NON-TRYPTIC TERMINUS SIMULATION FOR GENERALIZABLE *De Novo* PEPTIDE SEQUENCING

Anonymous authors
Paper under double-blind review

ABSTRACT

Tandem mass spectrometry (MS/MS)-based *de novo* peptide sequencing provides a high-throughput approach for identifying peptide sequences without relying on a predefined protein sequence database, making it valuable for discovering novel protein sequences. The performance of *de novo* peptide sequencing AI model is fundamentally dependent on the scale and diversity of training data. However, existing training datasets are dominated by tryptic peptides, whose C termini are strongly enriched for lysine (K) or arginine (R) due to the cleavage specificity of trypsin. This constrained cleavage pattern substantially limits the diversity of peptide sequences and fragmentation patterns represented in training data. Here, we propose Non-Tryptic Terminus Simulation (NTS), an interpretable and general data augmentation method that generates MS/MS spectra with non-tryptic C termini directly from existing tryptic peptide-spectrum matches. NTS removes the C-terminal residue from a tryptic peptide and transforms its corresponding spectrum by selectively removing or mass-shifting fragment ions dependent on the removed residue, while preserving their original intensities. This targeted and interpretable transformation substantially expands the diversity of training data without requiring additional wet-lab experiments or spectrum-generation models. Across multiple widely used datasets and three representative *de novo* sequencing models, Casanovo, π-HelixNovo, and π-HelixNovo2, NTS significantly improves model performance and generalization, yielding average relative peptide recall gains of 50.9%, 12.4%, and 17.2%, respectively. These results demonstrate that NTS provides an interpretable and broadly applicable solution to the training-data bottleneck in *de novo* peptide sequencing.

1 INTRODUCTION

Tandem mass spectrometry (MS/MS) is now widely used as the primary technology for large-scale proteomic analysis because of its high sensitivity, throughput, and capability to characterize peptides from complex biological samples (Aebersold & Mann, 2003; Guo et al., 2025). In a typical MS/MS-based proteomics workflow (Appendix A.1), proteins are enzymatically digested into peptides (Coon et al., 2005), whose sequences are subsequently inferred from their tandem mass spectra (He et al., 2015).

Two major approaches have been developed for peptide identification (Figure 1): database search (Eng et al., 2011) and *de novo* peptide sequencing (Seidler et al., 2010). Database search identifies experimental spectra by matching them to theoretical spectra derived from peptides in a reference protein database and is reliable when the database adequately represents the proteins present in a sample (Ramos-Fernandez et al., 2008). However, its search space is fundamentally constrained by the completeness and accuracy of the reference proteome, preventing the identification of novel peptides that fall outside its scope (Yates et al., 1995). Therefore, *de novo* peptide sequencing is proposed to overcome this limitation by directly inferring peptide sequences from MS/MS spectra without relying on a predefined sequence database (Seidler et al., 2010).

Despite substantial progress in *de novo* peptide sequencing, driven primarily by advances in model architectures from dynamic programming (Eddy, 2004) to deep neural networks (Yi et al., 2016; Sun et al., 2026), both model capacity and generalization remain constrained by the scale and diversity of training data. Specifically, most existing datasets are derived from tryptic digestion, whose prefer-

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.

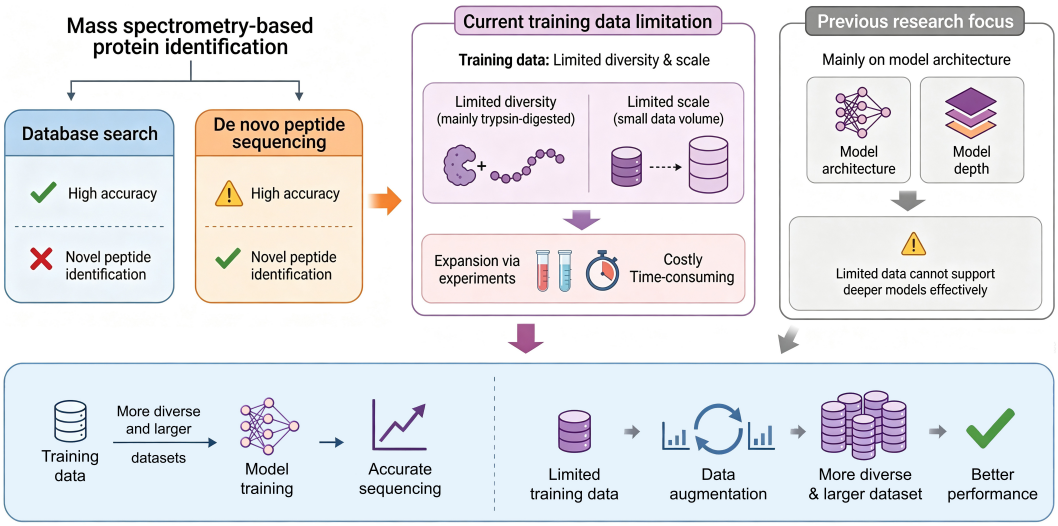

Figure 1: Overview of mass spectrometry-based protein identification, training data bottleneck in *de novo* peptide sequencing, and the proposed data augmentation strategy.

ence for cleavage after lysine (K) and arginine (R) (Appendix A.2) introduces a strong bias in peptide C-termini (Wang et al., 2018), thereby limiting the diversity of sequences and fragmentation patterns represented during training (Beslic et al., 2023). Meanwhile, acquiring large-scale peptide-spectrum matches (PSMs) with diverse non-tryptic termini remains costly and time-consuming, requiring extensive wet-lab experiments and database searches (Shouman et al., 2022). Therefore, an attractive alternative is to computationally generate MS/MS spectra corresponding to non-tryptic or otherwise diverse peptide termini.

However, directly generating realistic MS/MS spectra for arbitrary peptide sequences remains challenging. A generative model must capture not only b- and y-ion signals and their variable intensities (Gessulat et al., 2019), but also non-ideal peaks caused by internal fragmentation, neutral losses, chemical background, and other experimental effects (Lapin et al., 2024). Errors in reproducing these spectral features may cause generated spectra to differ substantially from real experimental spectra, limiting their effectiveness for training *de novo* sequencing models (Yang et al., 2026b).

In this work, we propose *Non-Tryptic Terminus Simulation* (NTS), an interpretable and general data augmentation method for simulating MS/MS spectra with diverse, non-tryptic C-termini directly from existing experimental PSMs (Figure 2). Specifically, NTS removes the C-terminal residue from a tryptic peptide and transforms its corresponding spectrum by selectively removing or mass-shifting fragment ions dependent on the removed residue, while preserving their original intensities. Meanwhile, unrelated fragment peaks are left unchanged, preserving the intrinsic spectral complexity and non-ideal characteristics of experimental MS/MS spectra. We use NTS to perform data augmentation in terms of data scale and diversity, and train multiple representative *de novo* peptide sequencing models. Across multiple widely used datasets, NTS consistently improves sequencing performance and substantially enhances generalization across species and digestion conditions. These results demonstrate the effectiveness of NTS for alleviating the training-data bottleneck in *de novo* peptide sequencing.

Our contributions can be summarized as follows:

- We reveal that the limited diversity of tryptic training data constitutes a fundamental bottleneck for the generalization of deep learning-based *de novo* peptide sequencing models, highlighting the importance of training-data diversity.
- We propose NTS, an interpretable data augmentation method that transforms real experimental PSMs into spectra corresponding to diverse non-tryptic C-termini without requiring additional wet-lab experiments or spectrum-generation models.

