SPECTRUM-GUIDED LABEL REFINEMENT FOR ROBUST *De Novo* PEPTIDE SEQUENCING

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural *de novo* peptide sequencing models learn from database-search labels that contain both false matches and systematic errors. We introduce Physics-Consistent Self-Correction (PCSC), which uses observed fragment peaks to refine these labels. A model predicts a peptide for each spectrum. The prediction replaces the current label when it receives a higher score from a fixed spectrum–peptide matcher. Models are then retrained on the updated labels. With 60% global corruption, five label updates recover 63.3% of corrupted labels, while four retraining rounds raise peptide recall on an independent nine-species benchmark from 0.3142 to 0.3970. On a separate corpus of over 35 million PSMs, one update improves recall in seven of nine species. The procedure also improves an autoregressive model, with gains of 6.88 percentage points under greedy decoding and 0.52 under mass-filtered beam search. These results show that fragment-based label refinement can improve sequencing accuracy across different models and training corpora.

1 INTRODUCTION

Tandem mass spectrometry (MS/MS) is central to large-scale proteomics (Aebersold & Mann, 2016; 2003; Miller & Smith, 2023). Database search identifies peptides by matching observed spectra against candidates from a reference proteome (Perkins et al., 1999; Eng et al., 1994; Cox & Mann, 2008; Kong et al., 2017). Its coverage depends on the database and search configuration, which can exclude unexpected modifications, variants, and novel peptides (Walsh et al., 2005; Chick et al., 2015). *De novo* sequencing predicts peptides directly from spectra (Ma et al., 2003; Frank & Pevzner, 2005). Neural methods have substantially improved its accuracy (Tran et al., 2017; Yilmaz et al., 2022; Eloff et al., 2025; Zhang et al., 2025; Chen et al., 2025), but their training labels still come from database-search peptide–spectrum matches (PSMs). Errors in these matches therefore enter the training data (Elias & Gygi, 2007; Savitski et al., 2015; Wen et al., 2025).

Deep networks often learn simple structure before memorizing noise (Arpit et al., 2017; Zhang et al., 2016). Peptide sequencing offers a way to evaluate model predictions independently of their training labels: each proposed sequence determines theoretical fragment masses that can be compared with observed peaks. Can a model trained on corrupted labels still predict the original sequences? If so, can fragment matching identify predictions that are useful for refining the training set?

We first study learning under synthetic label corruption. Starting from PSMs on which MassIVE-KB labels and OmniNovo predictions agree, we introduce local sequence changes or global perturbations. The model recovers the original sequence more often than it reproduces the injected label, although recovery and external benchmark accuracy decline as corruption becomes more severe.

This observation motivates Physics-Consistent Self-Correction (PCSC; Figure 1). A model predicts a candidate peptide for each spectrum. A fixed Hyperscore-style fragment matcher (Craig & Beavis, 2004) compares that candidate with the current label, and the higher-scoring sequence becomes the label for the next round of training. Under 60% global corruption, five updates recover 63.3% of corrupted labels, and four retraining rounds increase external peptide recall from 0.3142 to 0.3970.

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.

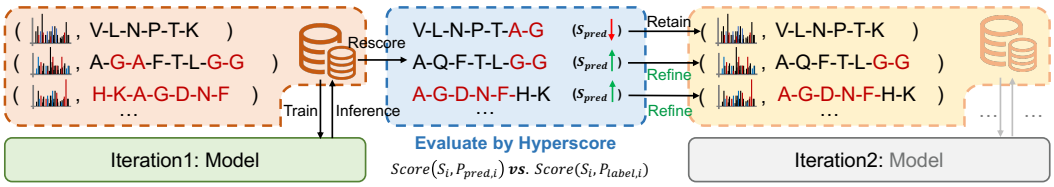


Figure 1: **Physics-Consistent Self-Correction.** A model trained on the current labels predicts a peptide for every spectrum. A fixed fragment matcher compares each prediction with its label. Higher-scoring predictions replace the labels, and a new model is trained from scratch on the updated dataset.

2 Related Work

De novo sequencing. Classical approaches use graph search or probabilistic models (Ma et al., 2003; Frank & Pevzner, 2005); neural systems have progressed from recurrent models (Tran et al., 2017; 2019) to Transformers and diffusion or parallel decoding (Yilmaz et al., 2022; Eloff et al., 2025; Zhang et al., 2025; Chen et al., 2025). These approaches improve sequence prediction through architecture and decoding. PCSC addresses the training labels: it compares model predictions with existing labels using fragment evidence, then retrains the sequencer on the accepted sequences.

Learning with noisy labels. Robust losses reduce the influence of mislabeled examples (Ghosh et al., 2017; Zhang & Sabuncu, 2018; Ma et al., 2020; Wang et al., 2019); sample-selection methods use small-loss examples or peer disagreement (Han et al., 2018; Yu et al., 2019; Li et al., 2020); and label-correction methods update targets from model outputs (Lu & He, 2022). Much of this work concerns classification with a fixed set of classes. Peptide sequencing produces variable-length sequences, and the spectrum provides evidence for evaluating each proposed sequence. PCSC uses this evidence to select labels.

Self-training and scientific constraints. Self-training learns from model-generated targets (Lee, 2013; Amini et al., 2022). Its success depends on selecting useful predictions for subsequent training. Physics-informed learning incorporates governing equations into the training objective (Raissi et al., 2019; Karniadakis et al., 2021). Here, established spectrum–peptide matching is applied after prediction, leaving the sequencer's training objective unchanged.

3 Controlled Study of Noisy Supervision

3.1 Data, Model, and Corruption

High-confidence seed. MassIVE-KB v2 contains approximately 3.6 million PSMs (Wang et al., 2018). We retain 2,487,323 PSMs for which an OmniNovo prediction (Chen et al., 2025) matches the database-search label. These agreed labels serve as the reference sequences for the corruption study.

Architecture. The base learner is a 31.9M-parameter non-autoregressive Transformer (NAT) with an encoder–decoder architecture related to π-PrimeNovo (Zhang et al., 2025): nine layers, model width 384, 12 attention heads, RMSNorm, SiLU, and connectionist temporal classification (CTC) loss. Peptides are decoded greedily. The NAT robust-learning baselines use the same backbone and preprocessing; Appendix A gives the shared training settings.

Synthetic corruption. We study local and global corruption at noise fractions from 10% to 60%. Local corruption swaps adjacent residues or makes a single near-isobaric replacement; global corruption shuffles the full residue sequence or makes multiple near-isobaric replacements (Figure 2). Swaps and shuffles alter residue order, whereas replacements can also change sequence length. These perturbations let us compare recovery across different types and amounts of label error.

In the 60% global-noise dataset, 266,983 reference peptides occur with both clean and corrupted PSMs. The corrupted labels also include 133,266 cases in which an N-terminal modification present

in the reference is omitted. Appendix A gives the dataset counts, and Appendix D describes the precursor-error distribution.

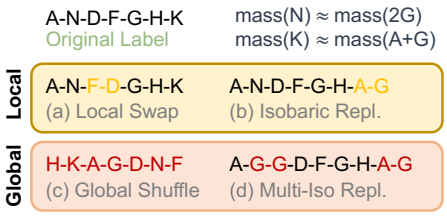
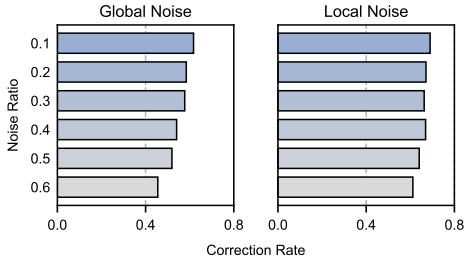
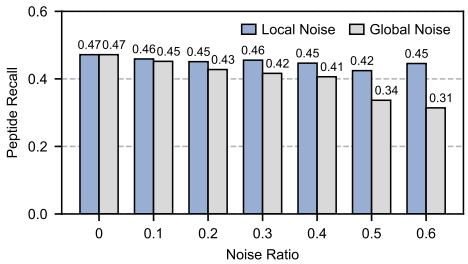


Figure 2: **Synthetic sequence perturbations.** Examples of adjacent-residue swapping, a single near-isobaric replacement, full-sequence shuffling, and multiple replacements. The replacements shown can substitute one residue with two residues of similar combined mass.



