.909 | 0.900 |
| Iona-400M, fine-tuned | 0.909 | 0.902 |

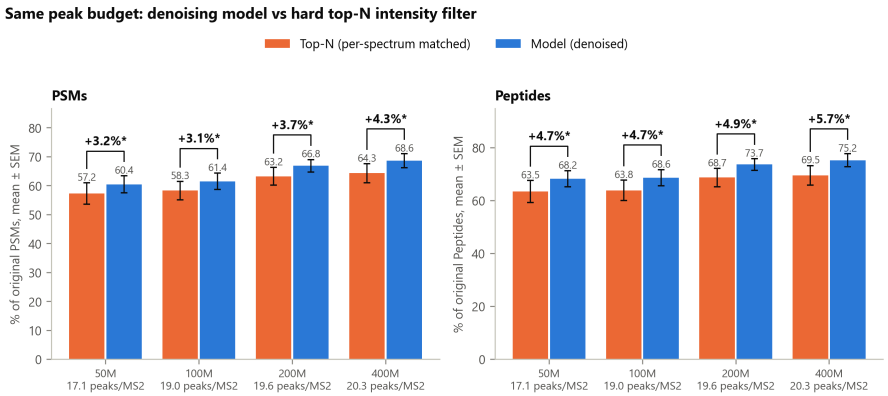



Figure 11: Peptides and PSMs identified after Iona denoising (threshold 0.1) and after per-spectrum matched top-$k$ intensity selection, as a percentage of each file's unfiltered search; mean ± SEM over 7 files; stars are paired $t$-tests across files, BH-adjusted.

Table 18: Model versus matched top-$k$, mean over 7 Q Exactive Plus files (PXD010540), MSFragger 4.1 + MSBooster + Percolator, 1% FDR. SNR $= 10\log_{10}(\sum$ matched b/y intensity$^2 / \sum$ unmatched intensity$^2)$ at ±20 ppm. Stars: BH-adjusted paired $t$-test, * $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$.

| | 50M | 100M | 200M | 400M |
|---|---|---|---|---|
| Peptides, % of original (top-$k$ → model) | 63.5 → 68.2* | 63.8 → 68.6* | 68.7 → 73.7* | 69.5 → 75.2* |
| PSMs, % of original (top-$k$ → model) | 57.2 → 60.4* | 58.3 → 61.4* | 63.2 → 66.8* | 64.3 → 68.6* |
| Short (7–10 aa) peptides kept, model / top-$k$ | 76 / 60%* | 74 / 60%* | 80 / 64%* | 81 / 63%* |
| b/y SNR, model − top-$k$ | +2.7 dB*** | +2.5 dB*** | +2.6 dB*** | +2.6 dB*** |
| Hyperscore, model − top-$k$ | +1.1** | +1.4*** | +1.4*** | +1.7*** |

Table 19: Single K562 cells (PXD070201; Orbitrap Astral, 8 Th DIA, dry injection): unique stripped peptides at 1% peptide-level $q$ on the human database with every MS2 scan included, and change relative to the original search.

| | blo5run12 | blo5run7 | blo5run3 | blo5run38 |
|---|---|---|---|---|
| Original peptides | 718 | 947 | 1,970 | 2,707 |
| IONA-soft ($\lambda = 0.1$) | 984 (**+37.0%**) | 1,518 (**+60.3%**) | 2,009 (+2.0%) | 2,694 ($-0.5\%$) |
| IONA-hard ($\tau = 0.95$) | 980 (+36.5%) | 1,405 (+48.4%) | 2,056 (+4.4%) | 2,686 ($-0.8\%$) |
| Intensity floor 1% | 935 (+30.2%) | 1,282 (+35.4%) | 1,957 ($-0.7\%$) | 2,609 ($-3.6\%$) |
| Gained / lost peptides, IONA-soft | +323 / $-57$ | +624 / $-53$ | +84 / $-45$ | +71 / $-84$ |
| Protein groups, original → IONA-soft | 495 → 596 | 569 → 816 | 975 → 1,052 | 1,246 → 1,323 |
| Spectronaut peptides (lab reference) | 5,108 | 8,239 | 10,530 | 10,825 |
| MSFragger / Spectronaut (plate percentile) | 0.141 (7th) | 0.115 (2nd) | 0.187 (26th) | 0.250 (87th) |