- We construct augmented training datasets with substantially increased C-terminal amino-acid diversity and demonstrate that NTS consistently improves the performance and cross-condition generalization of multiple *de novo* peptide sequencing models across diverse datasets.

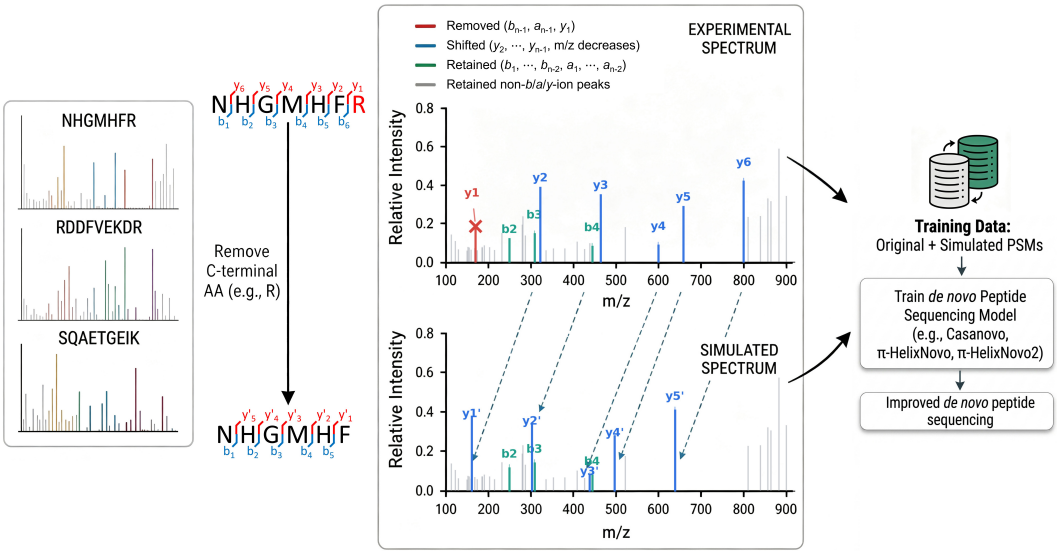

Figure 2: Overview of the NTS workflow. Multiple fragment charge states and isotope envelopes are considered when identifying peaks to be removed or shifted.

2 Related Work

2.1 *De novo* Peptide Sequencing Method

De novo peptide sequencing infers peptide sequences directly from tandem mass spectra without relying on sequence databases (Seidler et al., 2010). Early approaches, such as PEAKS (Ma et al., 2003), formulated the task using dynamic programming, while subsequent methods increasingly turned to deep neural networks. DeepNovo (Tran et al., 2017) pioneered neural sequence modeling for *de novo* sequencing, followed by CNN-based (Aghdam et al., 2017) and LSTM-based (Yu et al., 2019) approaches including pNovo (Chi et al., 2010), PointNovo (Qiao et al., 2021), and PepNet (Liu et al., 2023). More recently, Transformer-based models (Vaswani et al., 2017), pioneered by Casanovo (Yilmaz et al., 2022; 2024), have become the dominant paradigm, with representative methods including GraphNovo (Mao et al., 2023), π-HelixNovo (Yang et al., 2024), NovoB (Lee & Kim, 2024), π-PrimeNovo (Zhang et al., 2025), and π-HelixNovo2 (Yang et al., 2026a).

2.2 Data-Augmentation Method for Mass Spectrometry

Mass spectrum generation is a widely used approach for predicting theoretical MS/MS spectra from peptide sequences and precursor charge states by estimating the intensities of b- and y-ions (Gessulat et al., 2019; Zhou et al., 2017). It plays an important role in database search, where spectral similarity between experimental and theoretical spectra is used for peptide identification, with representative methods including Prosit (Gessulat et al., 2019) and pDeep (Zhou et al., 2017). However, these theoretical spectra remain substantially different from real experimental MS/MS spectra (Lapin et al., 2024). Beyond discrepancies in b/y-ion intensities, experimental spectra also contain diverse non-b/y-ion signals that are absent from these predictions. Consequently, theoretical spectra are often too idealized to serve as direct training data for *de novo* peptide sequencing, as they fail to capture the complex and non-ideal spectral characteristics of real experimental data (Yang et al., 2026b).

Complementary-ion reconstruction provides another approach for enhancing experimental MS/MS spectra (Mao et al., 2020; Yang et al., 2024). Given the precursor mass and an experimental spectrum, complementary ions can be calculated to recover fragment-ion signals missing due to incomplete ionization or fragmentation during acquisition. This approach effectively increases the usable fragment-ion information within individual spectra and has been applied to both database search (Mao et al., 2020) and *de novo* peptide sequencing, such as π-HelixNovo (Yang et al., 2024), yielding consistent performance improvements. However, complementary-ion reconstruction only enriches information within existing spectra—it neither increases the number of PSMs nor expands the diversity of peptide sequences represented in the training set.

3 METHODOLOGY

3.1 OVERVIEW

NTS is a chemistry-based data augmentation method for *de novo* peptide sequencing. Given a tryptic PSM (P, S, z), where $P = x_1x_2\cdots x_n$ is the peptide sequence, x_i denotes the i-th amino acid residue, $S = (mz_i, I_i)_{i=1}^{N}$ is the observed MS2 spectrum, and z is the precursor charge, NTS deterministically constructs a new PSM (P', S', z) for the truncated peptide

$$P' = x_1x_2\cdots x_{n-1}, \tag{1}$$

by removing the C-terminal residue x_n.

The key idea is to transform the experimental spectrum directly rather than generate a new spectrum from the peptide sequence. Peptide fragmentation produces sequence-informative $a/b/y$ ions, while experimental MS2 spectra also contain neutral-loss, internal-fragment, isotope, and background signals (Roepstorfe, 1984; Paizs & Suhai, 2005). NTS therefore utilizes the known mass shift associated with residue removal to selectively modify only the fragment ions affected by the removed residue. Unlike spectrum prediction methods such as pDeep and Prosit, which learn to predict b/y fragment-ion intensities from peptide sequences (Zhou et al., 2017; Gessulat et al., 2019), NTS preserves the experimentally measured intensities of peaks. Specifically, affected peaks are either deleted or shifted in m/z, while all other peaks remain unchanged. This design preserves the experimental characteristics of the original spectrum while extending the training distribution beyond canonical tryptic termini.

3.2 THEORETICAL FRAGMENT IONS

Let $r(x_i)$ denote the monoisotopic residue mass of amino acid x_i, m_N the mass of an optional N-terminal modification, and z_f the fragment charge state. The neutral masses of the standard b-, a-, and y-ion series are

$$b_k = m_N + \sum_{i=1}^{k} r(x_i), \qquad k = 1, \ldots, n-1, \tag{2}$$

$$a_k = b_k - m_{\mathrm{CO}}, \qquad k = 1, \ldots, n-1, \tag{3}$$

$$y_k = m_{\mathrm{H_2O}} + \sum_{i=n-k+1}^{n} r(x_i), \qquad k = 1, \ldots, n-1. \tag{4}$$

Here, $m_{\mathrm{CO}} = 27.994915$ Da and $m_{\mathrm{H_2O}} = 18.010565$ Da. The observed m/z of a neutral fragment with mass M and charge z_f is

$$mz(M, z_f) = \frac{M + z_f m_p}{z_f}, \tag{5}$$

where $m_p = 1.007276$ Da is the proton mass. The a-ion series is derived from the corresponding b-ion series by CO loss, and both are N-terminal fragment series (Roepstorfe, 1984; Johnson et al., 1988).