(a) Exact correction on corrupted training PSMs.



(b) Peptide recall on the independent nine-species set.

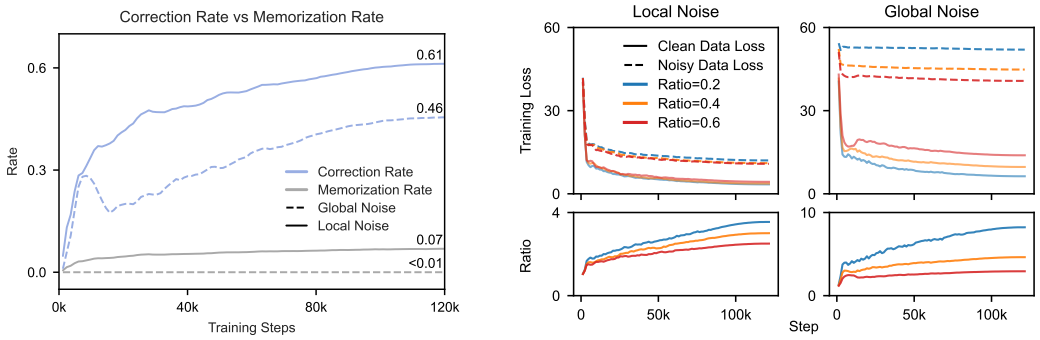
Figure 3: **Learning under controlled corruption.** Exact correction remains substantial as noise increases, while global corruption causes the main drop in transfer recall.

3.2 Correction Precedes Memorization

The model recovers many original sequences even at 60% corruption (Figure 3). Local corruption has little effect on nine-species recall over the studied range, while 60% global corruption lowers it to about 0.31. During training, exact recovery rises early and remains more frequent than reproduction of the injected labels. Clean and globally corrupted examples also develop different loss trajectories (Figure 4).

The loss curves show how this difference develops. Clean-label loss continues to fall, while globally corrupted targets retain high loss; the noisy-to-clean loss ratio consequently grows during training. Under local corruption, both losses decrease and their separation is smaller. The network can thus recover reference sequences even while assigning high loss to the corrupted targets used for training.

This observation makes model predictions useful candidates for label refinement. A prediction that differs from its training label can recover the reference sequence, but disagreement alone does not identify such cases. PCSC compares the fragment evidence for the two sequences before selecting the



(a) Correction and injected-label memorization.

(b) Clean/noisy loss separation and loss ratio.

Figure 4: **Learning dynamics under controlled corruption.** With 60% noise, reference-sequence recovery exceeds reproduction of injected labels under both local and global corruption (a). Clean and corrupted samples also separate in loss space, with a larger noisy-to-clean ratio under global perturbations (b).

next training label. We then evaluate the resulting labels by exact recovery and the retrained models by recall on independent spectra.

4 Physics-Consistent Self-Correction

Let $\mathcal{D}_t = \{(x_i, y_i^{(t)})\}_{i=1}^N$ contain spectrum x_i and current peptide label $y_i^{(t)}$. Training from scratch yields model f_t, whose decoded candidate is $\hat{y}_i^{(t)} = f_t(x_i)$. We compute a fixed fragmentation score $s_{\text{phys}}(x, y)$ using matched b/y ions and their intensities. The next label is

$$y_i^{(t+1)} = \begin{cases} \hat{y}_i^{(t)}, & s_{\text{phys}}(x_i, \hat{y}_i^{(t)}) > s_{\text{phys}}(x_i, y_i^{(t)}), \\ y_i^{(t)}, & \text{otherwise.} \end{cases} \tag{1}$$

The score of each accepted label is non-decreasing: $s_{\text{phys}}(x_i, y_i^{(t+1)}) \geq s_{\text{phys}}(x_i, y_i^{(t)})$. Ties retain the current label. We use a fixed Hyperscore that combines matched fragment counts and intensities, and initialize a new model at every round. Appendix B defines the score. At each update, selection is restricted to the current label and one decoded prediction per spectrum. Retraining changes the candidates available at the next round while the scoring rule remains fixed.

5 Experiments

Evaluation. We evaluate exact sequence recovery with I/L equivalence and measure generalization by peptide recall on the independent nine-species benchmark. Amino-acid precision and recall use mass-aware alignment. Physical metrics include the fraction of spectral intensity explained by matched b/y ions (TIC), matched peak counts, and mean signed precursor-mass error (PPM). The pre-corruption sequences provide the reference for recovery measurements. Metric definitions are in Appendix B.

5.1 Iterative restoration at 60% global corruption

Most recovery occurs in the first update, followed by further gains over successive rounds. The first update recovers 777,542 of the 1,493,219 corrupted labels (52.1%); after five updates, recovery reaches 944,697 (63.3%). Between these two points, labels still equal to the injected sequence fall from 145,519 to 35,137, while changed but incorrect labels fall from 570,158 to 513,385 (Figure 5).

The four evaluated retraining rounds raise peptide recall from 0.3142 to 0.3509, 0.3728, 0.3875, and 0.3970. Gains are largest early and decrease across rounds. Across all training PSMs, mean

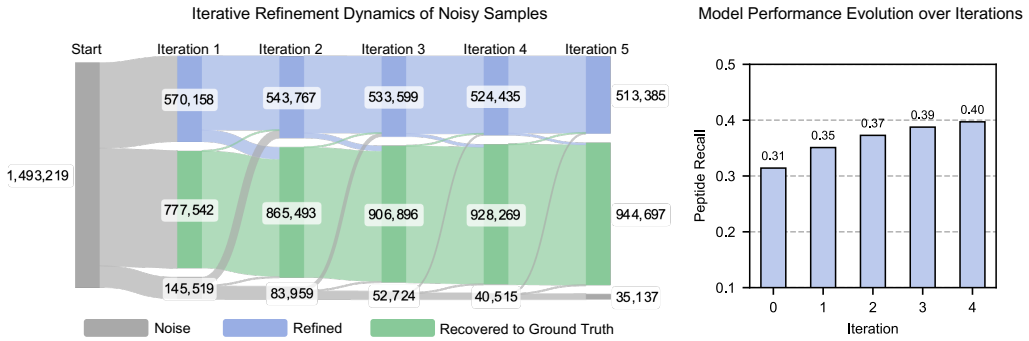
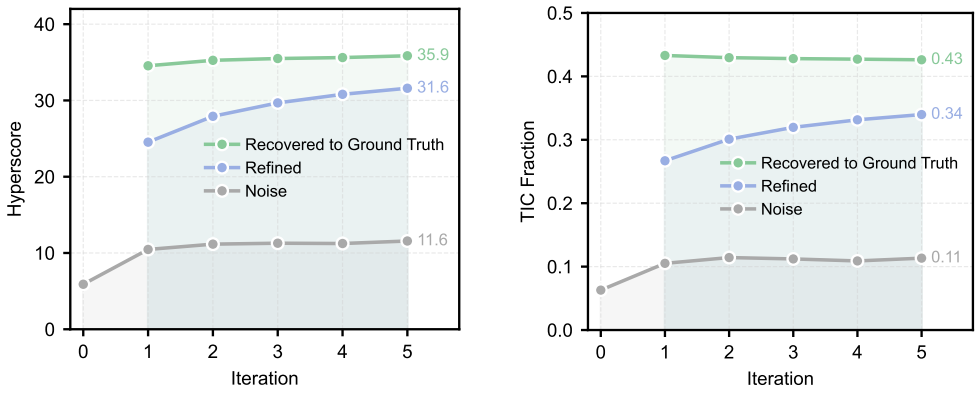


Figure 5: **Iterative restoration.** Left: status of the 1,493,219 initially corrupted PSMs. "Recovered" means exact reference-clean sequence recovery; "Refined" means a changed label that is still incorrect; "Noise" retains the original corrupted label. Right: peptide recall of successive models on the independent nine-species benchmark.