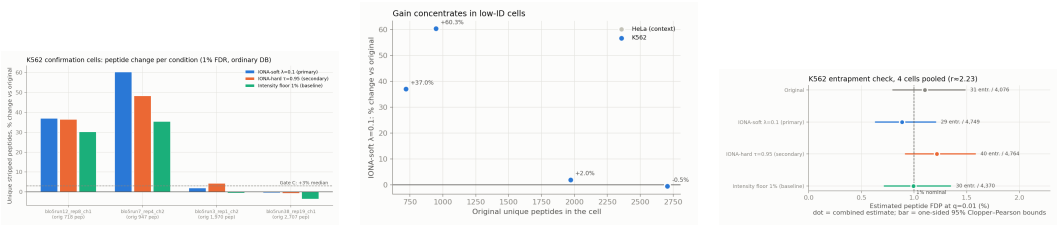



Figure 12: **Left:** per-cell peptide change by condition. **Middle:** IONA-soft gain against original peptide count for the 4 K562 and 4 HeLa confirmation cells; gains appear only below about 1,000 original peptides. **Right:** pooled K562 entrapment FDP at $q = 0.01$ per condition.

Table 20: Pooled K562 entrapment check with a FDRBench peptide-level shuffled entrapment database (v1.1.1; measured $r = 2.23$), same settings for every condition; combined FDP with one-sided 95% Clopper–Pearson bounds scaled by $(1 + 1/r)$.

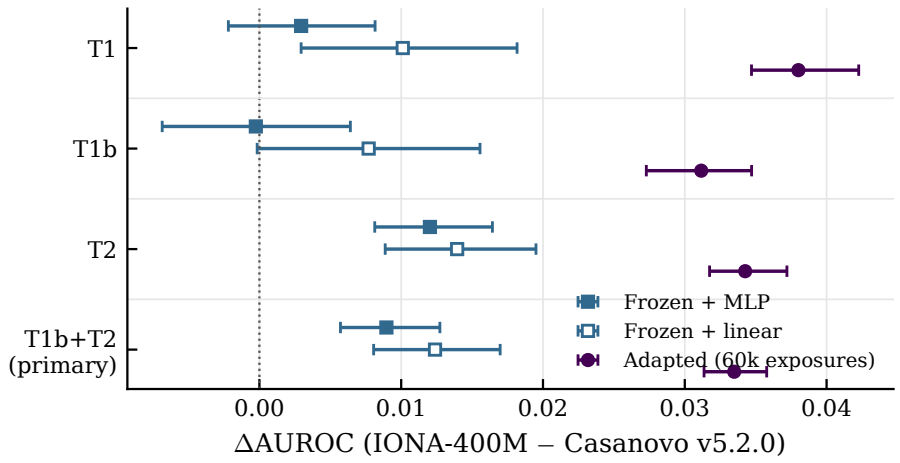
| Condition | Discoveries | Entrapment hits | Combined FDP | 95% bounds |
|---|---|---|---|---|
| Original | 4,076 | 31 | 1.10% | 0.80–1.48% |
| IONA-soft $\lambda = 0.1$ | 4,749 | 29 | 0.88% | 0.63–1.20% |
| IONA-hard $\tau = 0.95$ | 4,764 | 40 | 1.22% | 0.92–1.58% |
| Intensity floor 1% | 4,370 | 30 | 0.99% | 0.72–1.35% |