3.3 THE $b/a/y$ IONS AFFECTED BY C-TERMINAL RESIDUE REMOVAL

After removal of x_n, the $b/a/y$ ions of P can be categorized into three classes based on their relationship to the truncated peptide P'.

Unaffected ions. The ions $b_1, \ldots, b_{n-2}$ and $a_1, \ldots, a_{n-2}$ depend only on the residues $x_1, \ldots, x_{n-2}$ and are not affected by the removal of x_n. Therefore, their corresponding observed peaks are retained unchanged.

Ions to be removed. The ions b_{n-1}, a_{n-1}, and y_1 are removed from the simulated spectrum. The ions b_{n-1} and a_{n-1} are the highest-order N-terminal fragments of the original peptide and depend on the C-terminal residue x_n, while y_1 consists exclusively of the removed residue x_n. Consequently, none of these ions corresponds to a standard $a/b/y$ fragment of the truncated peptide P'.

Ions to be shifted. The ions $y_2, \ldots, y_{n-1}$ remain chemically valid C-terminal fragments after removal of x_n. However, each fragment loses the mass contribution of x_n, resulting in a corresponding decrease in neutral mass. Because the removal of x_n also reduces the number of residues in each y-ion by one, the original y_k ion corresponds to the y_{k-1} ion in the truncated peptide P'. Thus, the new neutral masses are given by

$$y'_{k-1} = y_k - r(x_n), \qquad k = 2, \ldots, n-1. \tag{6}$$

For a fragment with charge z_f, the corresponding m/z shift is

$$\Delta mz = \frac{r(x_n)}{z_f}. \tag{7}$$

Importantly, NTS modifies only the presence or m/z position of affected peaks, while preserving the measured intensity I_i of every retained peak. No intensity values are estimated or rescaled.

3.4 MULTI-CHARGE AND ISOTOPE-AWARE TARGET ENUMERATION

In experimental mass spectra, fragment ions may occur at multiple charge states (Tran et al., 2017) and may be observed as isotope envelopes rather than isolated monoisotopic peaks (Senko et al., 1995). To account for these experimental characteristics, NTS enumerates theoretical target m/z values for each accessible fragment charge state

$$z_f \in \{1, \ldots, Z\}, \qquad Z = \min(z_{\max}, z), \tag{8}$$

where $z_{\max}$ denotes the maximum fragment charge considered by NTS ($z_{\max} = 3$ in all experiments), while the upper bound is further constrained by the precursor charge z, as a fragment ion cannot carry a charge state higher than that of its precursor. NTS also enumerates the corresponding isotope peaks $M+0, M+1, \ldots, M+k_{\text{iso}}$ ($k_{\text{iso}} = 2$ in all experiments), where the m/z of the j-th isotope peak is offset from the monoisotopic peak by

$$j \cdot \Delta_{\text{iso}}(z_f), \qquad \Delta_{\text{iso}}(z_f) = \frac{1.003355}{z_f}\ \text{Da}, \qquad j = 0, \ldots, k_{\text{iso}}. \tag{9}$$

Combining both dimensions, the full set of theoretical targets to be *removed* is

$$\mathcal{D} = \bigcup_{z_f=1}^{Z} \bigcup_{j=0}^{k_{\text{iso}}} \Big\{ mz(b_{n-1}, z_f) + j\Delta_{\text{iso}}(z_f),\ mz(a_{n-1}, z_f) + j\Delta_{\text{iso}}(z_f),\ mz(y_1, z_f) + j\Delta_{\text{iso}}(z_f) \Big\}, \tag{10}$$

and the full set of targets to be *shifted* is

$$\mathcal{T} = \bigcup_{z_f=1}^{Z} \bigcup_{j=0}^{k_{\text{iso}}} \bigcup_{k=2}^{n-1} \Big\{ \big(mz(y_k, z_f) + j\Delta_{\text{iso}}(z_f) \big) \Big\}, \tag{11}$$

Considering multiple charge states and isotope envelopes enables NTS to identify experimentally observed peaks corresponding to the affected $b/a/y$ fragment ions, including their multiply charged and isotopic forms. This ensures that residual peaks originating from the same fragment are not incorrectly retained after the corresponding canonical fragment is removed or shifted.

All other observed peaks are preserved without modification. These peaks may correspond to neutral-loss fragments, internal fragments, unaffected isotope peaks, background signals, or other experimentally observed features unrelated to the affected canonical $a/b/y$ ions. Retaining these signals preserves the intrinsic spectral complexity and non-ideal characteristics of experimental MS2 spectra, rather than reducing them to an idealized collection of canonical fragment ions. As a result, NTS produces simulated spectra that remain faithful to the complexity of real experimental measurements while effectively extending the training distribution to peptides with non-tryptic termini.

3.5 Tolerance-Based Peak Matching

For each observed peak $(mz_i, I_i) \in S$, NTS searches the entire set of theoretical targets $\mathcal{D} \cup \mathcal{T}$ constructed in Section 3.4 and identifies the closest target,

$$t^\star = \arg \min_{t \in \mathcal{D} \cup \mathcal{T}} |mz_i - t|. \tag{12}$$

The peak is considered a match to $t^\star$ only if the minimum mass difference is within the corresponding ppm-based tolerance with an absolute floor,

$$|mz_i - t^\star| \;\le\; \mathrm{tol}(t^\star) = \max\left(t^\star \cdot \mathrm{ppm} \cdot 10^{-6},\; \mathrm{tol}_{\mathrm{floor}}\right), \tag{13}$$

where we set $\mathrm{ppm} = 20$ and $\mathrm{tol}_{\mathrm{floor}} = 0.2$ Da in all experiments. The corresponding peak is removed when $t^\star \in \mathcal{D}$, or its m/z is decreased by $r(x_n)/z_f$ when $t^\star \in \mathcal{T}$. If no target falls within the tolerance, the peak is left unchanged. Here, the relative ppm tolerance scales the matching window with ion mass, providing a consistent relative mass-error criterion across the m/z range. This avoids an overly permissive window at low m/z or an overly restrictive window at high m/z that can arise with a fixed Da tolerance. The absolute tolerance floor further prevents the matching window from becoming excessively narrow for low-mass ions.

3.6 Precursor and Sequence Update

After processing all affected peaks, the remaining peaks are sorted in ascending order of m/z, and the corresponding peptide sequence is defined as $P' = x_1 \cdots x_{n-1}$. The precursor m/z is then recalculated from the neutral mass of the truncated peptide using the original precursor charge z:

$$M'_{\mathrm{neutral}} = m_N + \sum_{i=1}^{n-1} r(x_i) + m_{\mathrm{H_2O}}, \tag{14}$$

$$mz'_{\mathrm{precursor}} = \frac{M'_{\mathrm{neutral}} + z m_p}{z}. \tag{15}$$

The resulting triplet (P', S', z) constitutes a valid simulated PSM with a non-tryptic C terminus and can be directly used to train *de novo* sequencing models.

3.7 Algorithm Summary

Algorithm 1 summarizes the procedure for generating simulated non-tryptic PSMs using the proposed NTS method.