Hyperscore rises from 17.99 to 34.81, TIC from 0.2002 to 0.3965, and matched peaks from 8.43 to 16.21 after five label updates. Appendix D gives the complete physical metrics.

Refinement also changes some initially correct labels. Among the 994,094 PSMs not selected for corruption, 15,791 have incorrect labels after the first update and 42,869 after the fifth (1.59% and 4.31%). Accounting for both recovery and these errors, the fraction of reference-matching labels across the full training set rises from 39.97% to 70.59% after one update and 76.22% after five. The later rounds add 140,077 correct labels overall. Thus, the smaller gains after the first update still yield a net improvement in supervision, even as some initially correct labels are lost.

Figure 6 separates PSMs by their label status after each update. At round 5, recovered labels have mean Hyperscore 35.9 and TIC 0.43, compared with 11.6 and 0.11 for unchanged noisy labels. For labels that have changed but remain incorrect, the group means rise from Hyperscore 24.5 to 31.6 and TIC 0.27 to 0.34 between rounds 1 and 5. After five updates, changed labels account for 93.6% of the errors remaining in the initially corrupted subset. Residual error is therefore dominated by replacements that improve fragment agreement without recovering the complete reference sequence.



(a) Hyperscore.

(b) TIC fraction.

Figure 6: Physical metrics for initially noisy PSMs, grouped by label status after each update. Recovered labels have the highest mean scores; unchanged noisy labels have the lowest.

5.2 Robust-learning and acceptance-rule baselines

GCE gives the highest recall among the single-training comparisons, at 0.3790 (Table 1). Among methods with the same four-update schedule, PCSC improves recall at every round and reaches

Table 1: Peptide recall under 60% global corruption. Top: single-training comparisons using the NAT backbone. Bottom: models retrained after each label update, with the best recall in each round in bold.

Single-training method				Training point		Recall
Noisy NAT training				100 epochs		0.3142
Co-teaching (Han et al., 2018)				100 epochs		0.3616
GCE (Zhang & Sabuncu, 2018)				100 epochs		0.3790
SCE-based NAT adaptation (Wang et al., 2019)				80 epochs		0.2728
SELC-based NAT adaptation (Lu & He, 2022)				10 / 20 epochs		0.2292 / 0.1598
Method	Initial	Round 1	Round 2	Round 3	Round 4	Best
Confidence-direct	0.3142	0.3216	0.3253	0.3073	0.3072	0.3253
Confidence-threshold	0.3142	0.1738	0.3030	0.3261	0.3510	0.3510
PCSC (Hyperscore)	0.3142	**0.3509**	**0.3728**	**0.3875**	**0.3970**	**0.3970**

0.3970. Direct confidence replacement peaks at round 2 and then declines, while threshold-based replacement recovers gradually from an initial drop. The physical-score rule produces higher recall than either confidence rule at every update.

5.3 Large-scale real labels and PTMs

One update on MassIVE-KB v1, containing over 35 million PSMs, changes approximately 0.17 million labels ($< 0.5\%$). Retraining improves peptide recall in seven of nine species across ten paired seeds (Table 2). Recall decreases by 0.29 percentage points for *B. subtilis* and 1.12 points for *S. lycopersicum*, while amino-acid precision and recall improve for both (Appendix E).

Table 2: Peptide recall (%) after one PCSC update on 35M real-world PSMs, mean over ten paired seeds. Full amino-acid metrics, standard deviations, and exact p-values are in Appendix E.

	A. mel.	*B. sub.*	*C. end.*	*H. sap.*	*M. maz.*	*M. mus.*	*S. cer.*	*S. lyc.*	*V. mun.*
Original	54.30	63.50	45.25	58.91	57.52	56.23	61.62	63.48	61.63
Refined	54.87	63.21	45.55	59.51	57.73	56.34	62.03	62.36	62.05
Δ	+0.57	-0.29	+0.30	+0.60	+0.21	+0.11	+0.41	-1.12	+0.42

We also assess fragment agreement after label refinement across nine PTM categories using 3,001,570 modified PSMs from 19 PRIDE projects and one iProX project (Appendix F). PCSC changes 3.58% of these labels. For changed PSMs, mean Hyperscore rises from 8.86 to 12.12 and the number of matched peaks from 4.16 to 5.71. Hyperscore increases in all nine categories, with relative gains from 15.4% to 111.5% (Table 3).

The replacement rate varies from 2.16% for acetylation to 34.57% for trimethylation, so refinement affects very different fractions of the modification-specific data. The Hyperscore gains describe the selected replacements: the selection rule itself requires this score to increase. Together with the matched-peak counts, these results establish improved fragment agreement for the revised assignments; exact sequence and modification-site accuracy are not measured in this analysis.

5.4 Attention and peptide length

Cross-attention alignment. For 200 randomly sampled nine-species spectra, we measure cross-attention on peaks matching the reference peptide's theoretical b/y ions, aggregating over heads within each decoder layer. Before refinement, every layer exceeds the random peak-occupancy baseline of 17.7%; layer 6 assigns 38.8% of its attention to matched peaks. After four updates, enrichment over this baseline increases from 1.15× to 1.30× at layer 0 and from 1.40× to 1.67× at layer 2 (Figure 7). Alignment is lower after refinement at layers 3, 6, 7, and 8, placing the clearest improvement in the early layers.

Table 3: PTM-stratified changes on 3,001,570 PSMs from 19 PRIDE projects and one iProX project. Categories are not necessarily mutually exclusive; Hyperscore is computed on changed PSMs before and after adjudication.

Modification	PSMs	Changed (%)	Hyperscore	Relative gain
Phosphorylation	2,160,374	3.44	6.53 → 8.98	+37.4%
Acetylation	555,662	2.16	12.23 → 16.71	+36.6%
Oxidation	237,783	4.46	9.42 → 13.96	+48.2%
Carbamidomethylation	233,308	4.22	8.48 → 12.14	+43.2%
Ubiquitination	197,059	2.93	25.09 → 28.96	+15.4%
Methylation	120,164	14.58	10.80 → 16.52	+52.9%
Dimethylation	57,127	15.47	8.63 → 15.03	+74.1%
Trimethylation	6,722	34.57	6.63 → 14.02	+111.5%
Deamidation	4,014	11.71	11.02 → 15.00	+36.1%
All PSMs	3,001,570	3.58	8.86 → 12.12	+36.9%

In both models, attention is more concentrated at layer 5 than at layer 0. In the model trained on noisy labels, mean entropy is 4.16 nats at layer 0 and 2.43 at layer 5; the corresponding values after refinement are 3.95 and 2.59. The individual heatmaps in Appendix H show how concentrated attention can coincide with either a complete or a partial sequence prediction. For example, GVPVEGSR recovers only the C-terminal segment of GAEAANVTGPDGVPVEGSR, even though its attention concentrates on matched fragment peaks (Figure 10).