| **Arm (60k labels)** | **T1** | **T1b** | **T2** | **T1b+T2 (primary)** | **AP (primary)** |
|---|---|---|---|---|---|
| **IONA-400M, frozen + MLP** | 0.811 ± 0.004 | 0.855 ± 0.004 | 0.797 ± 0.003 | 0.811 ± 0.003 | 0.475 |
| Casanovo v5.2.0, frozen + MLP | 0.808 ± 0.001 | 0.855 ± 0.004 | 0.785 ± 0.003 | 0.802 ± 0.003 | 0.470 |
| **IONA-400M, frozen + linear** | 0.793 ± 0.001 | 0.839 ± 0.002 | 0.779 ± 0.001 | 0.794 ± 0.002 | 0.449 |
| Casanovo v5.2.0, frozen + linear | 0.783 ± 0.001 | 0.832 ± 0.002 | 0.765 ± 0.001 | 0.782 ± 0.001 | 0.442 |
| **IONA-400M, adapted** | 0.857 ± 0.002 | 0.908 ± 0.004 | 0.836 ± 0.002 | 0.854 ± 0.002 | 0.527 |
| Casanovo v5.2.0, adapted | 0.819 ± 0.005 | 0.876 ± 0.002 | 0.802 ± 0.006 | 0.820 ± 0.005 | 0.494 |
| Peak statistics (HGB) | 0.704 | 0.719 | 0.678 | 0.688 | 0.338 |
| Binned spectrum (HGB) | 0.695 | 0.664 | 0.659 | 0.660 | 0.313 |
| *Paired ΔAUROC (IONA-400M − Casanovo v5.2.0), 95% conditional interval* | | | | | |
| Δ frozen + MLP | +0.003 [-0.002, +0.008] | -0.000 [-0.007, +0.006] | +0.012 [+0.008, +0.016] | +0.009 [+0.006, +0.013] | |
| Δ frozen + linear | +0.010 [+0.003, +0.018] | +0.008 [-0.000, +0.016] | +0.014 [+0.009, +0.020] | +0.012 [+0.008, +0.017] | |
| Δ adapted | +0.038 [+0.035, +0.042] | +0.031 [+0.027, +0.035] | +0.034 [+0.032, +0.037] | +0.033 [+0.031, +0.036] | |

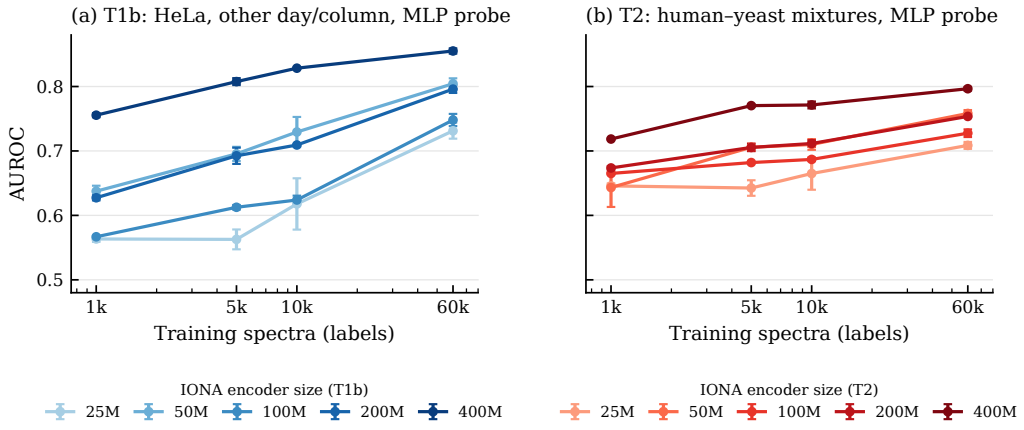
Table 21: Iona-400M against Casanovo v5.2.0, all roles and arms (T3 excluded). HGB = histogram gradient boosting.