Algorithm 1 Non-Tryptic Terminus Simulation (NTS)

Require: Peptide $P = x_1 \cdots x_n$, spectrum $S = (mz_i, I_i)_{i=1}^{N}$, precursor charge z
Compute the theoretical fragment ions $b_1, \ldots, b_{n-1}$, $a_1, \ldots, a_{n-1}$, and $y_1, \ldots, y_{n-1}$
Construct $\mathcal{D}$ and $\mathcal{T}$ by enumerating the theoretical m/z values of b_{n-1}, a_{n-1}, y_1 and $y_2, \ldots, y_{n-1}$, respectively, across all allowed fragment charges z_f and isotope states
for each observed peak $(mz_i, I_i) \in S$ **do**
  Find the nearest target $t^\star$ in $\mathcal{D} \cup \mathcal{T}$
  if $|mz_i - t^\star| > \mathrm{tol}(t^\star)$ **then**
    Keep the peak unchanged
  else if $t^\star \in \mathcal{D}$ **then**
    Remove the peak
  else
    $mz_i \leftarrow mz_i - r(x_n)/z_f$
  end if
end for
Recompute the precursor m/z for the truncated peptide $P' = x_1 \cdots x_{n-1}$ at charge z
Sort the remaining peaks in ascending order of m/z
return Simulated PSM (P', S', z)

4 EXPERIMENTS

4.1 DATASETS AND EVALUATION METRICS

In this study, we used the nine-species benchmark dataset (Trypsin) (Tran et al., 2017) as the training set for *de novo* peptide sequencing models. To systematically evaluate cross-species and cross-enzyme generalization, we further assessed the trained models on 11 independent test datasets:

(1) human HeLa cell-line samples from the Association of Biomolecular Resource Facilities (ABRF; Trypsin) (Qiao et al., 2021); (2) human lung cancer patient samples (PXD012645; Trypsin) (Sim et al., 2019); (3) *Mus musculus* brain tissue samples (PXD008844; Trypsin) (Zhu et al., 2018); (4) *Plasmodium berghei* ANKA strain samples (PXD010559; Trypsin) (Shears et al., 2019); (5) *Arabidopsis thaliana* young flower bud samples (PXD016315; Trypsin) (Fíla et al., 2020); (6) *Caenorhabditis elegans* samples (PXD013233; Trypsin) (Tharyan et al., 2020); (7) *Escherichia coli* samples (PXD016001; Trypsin) (Yu et al., 2020); (8) bottlenose dolphin (*Tursiops truncatus*) samples from six tissues, including brain, blubber, kidney, liver, muscle, and skin (PXD008808; Trypsin) (Neely et al., 2023); (9) human HeLa cell-line samples digested with four proteases (PXD004452; Chymotrypsin, Glu-C, Lys-C, and Trypsin) (Bekker-Jensen et al., 2017); (10) C-terminal and synthetic peptides (MSV000094014; non-enzymatic) (Yilmaz et al., 2024); and (11) human IgG1 samples digested with Asp-N, Chymotrypsin, Glu-C, Lys-C, Proteinase-K, and Trypsin, together with mouse IgG1 samples digested with Asp-N, Chymotrypsin, and Trypsin (MSV000079801) (Tran et al., 2016).

We evaluated *de novo* peptide sequencing performance using three widely adopted metrics: amino-acid precision (AA precision), amino-acid recall (AA recall), and peptide recall. These metrics are computed based on the number of correctly matched amino acids. Specifically, a predicted amino acid is regarded as correctly matched if the absolute mass difference between its predicted mass m_{pred} and the corresponding ground-truth mass m_{ground} falls within 0.1 Da when no prefix/suffix mass offset is present, or within 0.5 Da when such an offset is present. The three metrics are defined as follows:

$$\text{AA Precision} = \frac{N^a_{\mathrm{match}}}{N^a_{\mathrm{pred}}}, \tag{16}$$

$$\text{AA Recall} = \frac{N^a_{\mathrm{match}}}{N^a_{\mathrm{ground}}}, \tag{17}$$

$$\text{Peptide Recall} = \frac{N^p_{\mathrm{match}}}{N^p_{\mathrm{pred}}}. \tag{18}$$

Here, N^a_{pred} and N^a_{ground} denote the total numbers of predicted and ground-truth amino acids, respectively, while N^a_{match} denotes the number of amino acids correctly matched between the predicted and ground-truth peptide sequences. N^p_{match} denotes the number of fully matched peptides, defined as peptides for which all constituent amino acids are correctly matched, and N^p_{pred} denotes the total number of peptides predicted by the model for the corresponding MS2 spectra.

4.2 Baselines and Experimental Setups

We used three representative models as baselines in our experiments: Casanovo, π-HelixNovo, and π-HelixNovo2, all of which were introduced in the Related Work section. Each model was trained using the optimal hyperparameters specified in its respective codebase. All models were trained on a single NVIDIA RTX 4090 GPU for 600k steps with a batch size of 32, using the Adam optimizer (Ismail, 2019) with a weight decay of 10^{-5} and a cosine learning rate scheduler with 100k warm-up steps. For each *de novo* peptide sequencing model, we trained two variants that differed only in their training data. Specifically, the baseline variant was trained solely on the original dataset, whereas the augmented variant was trained on the original dataset supplemented with NTS-generated PSMs. Both variants were evaluated under identical settings, thereby isolating the effect of NTS augmentation on model performance.

4.3 Main Results

As shown in Table 1 and Appendix A.3, our proposed NTS consistently and significantly improves the performance and generalization of *de novo* peptide sequencing models, including Casanovo, π-HelixNovo, and π-HelixNovo2, across 11 independent test datasets. Specifically, NTS improves AA precision, AA recall, and peptide recall across diverse evaluation settings, including standard tryptic MS/MS datasets from multiple species (ABRF-PXD008808), non-tryptic datasets derived using alternative proteases or without enzymatic digestion (PXD004452–MSV000094014), and challenging multi-enzyme antibody datasets (MSV000079801). The latter comprises human IgG1 digested with six proteases and mouse IgG1 digested with three proteases, providing a stringent evaluation of cross-enzyme generalization.

Table 1: Peptide recall (%) of Casanovo, π-HelixNovo, and π-HelixNovo2 models trained with and without NTS. For each model, the better-performing variant is shown in **bold**. Relative changes are shown in parentheses, with red for increases and blue for decreases.