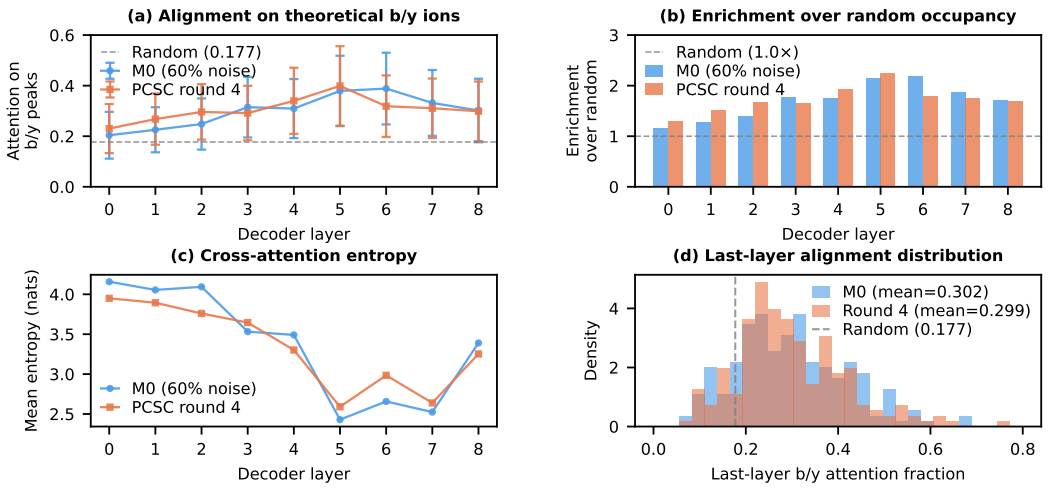


Figure 7: Cross-attention diagnostics on 200 randomly selected nine-species spectra. The four panels show attention on theoretical b/y peaks, enrichment over random occupancy, cross-attention entropy, and the per-sample last-layer alignment distribution.

Candidate-length bias. Because Hyperscore depends on matched-ion counts, we examine whether refinement shifts labels toward longer sequences. Among PSMs selected for corruption, mean label length decreases from 14.77 to 14.35 residues after five updates, close to the reference mean of 14.39; the median changes from 14 to 13 residues, matching the reference median. The Kolmogorov–Smirnov distance to the reference length distribution falls from 0.033 to 0.002 after the first update.

Fewer labels change in later updates, and their mean length change approaches zero. The first update changes 1,347,700 initially corrupted labels, with a mean length change of -0.42 residues; the fifth changes 81,733, with a mean change of $+0.02$. From round 0 to round 5, both Hyperscore and TIC increase among labels that become shorter, stay the same length, or become longer (Appendix G).

Recovery nevertheless varies strongly with reference length: after five updates, it is 94.9% for peptides of length 5–7 and 18.9% for lengths 25–44.

5.5 Architecture and scorer transfer

We examine two components of the procedure separately. Applying spectral angle to the fixed first-round NAT candidates tests how the acceptance rule changes label selection. Training an autoregressive (AR) model before and after a label update tests the complete refinement procedure with a different sequence decoder.

Table 4: AR retraining after one PCSC update. Both models use the same training budget and fixed decoding parameters. Recall is micro-averaged over all 89,924 nine-species spectra; intervals resample peptide clusters for these fixed models.

Decoding	Noisy training	PCSC round 1	Δ (pp)	95% interval (pp)
Greedy	0.2275	0.2963	+6.88	[6.49, 7.25]
Mass-filtered beam	0.3562	0.3615	+0.52	[0.17, 0.88]

Spectral-angle scoring. On the fixed first-round candidate pool, spectral angle and Hyperscore make the same selection for 97.11% of PSMs. Spectral angle recovers 776,821 corrupted labels (52.02%), compared with 777,542 (52.07%) for Hyperscore. Their reference-match rates also remain close across peptide-length and precursor-charge groups, with maximum differences of 0.20 and 0.07 percentage points, respectively (Appendix C).

The NAT predictions already match the reference for 780,962 corrupted PSMs (52.30%). Hyperscore retains 99.56% of these correct candidates, rejecting 3,420. Because selection chooses between the supplied label and this single prediction, 52.30% is the maximum recovery attainable from this pool. Re-scoring can alter which incorrect predictions replace labels; recovering the remaining reference sequences requires new candidates.

Requiring both scorers to favor replacement reduces wrong replacements from 570,158 to 551,389 and raises replacement precision from 57.69% to 58.48%, with recovery of 52.00%. Relative to Hyperscore, this rule rejects 18,769 wrong replacements and 1,060 correct ones. A spectral-angle margin of 0.05 increases these counts to 71,627 and 4,694, respectively, raising precision to 60.79% and lowering recovery to 51.76%. Stricter acceptance rejects more incorrect replacements, but every rejected PSM retains its supplied label. On this initially corrupted subset, the resulting increase in replacement precision comes with fewer exact recoveries. Appendix C gives the score definition, uncertainty estimates, and stratified results.

Autoregressive model. The AR model uses a 29.3M-parameter encoder–decoder Transformer with a causal decoder and teacher-forced cross-entropy. We train one model on the noisy labels and another from scratch after one Hyperscore update. Both receive 100 epochs, each sampling 2,487,323 PSMs with replacement, and share optimization settings, seed 42, and global batch size 512.

We select mass-filtered beam search on 1,024 development spectra: width 10, a 50-ppm precursor tolerance, and no length normalization. This configuration is fixed for generating training candidates and evaluating both models. The label update recovers 1,099,797 corrupted labels (73.65%) at 79.96% replacement precision, and changes 3,205 initially correct labels to incorrect sequences. Greedy and beam decoding are then evaluated separately on the same 89,924 external spectra.

Retraining raises micro-averaged recall by 6.88 percentage points with greedy decoding and 0.52 points with beam search (Table 4). Paired resampling of reference-peptide clusters within species gives 95% intervals of [6.49, 7.25] and [0.17, 0.88] points. These intervals describe test-sample uncertainty for the two models trained with seed 42. Greedy recall improves in all nine species; beam recall improves in six and declines in three, by 0.12–0.62 points (Appendix C).

Paired prediction outcomes explain why the net gains differ (Table 5). Greedy decoding gains 13,815 correct predictions and loses 7,628, for 6,187 additional matches. With beam search, 10,100 gains are nearly offset by 9,629 losses, leaving 471 additional matches. The smaller beam gain therefore accompanies substantial changes in which spectra are sequenced correctly. Beam search remains

more accurate for both models, but its advantage over greedy decoding narrows from 12.87 to 6.52 percentage points after retraining.

Table 5: Paired AR prediction outcomes on the same 89,924 external spectra before and after one label update. Correctness is exact reference agreement with I/L equivalence. The four outcome columns partition all evaluated spectra.

Decoding	Both correct	Wrong → correct	Correct → wrong	Both wrong	Net gain
Greedy	12,828	13,815	7,628	55,653	+6,187
Mass-filtered beam	22,404	10,100	9,629	47,791	+471

On the same 20,000 initially corrupted training PSMs, retraining increases both reference-sequence recovery and reproduction of injected labels. With greedy decoding, recovery rises from 51.015% to 74.120% and injected-label reproduction from 0.275% to 1.970%. With beam search, the corresponding rates rise from 73.740% to 74.670% and from 0.125% to 2.005%.

6 Discussion and Conclusion

PCSC separates sequence prediction from label selection: a neural model proposes a peptide, and observed fragment peaks determine whether it replaces the current target. Across NAT and AR models, retraining on the selected labels improves external sequencing recall. The controlled study connects this gain to a measurable increase in reference-matching training labels, while the 35M-PSM experiment shows that changing fewer than 0.5% of naturally occurring labels can also improve recall in most evaluated species.

Candidate availability and selection quality affect different parts of this process. In the first NAT update, the two scorers recover almost all correct sequences already present among the predictions, so their similar recovery rates follow from the shared candidate pool. Their acceptance rules still differ in how many incorrect replacements they admit. Multiple rounds address a different constraint by generating new candidates after retraining. The AR results further show that the resulting gain depends on the decoder used at evaluation: improvement under greedy decoding coexists with smaller aggregate gains and species-level declines under beam search.

The controlled study evaluates recovery in an agreement-selected subset of database-search and OmniNovo identifications, where clean and corrupted examples of the same peptide occur in training. The AR experiment extends this comparison to a causal decoder, using one label update and one training seed.

Each additional refinement round incurs the cost of training a new model from scratch. Training times for the 2.4M- and 35M-PSM corpora are given in Appendix A.

Reproducibility statement

Appendices A–C provide the training settings, metric definitions, and AR decoding procedure. Subsequent appendices give detailed results and data sources. Code and checkpoints will be released upon publication.

Ethics statement

This study uses publicly available proteomics data and involves no new human-subject recruitment or intervention. Biological or clinical applications of the predicted peptide assignments require independent validation and false-discovery control.