Table 22: Frozen family sweep, MLP probe, 60k labels, primary T1b+T2 AUROC (mean ± SD over 3 head seeds), paired adjacent-size Δ and Δ versus Casanovo with 95% conditional intervals, and frozen-extraction throughput.

| Encoder | Primary AUROC | Δ vs previous size | Δ vs Casanovo | Throughput (spectra/s) |
|---|---|---|---|---|
| Iona-25M | 0.714 ± 0.007 | – | $-0.088$ [$-0.093$, $-0.083$] | 938 (Ada) |
| Iona-50M | 0.770 ± 0.006 | $+0.055$ [$+0.051$, $+0.059$] | $-0.033$ [$-0.037$, $-0.029$] | 727 (Ada) |
| Iona-100M | 0.733 ± 0.007 | $-0.037$ [$-0.041$, $-0.033$] | $-0.070$ [$-0.074$, $-0.065$] | 354 (A6000) |
| Iona-200M | 0.764 ± 0.002 | $+0.032$ [$+0.028$, $+0.035$] | $-0.038$ [$-0.042$, $-0.034$] | 204 (A6000) |
| Iona-400M | 0.811 ± 0.003 | $+0.047$ [$+0.043$, $+0.051$] | $+0.009$ [$+0.006$, $+0.013$] | 129 (A6000) |
| Casanovo v5.2.0 | 0.802 ± 0.003 | – | – | – |




60k labels. Paired within-acquisition block bootstrap (100-scan blocks, 1,000 resamples), 95% conditional interval.




Frozen encoders; mean ± seed SD over 3 head seeds; T2 = equal-weight mean of 3 acquisitions. Nested label subsets. T3 excluded.

Figure 13: **Top:** paired ΔAUROC (Iona-400M − Casanovo) with 95% conditional intervals per role and arm. **Bottom:** label efficiency of the frozen family (1k–60k labels, MLP probe) on T1b and T2.

Table 23: Two-component chimera decomposition against size. Synthetic: 5,000 held-out MSConsensus mixtures (2,467 single-source, 2,533 two-component); AUROC over all spectra, macro F1 over the pairs. HCT116: F1 on exclusive supported fragments of 1,376 search-corroborated pairs. Fine-tuned (FT): mean ± SD over seeds 17/29/43; frozen: one fit. Synthetic Δ intervals resample 1,663 parent-connected groups; HCT116 intervals resample only 3 acquisitions (295, 1,080 and 1 pairs) and are not credible population intervals.

| Encoder | FT syn. AUROC | FT syn. F1 | FT HCT116 F1 | Frozen syn. F1 | Frozen HCT116 F1 | Δ syn. F1 vs prev. | FT − frozen, syn. | FT − frozen, HCT116 |
|---|---|---|---|---|---|---|---|---|
| Iona-25M | 0.994 ± 0.001 | 0.851 ± 0.008 | 0.657 ± 0.007 | 0.668 | 0.570 | – | +0.183 | +0.087 |
| Iona-50M | 0.995 ± 0.001 | 0.879 ± 0.004 | 0.692 ± 0.012 | 0.695 | 0.590 | +0.028 [+0.025, +0.030] | +0.183 | +0.102 |
| Iona-100M | 0.997 ± 0.000 | 0.896 ± 0.002 | 0.701 ± 0.027 | 0.690 | 0.593 | +0.017 [+0.015, +0.019] | +0.206 | +0.108 |
| Iona-200M | 0.997 ± 0.001 | 0.913 ± 0.002 | 0.757 ± 0.003 | 0.714 | 0.603 | +0.017 [+0.015, +0.019] | +0.199 | +0.155 |
| Iona-400M | 0.998 ± 0.000 | 0.913 ± 0.006 | 0.767 ± 0.004 | 0.764 | 0.600 | +0.000 [−0.002, +0.002] | +0.148 | +0.168 |
| Random backbone (200M arch.) | 0.951 ± 0.002 | 0.716 ± 0.029 | 0.643 ± 0.010 | – | – | – | – | – |
| Peak statistics / intensity split | 0.913 | 0.635 | 0.493 | – | – | – | – | – |