Dataset	*Species*	Sample / Tissue	Digestion	Casanovo / +NTS	π-HelixNovo / +NTS	π-HelixNovo2 / +NTS
ABRF data	*H. sapiens*	HeLa cell-line	Trypsin	35.9 / **44.4** (+23.7%)	46.7 / **49.3** (+5.6%)	52.8 / **58.5** (+10.8%)
PXD012645	*H. sapiens*	Lung cancer	Trypsin	30.2 / **38.7** (+28.1%)	39.3 / **39.5** (+0.5%)	42.3 / **46.8** (+10.6%)
PXD008844	*M. musculus*	Brain	Trypsin	44.2 / **50.1** (+13.3%)	51.1 / **52.7** (+3.1%)	54.6 / **57.3** (+4.9%)
PXD010559	*P. berghei*	ANKA strain	Trypsin	27.6 / **35.7** (+29.3%)	41.6 / **43.1** (+3.6%)	46.1 / **49.4** (+7.2%)
PXD016315	*A. Thaliana*	Flower buds	Trypsin	33.0 / **40.8** (+23.6%)	48.8 / **50.4** (+3.3%)	56.8 / **59.0** (+3.9%)
PXD013233	*C. Elegans*	Whole organism	Trypsin	31.9 / **42.2** (+32.3%)	47.0 / **49.6** (+5.5%)	53.7 / **58.0** (+8.0%)
PXD016001	*E. Coli*	Whole cells	Trypsin	44.0 / **55.1** (+25.2%)	61.9 / **64.5** (+4.2%)	69.9 / **72.2** (+3.3%)
PXD008808	*T. truncatus*	Blubber	Trypsin	37.6 / **44.0** (+17.0%)	**46.0** / 45.9 (-0.2%)	50.1 / **54.1** (+8.0%)
		Brain	Trypsin	37.7 / **46.0** (+22.0%)	46.9 / **47.7** (+1.7%)	51.4 / **56.0** (+8.9%)
		Kidney	Trypsin	39.0 / **47.4** (+21.5%)	48.6 / **50.5** (+3.9%)	53.8 / **57.9** (+7.6%)
		Liver	Trypsin	37.4 / **46.0** (+23.0%)	48.8 / **50.3** (+3.1%)	53.4 / **58.5** (+9.6%)
		Muscle	Trypsin	39.0 / **47.4** (+21.5%)	48.5 / **50.2** (+3.5%)	53.9 / **58.7** (+8.9%)
		Skin	Trypsin	40.1 / **48.0** (+19.7%)	49.3 / **50.5** (+2.4%)	53.8 / **58.2** (+8.2%)
PXD004452	*H. sapiens*	HeLa cell-line	Chymotrypsin	5.2 / **10.6** (+103.8%)	12.3 / **16.0** (+30.1%)	19.6 / **24.2** (+23.5%)
			Glu-C	4.4 / **9.5** (+115.9%)	10.9 / **13.6** (+24.8%)	15.9 / **19.9** (+25.2%)
			Lys-C	20.5 / **27.7** (+35.1%)	29.4 / **31.2** (+6.1%)	34.2 / **37.5** (+9.6%)
			Trypsin	29.2 / **38.8** (+32.9%)	40.2 / **43.1** (+7.2%)	47.3 / **51.7** (+9.3%)
MSV000094014	*H. sapiens*	Synthetic & C-terminal peptides	Non-enzymatic	8.0 / **17.4** (+117.5%)	17.3 / **22.7** (+31.2%)	23.8 / **32.6** (+37.0%)
MSV000079801	*H. sapiens*	IgG1	Asp-N	1.5 / **5.2** (+246.7%)	3.5 / **4.7** (+34.3%)	6.9 / **12.7** (+84.1%)
			Chymotrypsin	0.5 / **1.0** (+100.0%)	2.7 / **3.3** (+22.2%)	5.1 / **7.9** (+54.9%)
			Glu-C	5.5 / **5.9** (+7.3%)	6.2 / **7.0** (+12.9%)	14.0 / **17.0** (+21.4%)
			Lys-C	11.8 / **13.4** (+13.6%)	14.2 / **16.4** (+15.5%)	20.2 / **21.8** (+7.9%)
			Proteinase-K	2.0 / **3.4** (+70.0%)	4.3 / **4.5** (+4.7%)	6.7 / **7.9** (+17.9%)
			Trypsin	11.2 / **15.8** (+41.1%)	16.9 / **19.0** (+12.4%)	23.9 / **27.5** (+15.1%)
	M. musculus	IgG1	Asp-N	8.5 / **14.1** (+65.9%)	17.3 / **17.9** (+3.5%)	23.3 / **27.9** (+19.7%)
			Chymotrypsin	5.1 / **9.7** (+90.2%)	10.2 / **17.6** (+72.5%)	18.5 / **24.4** (+31.9%)
			Trypsin	21.5 / **28.7** (+33.5%)	31.7 / **37.5** (+18.3%)	43.3 / **46.8** (+8.1%)

Collectively, these results demonstrate that NTS improves cross-species and cross-enzyme generalization, enhances the robustness of *de novo* peptide sequencing models to unseen digestion conditions and complex proteolytic regimes.

4.4 THE INTERPRETATION OF NTS

NTS expands the scale and diversity of the training data by generating an additional dataset with the same number of PSMs but substantially more diverse C-terminal amino acid compositions. As shown in Figure 3, the C-terminal amino acids in the original training set are dominated by lysine (K) and arginine (R), which together account for 95.8% of all occurrences, reflecting the strong bias introduced by tryptic digestion. The remaining 18 amino acids account for only 4.2% of occurrences, with cysteine (C), proline (P), and tryptophan (W) each appearing fewer than 1,000 times among more than 1.5 million PSMs. After NTS augmentation, the proportion of K and R decreases to 49.8%, while the frequencies of the other amino acids increase by an average of approximately 24-fold. Notably, the occurrences of threonine (T) and valine (V) increase by approximately 49-fold and 47-fold, respectively. Therefore, NTS provides substantially more diverse fragmentation patterns in the training data, contributing to the improved performance on non-tryptic and multi-enzyme datasets reported in section 4.3.

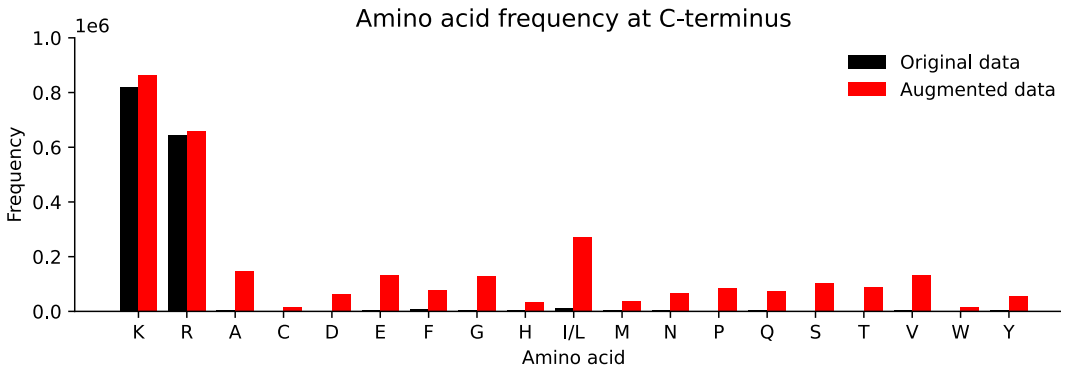

Figure 3: Comparison of C-terminal amino acid frequency distributions in the training dataset before and after NTS augmentation.

5 CONCLUSION

In this study, we proposed NTS, an interpretable and general-purpose data augmentation method for improving the generalization of *de novo* peptide sequencing models. By removing the C-terminal residue, updating precursor information, selectively removing or shifting the affected fragment ions, and reconstructing the resulting spectrum, NTS simulates training PSMs with diverse non-tryptic termini directly from existing experimental data, without requiring additional wet-lab experiments or spectrum-generation models. Extensive experiments across multiple species, tissues, proteases, and peptide types demonstrate that NTS consistently improves the performance and generalization of representative *de novo* peptide sequencing models, underscoring the importance of training-data diversity. These results further suggest that data augmentation offers a promising data-centric strategy for advancing *de novo* peptide sequencing. Future work could focus on improving the chemical fidelity of spectrum simulation, enabling more diverse training distributions and facilitating the exploration of the dark proteome.

AI USE STATEMENT

In this work, we used generative AI tools, including ChatGPT and Claude, to assist with manuscript preparation, mainly for English language polishing, grammar checking, and improving sentence clarity and readability. We did not use generative AI tools to generate research ideas, formulate hypotheses, design or conduct experiments, analyze or interpret experimental results, or produce experimental results. All AI-assisted text was carefully reviewed and revised by the authors to ensure its accuracy and consistency with the scientific content. The authors take full responsibility for the final content of this work, including all text and claims that were produced or revised with the assistance of generative AI.