AI use statement

Generative AI tools assisted manuscript language editing, experiment implementation, result analysis, and cluster-job orchestration. The authors take responsibility for the final content.

REFERENCES

Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928): 198–207, 2003.

Ruedi Aebersold and Matthias Mann. Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620):347–355, 2016.

Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Emilie Devijver, and Yury Maximov. Self-training: A survey. *arXiv preprint arXiv:2202.12040*, 2022.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 233–242. PMLR, 2017.

Yuhan Chen, Shang Qu, Zhiqiang Gao, Yuejin Yang, Xiang Zhang, Sheng Xu, Xinjie Mao, Liujia Qian, Jiaqi Wei, Zijie Qiu, et al. Accurate de novo sequencing of the modified proteome with omninovo. *arXiv preprint arXiv:2512.12272*, 2025.

Joel M Chick, Deepak Kolippakkam, David P Nusinow, Bo Zhai, Ramin Rad, Edward L Huttlin, and Steven P Gygi. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature biotechnology*, 33(7):743–749, 2015.

Jürgen Cox and Matthias Mann. Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26 (12):1367–1372, 2008.

Robertson Craig and Ronald C. Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, 2004.

Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, 4(3):207–214, 2007.

Kevin Eloff, Konstantinos Kalogeropoulos, Amandla Mabona, Oliver Morell, Rachel Catzel, Esperanza Rivera-de Torre, Jakob Berg Jespersen, Wesley Williams, Sam PB van Beljouw, Marcin J Skwark, et al. Instanovo enables diffusion-powered de novo peptide sequencing in large-scale proteomics experiments. *Nature Machine Intelligence*, pp. 1–15, 2025.

Jimmy K Eng, Ashley L McCormack, and John R Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the american society for mass spectrometry*, 5(11):976–989, 1994.

Ari Frank and Pavel Pevzner. Pepnovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry*, 77(4):964–973, 2005.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

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.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, 2013.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

Yangdi Lu and Wenbo He. SELC: Self-ensemble label correction improves learning with noisy labels. *arXiv preprint arXiv:2205.01156*, 2022.

Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry*, 17(20):2337–2342, 2003.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*, pp. 6543–6553. PMLR, 2020.

Rachel M Miller and Lloyd M Smith. Overview and considerations in bottom-up proteomics. *Analyst*, 148(3):475–486, 2023.

David N Perkins, Darryl JC Pappin, David M Creasy, and John S Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS: An International Journal*, 20(18):3551–3567, 1999.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Mikhail M Savitski, Mathias Wilhelm, Hannes Hahne, Bernhard Kuster, and Marcus Bantscheff. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Molecular & Cellular Proteomics*, 14(9):2394–2404, 2015.

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.

Ngoc Hieu Tran, Rui Qiao, Lei Xin, Xin Chen, Chuyi Liu, Xianglilan Zhang, Baozhen Shan, Ali Ghodsi, and Ming Li. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nature methods*, 16(1):63–66, 2019.

Christopher T Walsh, Sylvie Garneau-Tsodikova, and Gregory J Gatto Jr. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angewandte Chemie International Edition*, 44(45):7342–7372, 2005.

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.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 322–330, 2019.

Bo Wen, Jack Freestone, Michael Riffle, Michael J MacCoss, William S Noble, and Uri Keich. Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment. *Nature Methods*, pp. 1–10, 2025.

Melih Yilmaz, William Fondrie, Wout Bittremieux, Sewoong Oh, and William S Noble. De novo mass spectrometry peptide sequencing with a transformer model. In *International Conference on Machine Learning*, pp. 25514–25522. PMLR, 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.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International conference on machine learning*, pp. 7164–7173. PMLR, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Xiang Zhang, Tianze Ling, Zhi Jin, Sheng Xu, Zhiqiang Gao, Boyan Sun, Zijie Qiu, Jiaqi Wei, Nanqing Dong, Guangshuai Wang, et al. π-primenovo: an accurate and efficient non-autoregressive deep learning model for de novo peptide sequencing. *Nature Communications*, 16(1):267, 2025.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

A Training and Implementation Details

Table 6: Default model and optimization configuration.

Parameter	Value	Parameter	Value
Model width	384	Optimizer	AdamW
Feed-forward width	768	Learning rate	3.5×10^{-4}
Total layers	9	Weight decay	3.0×10^{-3}
Attention heads / KV heads	12 / 6	Gradient clipping	2.5
Head dimension	32	Max epochs	100
Dropout	0.2	Warm-up	1 epoch
Activation / norm	SiLU / RMSNorm	Schedule	cosine decay
Loss	CTC	Precision	bfloat16
Max peaks	180	Precursor tolerance	50 ppm
Fragment range	1–6500 Da	Max charge	10+
Remove precursor window	1.0 Da	Isotope error	$[0, 1]$

Experiments use PyTorch 2.3.1, CUDA 12.1, flash-attn 2.7.4.post1, and NumPy 1.23.6 on Ubuntu 22.04. The default batch size is 4096 spectra; the controlled NAT baseline and Co-teaching use 2048. The default NAT training budget is 100 epochs with shuffled data; Table 1 specifies the evaluation points for the baseline comparisons. Controlled comparisons use seed 42, and the MassIVE-KB v1 experiment uses ten paired seeds for training on original and refined labels. Models are initialized from scratch at each refinement round. Training one round on 2.4M PSMs takes approximately 24 hours on two NVIDIA H200 GPUs or 8 hours on eight H200s; a 35M-PSM round takes approximately 72 hours on eight H200s.

Inputs exclude peaks in the precursor window and retain at most 180 peaks. NAT models use greedy decoding with precursor-mass checks on candidate sequences. The vocabulary comprises the standard amino acids with I/L equivalence and the modification tokens used by the sequencer. AR training and decoding are specified in Appendix C.

Corrupted labels. Of the 2,487,323 training PSMs, 1,493,229 are selected for corruption and 994,094 retain their reference labels. Of the selected PSMs, 1,493,219 have labels that differ from the reference under I/L equivalence; the remaining ten retain an equivalent sequence. Reference sequences are used to measure recovery and are excluded from the label-selection rule.

B Evaluation Metrics

Peptide recall. Peptide recall is the fraction of evaluated spectra whose predicted sequence matches the reference sequence after treating I/L as equivalent.

Amino-acid precision and recall. Predicted and target sequences are aligned by residue and cumulative prefix mass. A residue is matched if its mass differs by at most 0.1 Da and the prefix mass differs by at most 0.5 Da. Precision divides matched residues by predicted length; recall divides by target length.

Hyperscore. For matched b and y ions, we use

$$H = \log\left(n_b!\, n_y! \sum_{i=1}^{n_b} I_{b,i} \sum_{j=1}^{n_y} I_{y,j}\right), \tag{2}$$

with fragment intensities normalized to $[1, 100]$. An empty ion-series intensity sum is set to one, and $0! = 1$.

TIC fraction and matched peaks. After removing the precursor (± 1.5 Da), peaks below 1% of base intensity, and all but the top 150 remaining peaks, TIC is the intensity of matched *b*/*y* ions within 20 ppm divided by total retained intensity. Matched peaks is the corresponding ion count; matched intensity is its unnormalized intensity sum.

Precursor error. $\mathrm{PPM} = 10^6(m/z_{\mathrm{obs}} - m/z_{\mathrm{theo}})/(m/z_{\mathrm{theo}})$. We report the mean of these signed errors over PSMs. Positive values indicate that observed precursor m/z exceeds theoretical m/z; negative values indicate the reverse.

C Architecture and Scorer Transfer

C.1 Spectral-angle scoring

For a candidate with K theoretical b/y ions (fragment charges up to 2+), let z_j be the square root of the most intense observed peak within 20 ppm of ion j, or zero when no peak matches. The independent score is the spectral angle to a unit theoretical-ion vector,

$$S_{\mathrm{SA}} = 1 - \frac{2}{\pi} \arccos\left(\frac{\sum_{j=1}^{K} z_j}{\sqrt{K}\sqrt{\sum_{j=1}^{K} z_j^2}}\right). \tag{3}$$

We set $S_{\mathrm{SA}} = 0$ when $K = 0$ or all $z_j = 0$. Both scorers compare the same candidate peptides on the same spectra, without additional precursor filtering. Consensus accepts a prediction when both scores exceed those of the current label. The margin rule also requires $S_{\mathrm{SA}}(x, \hat{y}) - S_{\mathrm{SA}}(x, y) > 0.05$.

Table 7: Scoring rules applied to the first-round NAT candidate pool. Recovery is measured on 1,493,219 corrupted PSMs with I/L equivalence. Precision is the fraction of accepted replacements matching the reference sequence.

Adjudicator	Replaced	Exact recovered (%)	Precision (%)	Wrong replacements
Hyperscore	1,347,700	777,542 (52.07)	57.69	570,158
Spectral angle	1,351,260	776,821 (52.02)	57.49	574,439
Consensus	1,327,871	776,482 (52.00)	58.48	551,389
Consensus + 0.05 margin	1,271,379	772,848 (51.76)	60.79	498,531

The candidate pool contains 1,493,229 PSMs selected for corruption, including ten whose labels are I/L-equivalent to their references. Selection agreement uses the full pool; recovery uses the 1,493,219 non-equivalent labels.

Uncertainty and stratification. PSM-level Wilson 95% intervals for recovery are [51.99, 52.15]% for Hyperscore and [51.94, 52.10]% for spectral angle, using the 1,493,219 non-equivalent labels. These binomial intervals treat PSMs as independent observations. Table 8 gives reference-match rates across length and charge groups. The largest absolute difference between scorers is 0.20 percentage points for length and 0.07 points for charge.

Table 8: Reference matches after label selection, stratified by reference-peptide length and precursor charge. Each grouping covers the full 1,493,229-PSM candidate pool, including the ten I/L-equivalent initial labels.

Group	PSMs	Hyperscore (%)	Spectral angle (%)
Reference length			
5–7	27,521	90.63	90.43
8–12	604,370	71.70	71.61
13–18	592,768	45.52	45.49
19–24	198,165	22.92	22.92
25+	70,405	5.75	5.75
Precursor charge			
1+	8,528	68.06	68.03
2+	775,442	64.02	63.95
3+	536,713	45.26	45.23
4+	142,854	21.72	21.70
5+	29,692	4.69	4.69

Scores of reference sequences. Across the full candidate pool, mean spectral angle is 0.1297 for the supplied labels and 0.3577 for model predictions. We also score reference peptides on a systematic sample of 149,323 PSMs, taking every tenth record. Their mean spectral angle is 0.4065, and the reference out-scores the supplied label in 98.62% of cases. This comparison shows that the score strongly favors reference sequences over corrupted labels in this setting.

C.2 AUTOREGRESSIVE EXPERIMENT

Training. The AR model has a nine-layer spectrum encoder of width 384 and a nine-layer causal Transformer decoder with 12 heads, totaling 29.3M parameters. Both models use teacher forcing, BOS/EOS tokens, label smoothing of 0.1, dropout of 0.2, and AdamW with learning rate 3.5×10^{-4} and weight decay 0.003. Each is initialized with seed 42 and trained for 100 epochs on eight GPUs with global batch size 512. Each epoch samples 2,487,323 PSMs with replacement. Evaluation uses the models at the end of training.

Decoder selection. Greedy decoding uses a 0.1-Da tolerance for EOS acceptance and mass-overshoot checks, with isotope offsets 0 and 1. Beam search uses precursor-mass filtering inspired by Casanovo (Yilmaz et al., 2024). It preserves model log probabilities before token masking, restricts modification tokens to legal positions, and requires at least five residues before EOS, with at most 79 pre-EOS tokens. Completed sequences must fall within the selected ppm tolerance of observed precursor m/z, after isotope offsets 0 or 1. The error is normalized by observed m/z. A spectrum with no admissible completion counts as incorrect. Completed sequences are ranked by the sum of token log probabilities divided by token count to the power α, including EOS.

We choose the beam configuration by exact recall on 1,024 randomly selected development spectra. The selected configuration uses width 10, a 50-ppm tolerance, and $\alpha = 0$, with development recall of 0.7305 compared with 0.6064 for greedy decoding. It is fixed for label generation and external evaluation. Ties are resolved by smaller beam width, then tighter tolerance, then $\alpha = 1$. The development set contains 130,998 spectra and has no exact spectrum or I/L-normalized reference-peptide overlap with training. Spectrum identity is defined by precursor m/z, charge, and the ordered peak list.

Table 9: Beam configurations evaluated on the development subset. The selected configuration is in bold.

Beam width	Tolerance (ppm)	α
1	50	1
5	20	1
5	50	1
5	100	1
5	50	0
10	50	1
10	**50**	**0**

Label refinement. Beam decoding generates a candidate for every training spectrum. Hyperscore accepts 1,375,451 replacements, including 1,099,797 exact recoveries. The updated dataset contains 2,090,696 correct labels.

Table 10: AR peptide recall before and after one PCSC update, evaluated on all 89,924 nine-species spectra.

Species	n	Greedy	Mass-filtered beam
Bacillus	9,804	$0.2470 \rightarrow 0.3221$	$0.4013 \rightarrow 0.4000$
Clambacteria	9,978	$0.2087 \rightarrow 0.2665$	$0.3097 \rightarrow 0.3158$
Honeybee	9,961	$0.1947 \rightarrow 0.2496$	$0.3129 \rightarrow 0.3137$
Human	10,066	$0.1607 \rightarrow 0.2348$	$0.2638 \rightarrow 0.2885$
M. mazei	10,060	$0.2171 \rightarrow 0.2865$	$0.3494 \rightarrow 0.3511$
Mouse	10,091	$0.2262 \rightarrow 0.2801$	$0.3337 \rightarrow 0.3296$
Ricebean	10,159	$0.2574 \rightarrow 0.3096$	$0.3980 \rightarrow 0.3918$
Tomato	9,729	$0.2869 \rightarrow 0.3705$	$0.4326 \rightarrow 0.4430$
Yeast	10,076	$0.2503 \rightarrow 0.3492$	$0.4072 \rightarrow 0.4222$
Macro	89,924	$0.2277 \rightarrow 0.2965$	$0.3565 \rightarrow 0.3617$
Micro	89,924	$0.2275 \rightarrow 0.2963$	$0.3562 \rightarrow 0.3615$

Paired inference and uncertainty. Both AR models are evaluated on the same spectra with fixed decoding parameters. We estimate uncertainty by resampling reference-peptide clusters within each species, using 2,000 bootstrap replicates. The 95% intervals in Table 4 describe test-sample uncertainty for this pair of models trained with seed 42.

Predictions on corrupted training examples. Table 11 compares recovery and reproduction of injected labels on the same 20,000 corrupted training PSMs.

Table 11: AR predictions on 20,000 corrupted training PSMs before and after refinement (%).

Decoding	Reference sequence recovered	Injected label reproduced
Greedy	$51.015 \rightarrow 74.120$	$0.275 \rightarrow 1.970$
Mass-filtered beam	$73.740 \rightarrow 74.670$	$0.125 \rightarrow 2.005$

D PHYSICAL METRICS ACROSS UPDATES

Table 12: Mean physical metrics for all PSMs across PCSC rounds. PPM is averaged with its sign.

Round	Hyperscore	TIC	Matched peaks	Matched intensity	Signed PPM
0	17.9868	0.2002	8.43	215.64	1612.77
1	31.5011	0.3653	14.73	388.18	76.33
2	33.2284	0.3818	15.51	407.92	41.83
3	34.0356	0.3898	15.87	417.19	22.17
4	34.4680	0.3939	16.06	422.03	16.51
5	34.8053	0.3965	16.21	425.41	14.49

Table 13: Mean physical metrics for PSMs selected for corruption. PPM is averaged with its sign.