Reproducibility Statement

We have taken several steps to facilitate the reproducibility of our results. Specifically, the proposed NTS augmentation method is described in detail in Section 3 and illustrated in Figure 2, the training dataset used in this work is publicly available at `https://zenodo.org/record/6791263/files/casanovo_preprocessed_data.zip?download=1`, and the codes of the de novo peptide sequencing models are available at their corresponding GitHub repositories. Furthermore, the source code of NTS, all model weights trained in this work, and the processed test datasets are provided via an anonymous, downloadable link at `https://zenodo.org/records/22787952`.

References

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

Hamed Habibi Aghdam, Elnaz Jahani Heravi, et al. Guide to convolutional neural networks. *New York, NY: Springer*, 10(978-973):51, 2017.

Dorte B Bekker-Jensen, Christian D Kelstrup, Tanveer S Batth, Sara C Larsen, Christa Haldrup, Jesper B Bramsen, Karina D Sørensen, Søren Høyer, Torben F Ørntoft, Claus L Andersen, et al. An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell systems*, 4(6):587–599, 2017.

Denis Beslic, Georg Tscheuschner, Bernhard Y Renard, Michael G Weller, and Thilo Muth. Comprehensive evaluation of peptide de novo sequencing tools for monoclonal antibody assembly. *Briefings in bioinformatics*, 24(1):bbac542, 2023.

Hao Chi, Rui-Xiang Sun, Bing Yang, Chun-Qing Song, Le-Heng Wang, Chao Liu, Yan Fu, Zuo-Fei Yuan, Hai-Peng Wang, Si-Min He, et al. pnovo: de novo peptide sequencing and identification using hcd spectra. *Journal of proteome research*, 9(5):2713–2724, 2010.

Joshua J Coon, John EP Syka, Jeffrey Shabanowitz, and Donald F Hunt. Tandem mass spectrometry for peptide and protein sequence analysis. *Biotechniques*, 38(4):519–523, 2005.

Sean R Eddy. What is dynamic programming? *Nature biotechnology*, 22(7):909–910, 2004.

Jimmy K Eng, Brian C Searle, Karl R Clauser, and David L Tabb. A face in the crowd: recognizing peptides through database search. *Molecular & Cellular Proteomics*, 10(11), 2011.

Jan Fíla, Božena Klodová, David Potěšil, Miloslav Juříček, Petr Šesták, Zbyněk Zdráhal, and David Honys. The beta subunit of nascent polypeptide associated complex plays a role in flowers and siliques development of arabidopsis thaliana. *International Journal of Molecular Sciences*, 21(6): 2065, 2020.

Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods*, 16(6):509–518, 2019.

Tiannan Guo, Judith A Steen, and Matthias Mann. Mass-spectrometry-based proteomics: from single cells to clinical applications. *Nature*, 638(8052):901–911, 2025.

Lin He, Jolene Diedrich, Yen-Yin Chu, and John R Yates III. Extracting accurate precursor information for tandem mass spectra by rawconverter. *Analytical chemistry*, 87(22):11361, 2015.

Amelia Ritahani Ismail. Adam optimization algorithm for wide and deep neural network. *Knowledge Engineering and Data Science*, 2019.

Richard S Johnson, Stephen A Martin, and Klaus Biemann. Collision-induced fragmentation of (m+ h)+ ions of peptides. side chain specific sequence ions. *International Journal of Mass Spectrometry and Ion Processes*, 86:137–154, 1988.

Joel Lapin, Xinjian Yan, and Qian Dong. Unispec: Deep learning for predicting the full range of peptide fragment ion series to enhance the proteomics data analysis workflow. *Analytical Chemistry*, 96(7):2783–2790, 2024.

Sangjeong Lee and Hyunwoo Kim. Bidirectional de novo peptide sequencing using a transformer model. *PLOS Computational Biology*, 20(2):e1011892, 2024.

Kaiyuan Liu, Yuzhen Ye, Sujun Li, and Haixu Tang. Accurate de novo peptide sequencing using fully convolutional neural networks. *Nature Communications*, 14(1):7974, 2023.

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.

Jiawei Mao, Keyun Wang, Luyao Liu, Hongqiang Qin, and Ming-liang Ye. Complementary ion spectrum strategy for improving the sensitivity of open search. *Journal of Chinese Mass Spectrometry Society*, 41(1):11–18, 2020.

Zeping Mao, Ruixue Zhang, Lei Xin, and Ming Li. Mitigating the missing-fragmentation problem in de novo peptide sequencing with a two-stage graph-based deep learning model. *Nature Machine Intelligence*, 5(11):1250–1260, 2023.

Benjamin A Neely, Debra L Ellisor, and W Clay Davis. Proteomics as a metrological tool to evaluate genome annotation accuracy following de novo genome assembly: a case study using the atlantic bottlenose dolphin (tursiops truncatus). *Genes*, 14(9):1696, 2023.

Jesper V Olsen, Shao-En Ong, and Matthias Mann. Trypsin cleaves exclusively c-terminal to arginine and lysine residues. *Molecular & cellular proteomics*, 3(6):608–614, 2004.

Béla Paizs and Sándor Suhai. Fragmentation pathways of protonated peptides. *Mass spectrometry reviews*, 24(4):508–548, 2005.

Rui Qiao, Ngoc Hieu Tran, Lei Xin, Xin Chen, Ming Li, Baozhen Shan, and Ali Ghodsi. Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nature Machine Intelligence*, 3(5):420–425, 2021.

Antonio Ramos-Fernandez, Alberto Paradela, Rosana Navajas, and Juan Pablo Albar. Generalized method for probability-based peptide and protein identification from tandem mass spectrometry data and sequence database searching. *Molecular & Cellular Proteomics*, 7(9):1748–1754, 2008.

Jesse Rodriguez, Nitin Gupta, Richard D Smith, and Pavel A Pevzner. Does trypsin cut before proline? *Journal of proteome research*, 7(1):300–305, 2008.

P Roepstorfe. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.*, 11:601–605, 1984.

Joerg Seidler, Nico Zinn, Martin E Boehm, and Wolf D Lehmann. De novo sequencing of peptides by ms/ms. *Proteomics*, 10(4):634–649, 2010.

Michael W Senko, Steven C Beu, and Fred W McLaffertycor. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 6(4):229–233, 1995.

Melanie J Shears, Raja Sekhar Nirujogi, Kristian E Swearingen, Santosh Renuse, Satish Mishra, Panga Jaipal Reddy, Robert L Moritz, Akhilesh Pandey, and Photini Sinnis. Proteomic analysis of plasmodium merosomes: the link between liver and blood stages in malaria. *Journal of proteome research*, 18(9):3404–3418, 2019.

Omar Shouman, Wassim Gabriel, Victor-George Giurcoiu, Vitor Sternlicht, and Mathias Wilhelm. Prospect: Labeled tandem mass spectrometry dataset for machine learning in proteomics. *Advances in Neural Information Processing Systems*, 35:32882–32896, 2022.

Seo Young Sim, Yu Ri Choi, Jun Hyung Lee, Jae Min Lim, Seung-Eun Lee, Kwang Pyo Kim, Jin Young Kim, Seung Hyeun Lee, and Min-Sik Kim. In-depth proteomic analysis of human bronchoalveolar lavage fluid toward the biomarker discovery for lung cancers. *PROTEOMICS–Clinical Applications*, 13(5):1900028, 2019.

Yingying Sun, Jun A, Zhiwei Liu, Rui Sun, Liujia Qian, Samuel H Payne, Wout Bittremieux, Markus Ralser, Chen Li, Yi Chen, et al. Ai proteomics: from protein identification to virtual cells. *Nature Methods*, pp. 1–13, 2026.