Round	Hyperscore	TIC	Signed PPM
0	5.8991	0.0629	2685.61
1	28.3746	0.3376	126.34
2	31.2330	0.3650	68.88
3	32.5620	0.3781	36.14
4	33.2689	0.3849	26.73
5	33.8207	0.3891	23.36

At round 0, PSMs selected for corruption have a median absolute precursor error of 8.48 ppm. The 133,266 labels that omit an N-terminal modification present in the reference account for 99.38% of total absolute precursor error.

E NINE-SPECIES RESULTS

Table 14: MassIVE-KB v1 peptide recall across ten paired seeds. Values are mean $\pm$ standard deviation in percentage points; p is from a paired t-test.

Species	Original	Refined	p
A. mellifera	54.30 ± 0.12	54.87 ± 0.04	1.99×10^{-8}
B. subtilis	63.50 ± 0.11	63.21 ± 0.06	1.81×10^{-5}
C. endoloripes	45.25 ± 0.06	45.55 ± 0.06	2.85×10^{-6}
H. sapiens	58.91 ± 0.08	59.51 ± 0.04	4.35×10^{-8}
M. mazei	57.52 ± 0.10	57.73 ± 0.05	6.93×10^{-4}
M. musculus	56.23 ± 0.07	56.34 ± 0.05	3.94×10^{-3}
S. cerevisiae	61.62 ± 0.16	62.03 ± 0.03	9.31×10^{-6}
S. lycopersicum	63.48 ± 0.07	62.36 ± 0.05	8.84×10^{-11}
V. mungo	61.63 ± 0.10	62.05 ± 0.12	5.73×10^{-5}

Table 15: Amino-acid precision and recall (%) for the same runs. Each cell is Original $\rightarrow$ Refined.

Species	AA precision	AA recall
A. mellifera	$71.19 \rightarrow 72.00$	$72.00 \rightarrow 72.70$
B. subtilis	$77.50 \rightarrow 77.97$	$78.45 \rightarrow 78.90$
C. endoloripes	$63.92 \rightarrow 64.24$	$65.74 \rightarrow 66.00$
H. sapiens	$71.29 \rightarrow 71.60$	$72.38 \rightarrow 72.63$
M. mazei	$74.81 \rightarrow 75.65$	$75.76 \rightarrow 76.52$
M. musculus	$77.16 \rightarrow 77.91$	$78.34 \rightarrow 78.93$
S. cerevisiae	$71.76 \rightarrow 72.36$	$73.19 \rightarrow 73.80$
S. lycopersicum	$77.19 \rightarrow 77.38$	$78.63 \rightarrow 78.71$
V. mungo	$73.17 \rightarrow 73.68$	$74.63 \rightarrow 74.95$

F Post-Translational Modification Study

We use the modified PSMs from the training partitions of the projects listed in Table 16. The 3,001,570 PSMs include both singly and multiply modified peptides.

Table 16: Public projects used in the PTM analysis.

PRIDE			
PXD037285	PXD039441	PXD020271	PXD052627
PXD002425	PXD012007	PXD000559	PXD005536
PXD021188	PXD046571	PXD003700	PXD011926
PXD003281	PXD014719	PXD027949	PXD009994
PXD002800	PXD009070	PXD005252	
iProX: IPX001804000			

A PSM can contribute to more than one modification category in Table 3. Hyperscore and matched peaks are measured on the PSMs whose labels change.

G Hyperscore Length-Bias Analysis

Table 17: Label-length and score changes from round 0 to round 5. Each of the 1,458,082 initially corrupted PSMs with a changed final label is counted once, using I/L equivalence.

Length direction	Count (%)	Δ Hyperscore	Δ TIC
Shorter	470,546 (32.3%)	+26.4	+0.292
Same length	932,385 (63.9%)	+29.8	+0.359
Longer	55,151 (3.8%)	+26.5	+0.282

For round 0→1, the 1,347,700 replacements comprise 63.1% unchanged-length, 32.4% shorter, and 4.5% longer labels. For round 4→5, the corresponding proportions among 81,733 replacements are 69.7%, 13.4%, and 16.9%. These summaries count changes within each update; Table 17 compares the initial and final labels of each PSM.

H Attention Diagnostics

We match observed peaks to theoretical b/y ions of the reference peptide, with charge 1+ or 2+, within 0.5 Da. For each spectrum and decoder layer, alignment is the attention assigned to matched peaks, summed over heads and decoder positions, divided by total attention to all encoder positions, including the latent token. Figure 7 reports the mean and standard deviation across spectra. Enrichment divides mean alignment by the mean fraction of peaks matching theoretical ions (0.177). Entropy is $-\sum_j a_j \log a_j$ over encoder positions, averaged over heads, decoder positions, and spectra. M_0 denotes training with 60% global noise; Iter 4 denotes training after four updates.

Figures 8–10 show three examples from M_0: an exact match, a prediction differing only by I/L, and a partial C-terminal match. Each includes the annotated spectrum and cross-attention at layers 0, 5, and 6. Heatmap cells show the maximum attention weight across heads for each decoder-position/peak pair. At each displayed position, the right-hand bar sums these maximum weights on matched b/y peaks and divides by their sum over all observed peaks, excluding the latent token. The figures show the first $\min(L + 10, T)$ decoder positions, where L is the reference length and T the number of decoder positions. Caption values average the bars over these displayed positions. The dashed line marks the fraction of observed peaks matched to theoretical ions for that spectrum.

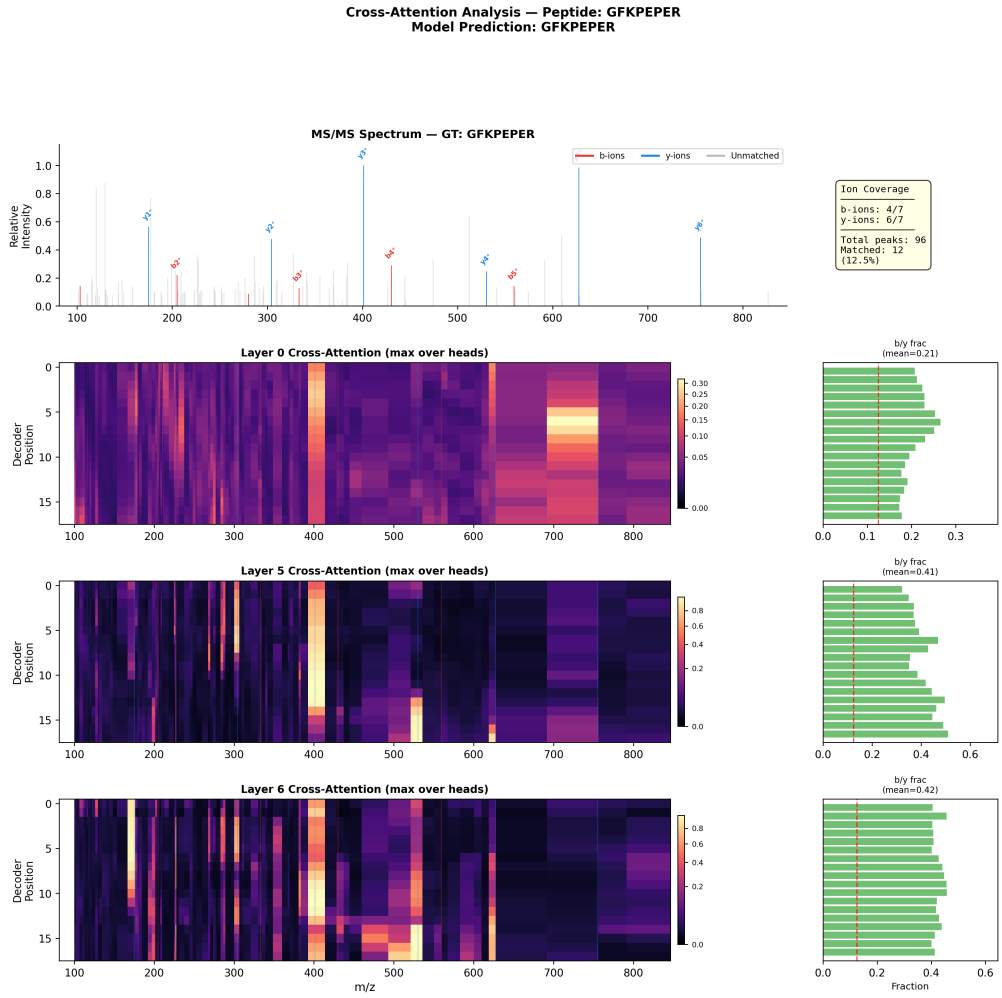


Figure 8: **Exact sequence recovery: GFKPEPER.** The prediction matches the reference peptide. Attention becomes more concentrated between layer 0 and layer 5, with mean displayed b/y fractions of 0.21 and 0.41, respectively. Colored peaks mark ions of the reference sequence.