Rebecca George Tharyan, Andrea Annibal, Isabelle Schiffer, Raymond Laboy, Ilian Atanassov, Anna Luise Weber, Birgit Gerisch, and Adam Antebi. Nfyb-1 regulates mitochondrial function and longevity via lysosomal prosaposin. *Nature metabolism*, 2(5):387–396, 2020.

Ngoc Hieu Tran, M Ziaur Rahman, Lin He, Lei Xin, Baozhen Shan, and Ming Li. Complete de novo assembly of monoclonal antibody sequences. *Scientific reports*, 6(1):31730, 2016.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Mingxun Wang, Jian Wang, Jeremy Carver, Benjamin S Pullman, Seong Won Cha, and Nuno Bandeira. Assembling the community-scale discoverable human proteome. *Cell systems*, 7(4):412–421, 2018.

Tingpeng Yang, Tianze Ling, Boyan Sun, Zhendong Liang, Fan Xu, Xiansong Huang, Linhai Xie, Yonghong He, Leyuan Li, Fuchu He, et al. Introducing π-helixnovo for practical large-scale de novo peptide sequencing. *Briefings in Bioinformatics*, 25(2):bbae021, 2024.

Tingpeng Yang, Tianze Ling, Boyan Sun, Zhendong Liang, Cheng Lai, Jiangli Hu, Zexuan Yi, Yonghong He, Leyuan Li, Yue Yu, et al. π-helixnovo2: Making accurate online de novo peptide sequencing available to all. *Genomics, Proteomics & Bioinformatics*, pp. qzag049, 2026a.

Zhiwen Yang, Pan Liu, Yifan Li, Yunhua Zhong, and Jun Xia. Pepspecbench: A unified evaluation benchmark for peptide tandem mass spectrometry prediction. *arXiv preprint arXiv:2605.01945*, 2026b.

John R Yates, Jimmy K Eng, Ashley L McCormack, and David Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical chemistry*, 67(8):1426–1436, 1995.

Huang Yi, Sun Shiyu, Duan Xiusheng, and Chen Zhigang. A study on deep neural networks framework. In *2016 IEEE advanced information management, communicates, electronic and automation control conference (IMCEC)*, pp. 1519–1522. IEEE, 2016.

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(1):6427, 2024.

Yanbao Yu, Aubrie O'Rourke, Yi-Han Lin, Harinder Singh, Rodrigo Vargas Eguez, Sinem Beyhan, and Karen E Nelson. Predictive signatures of 19 antibiotic-induced escherichia coli proteomes. *ACS infectious diseases*, 6(8):2120–2129, 2020.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.

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.

Xie-Xuan Zhou, Wen-Feng Zeng, Hao Chi, Chunjie Luo, Chao Liu, Jianfeng Zhan, Si-Min He, and Zhifei Zhang. pdeep: predicting ms/ms spectra of peptides with deep learning. *Analytical chemistry*, 89(23):12690–12697, 2017.

Ying Zhu, Maowei Dou, Paul D Piehowski, Yiran Liang, Fangjun Wang, Rosalie K Chu, William B Chrisler, Jordan N Smith, Kaitlynn C Schwarz, Yufeng Shen, et al. Spatially resolved proteome mapping of laser capture microdissected tissue with automated sample transfer to nanodroplets. *Molecular & Cellular Proteomics*, 17(9):1864–1874, 2018.

A Appendix

A.1 MS/MS-based proteomics workflow

In a typical MS/MS-based proteomics workflow (Figure 4), proteins are first digested by proteases to produce a mixture of peptides. Subsequently, the resulting peptides are separated by liquid chromatography (LC) and introduced into a mass spectrometer for tandem mass spectrometry analysis (Guo et al., 2025; Coon et al., 2005). During data acquisition, an MS1 survey scan measures the m/z values of precursor ions, from which their charge states and precursor masses are determined (Senko et al., 1995; He et al., 2015). The precursor ions are then isolated and fragmented to produce MS2 spectra containing sequence-informative fragment ions (Roepstorfe, 1984; Johnson et al., 1988). Peptide identification is subsequently performed by analyzing the precursor information together with the corresponding MS2 spectrum, either by matching the observed spectrum against peptide sequences derived from a protein database (Yates et al., 1995; Eng et al., 2011) or by directly inferring the peptide sequence through *de novo* sequencing (Seidler et al., 2010; Ma et al., 2003). Finally, the resulting peptide identifications provide evidence for protein inference, enabling the identification and characterization of proteins (Guo et al., 2025).

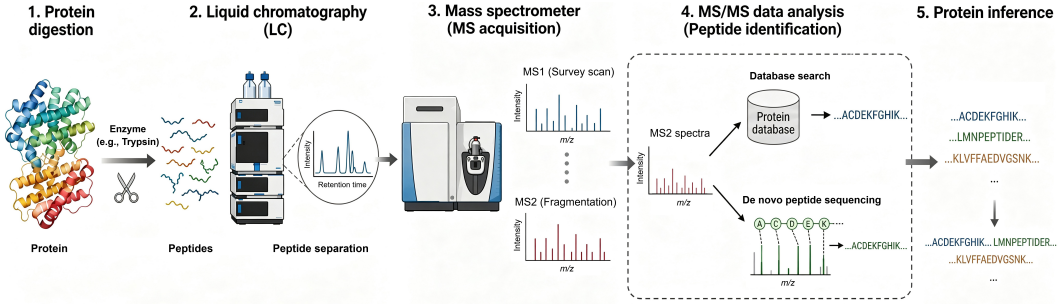

Figure 4: Overview of the tandem mass spectrometry workflow.

A.2 Trypsin digestion bias and restricted C-terminal amino acid diversity

Trypsin is the most widely used protease in MS/MS-based proteomics because of its high cleavage specificity and efficiency (Guo et al., 2025; Bekker-Jensen et al., 2017). As illustrated in Figure 5, trypsin preferentially cleaves peptide bonds at the C-terminal side of lysine (K) or arginine (R), except when K/R is followed by proline (P) (Olsen et al., 2004; Rodriguez et al., 2008). This cleavage specificity produces tryptic peptides of diverse lengths but imposes a strong constraint on their C-terminal residues. Specifically, except for the terminal peptide of a protein, which can terminate in any amino acid, tryptic peptides predominantly end with K or R (Olsen et al., 2004). As a result, K and R occur substantially more frequently than the other 18 amino acids at the peptide C-terminus (Beslic et al., 2023).

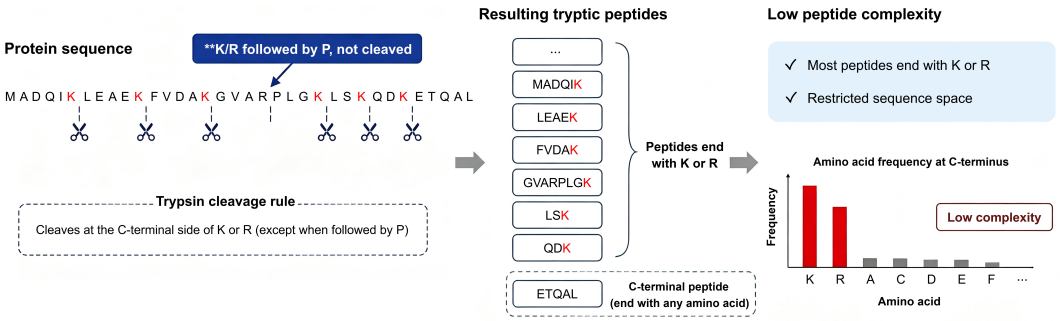

Figure 5: Trypsin cleavage specificity and the resulting bias in peptide C-terminal composition.