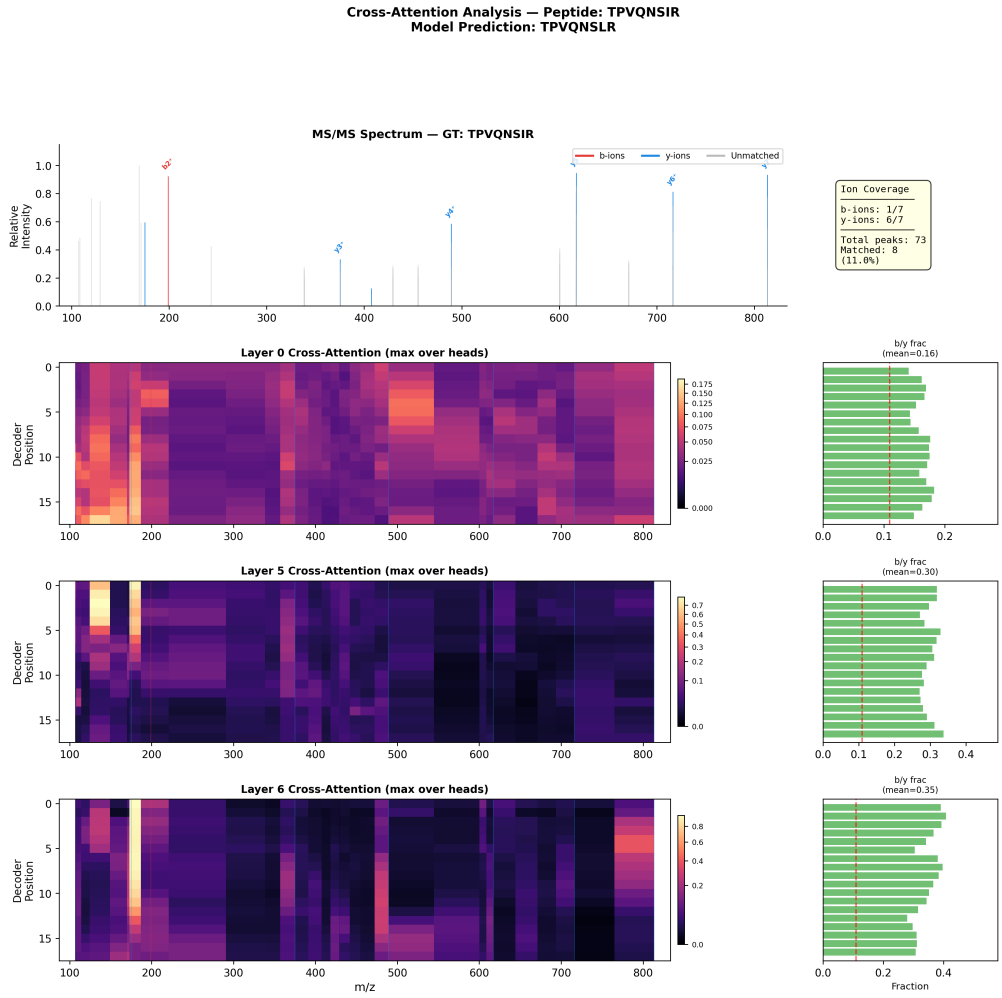


Figure 9: **I/L-equivalent prediction: TPVQNSLR for TPVQNSIR.** The sequences differ only at the I/L position and count as a match under the evaluation rule. The mean displayed b/y fraction reaches 0.30 at layer 5, with concentration on a small number of peaks.

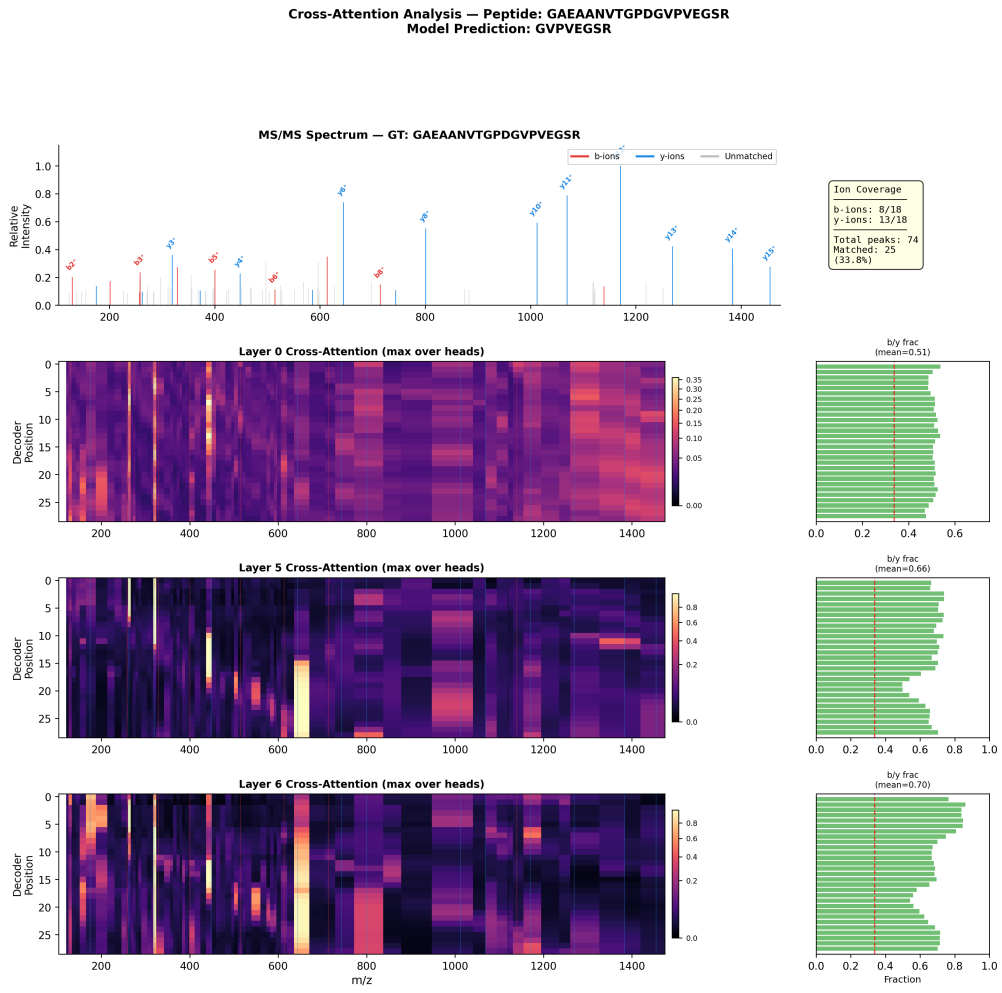


Figure 10: **Partial sequence recovery: GVPVEGSR for GAEAANVTGPDGVPVEGSR.** The prediction recovers the C-terminal segment. Attention aligns strongly with reference fragment peaks, reaching a mean displayed b/y fraction of 0.70 at layer 6, while the N-terminal segment is absent from the predicted sequence.