A.3 Additional Results

As shown in Table 2 and Table 3, this section presents amino acid-level comparisons of *de novo* peptide sequencing models trained with and without NTS, evaluated in terms of AA recall and AA precision across 11 test datasets covering seven species and a range of tryptic and non-tryptic digestion protocols.

Table 2: AA recall (%) of Casanovo, π-HelixNovo, and π-HelixNovo2 models trained with and without NTS. For each model, the better-performing variant is shown in **bold**. Relative changes are shown in parentheses, with red for increases and blue for decreases.

Dataset	*Species*	Sample / Tissue	Digestion	Casanovo / +NTS	π-HelixNovo / +NTS	π-HelixNovo2 / +NTS
ABRF data	*H. sapiens*	HeLa cell-line	Trypsin	58.3 / **65.7** (+12.7%)	69.4 / **72.3** (+4.2%)	76.1 / **79.6** (+4.6%)
PXD012645	*H. sapiens*	Lung cancer	Trypsin	51.3 / **59.5** (+16.0%)	61.7 / **62.8** (+1.8%)	65.7 / **69.1** (+5.2%)
PXD008844	*M. musculus*	Brain	Trypsin	64.4 / **69.3** (+7.6%)	72.6 / **73.5** (+1.2%)	75.8 / **77.9** (+2.8%)
PXD010559	*P. berghei*	ANKA strain	Trypsin	49.4 / **57.0** (+15.4%)	64.9 / **66.5** (+2.5%)	70.3 / **73.7** (+4.8%)
PXD016315	*A. Thaliana*	Flower buds	Trypsin	55.3 / **61.7** (+11.6%)	69.6 / **71.8** (+3.2%)	76.3 / **78.3** (+2.6%)
PXD013233	*C. Elegans*	Whole organism	Trypsin	54.5 / **63.2** (+16.0%)	67.7 / **70.1** (+3.5%)	72.6 / **75.6** (+4.1%)
PXD016001	*E. Coli*	Whole cells	Trypsin	66.8 / **74.6** (+11.7%)	80.6 / **82.2** (+2.0%)	86.0 / **87.1** (+1.3%)
PXD008808	*T. truncatus*	Blubber	Trypsin	57.6 / **63.4** (+10.1%)	66.4 / **67.3** (+1.4%)	71.3 / **73.8** (+3.5%)
		Brain	Trypsin	58.9 / **65.6** (+11.4%)	68.4 / **70.1** (+2.5%)	73.7 / **76.7** (+4.1%)
		Kidney	Trypsin	61.1 / **67.8** (+11.0%)	70.8 / **73.3** (+3.5%)	76.8 / **79.8** (+3.9%)
		Liver	Trypsin	60.0 / **67.1** (+11.8%)	71.4 / **73.7** (+3.2%)	76.9 / **80.0** (+4.0%)
		Muscle	Trypsin	61.6 / **68.6** (+11.4%)	71.1 / **73.1** (+2.8%)	77.5 / **80.3** (+3.6%)
		Skin	Trypsin	61.5 / **67.9** (+10.4%)	71.1 / **73.1** (+2.8%)	76.6 / **79.4** (+3.7%)
PXD004452	*H. sapiens*	HeLa cell-line	Chymotrypsin	22.2 / **31.6** (+42.3%)	35.0 / **40.3** (+15.1%)	42.8 / **48.1** (+12.4%)
			Glu-C	20.2 / **30.1** (+49.0%)	30.8 / **35.7** (+15.9%)	35.4 / **41.0** (+15.8%)
			Lys-C	40.7 / **48.9** (+20.1%)	50.0 / **51.8** (+3.6%)	54.0 / **57.1** (+5.7%)
			Trypsin	52.0 / **61.6** (+18.5%)	61.9 / **65.0** (+5.0%)	67.5 / **71.5** (+5.9%)
MSV000094014	*H. sapiens*	Synthetic & C-terminal peptides	Non-enzymatic	26.8 / **42.1** (+57.1%)	42.2 / **49.5** (+17.3%)	51.0 / **59.2** (+16.1%)
MSV000079801	*H. sapiens*	IgG1	Asp-N	19.3 / **29.3** (+51.8%)	29.6 / **33.0** (+11.5%)	34.6 / **40.7** (+17.6%)
			Chymotrypsin	8.7 / **14.2** (+63.2%)	16.8 / **21.2** (+26.2%)	25.2 / **28.3** (+12.3%)
			Glu-C	14.7 / **26.2** (+78.2%)	26.0 / **34.0** (+30.8%)	31.9 / **39.8** (+24.8%)
			Lys-C	29.2 / **35.9** (+22.9%)	35.4 / **40.2** (+13.6%)	40.9 / **44.4** (+8.6%)
			Proteinase-K	10.4 / **12.8** (+23.1%)	15.5 / **17.1** (+10.3%)	18.1 / **19.8** (+9.4%)
			Trypsin	34.9 / **42.5** (+21.8%)	43.0 / **45.4** (+5.6%)	48.5 / **51.8** (+6.8%)
	M. musculus	IgG1	Asp-N	25.3 / **32.9** (+30.0%)	37.8 / **40.9** (+8.2%)	46.6 / **53.3** (+14.4%)
			Chymotrypsin	19.1 / **26.2** (+37.2%)	28.6 / **34.9** (+22.0%)	35.7 / **42.0** (+17.6%)
			Trypsin	42.9 / **49.0** (+14.2%)	51.1 / **55.5** (+8.6%)	59.7 / **64.6** (+8.2%)

Consistent improvements in both AA recall and AA precision across Casanovo, π-HelixNovo, and π-HelixNovo2 further confirm the effectiveness of NTS in improving the performance and generalization of *de novo* peptide sequencing models. Notably, these improvements are substantially larger on the non-tryptic datasets PXD004452 and MSV000094014, as well as on the multi-enzyme antibody dataset MSV000079801. Specifically, the relative average AA precision gains from NTS on tryptic versus non-tryptic datasets are 12.0% vs. 37.6% for Casanovo, 2.6% vs. 13.5% for π-HelixNovo, and 3.7% vs. 12.8% for π-HelixNovo2. These results indicate that NTS helps *de novo* sequencing models learn features that better represent non-tryptic PSMs.

Another notable observation is that the gains from NTS are consistently larger for Casanovo than for π-HelixNovo and π-HelixNovo2. Specifically, the relative average improvement in AA precision across all test datasets is 25.3% for Casanovo, compared with 8.2% for π-HelixNovo and 8.4% for π-HelixNovo2. This pattern suggests that model architecture and NTS-based data augmentation may

Table 3: AA precision (%) of Casanovo, π-HelixNovo, and π-HelixNovo2 models trained with and without NTS. For each model, the better-performing variant is shown in **bold**. Relative changes are shown in parentheses, with red for increases and blue for decreases.

Dataset	*Species*	Sample / Tissue	Digestion	Casanovo / +NTS	π-HelixNovo / +NTS	π-HelixNovo2 / +NTS
ABRF data	*H. sapiens*	HeLa cell-line	Trypsin	58.0 / **65.5** (+12.9%)	69.2 / **72.1** (+4.2%)	75.9 / **79.4** (+4.6%)
PXD012645	*H. sapiens*	Lung cancer	Trypsin	51.6 / **59.7** (+15.7%)	61.9 / **63.1** (+1.9%)	66.0 / **69.4** (+5.2%)
PXD008844	*M. musculus*	Brain	Trypsin	64.4 / **69.4** (+7.8%)	72.6 / **73.5** (+1.2%)	75.9 / **77.9** (+2.6%)
PXD010559	*P. berghei*	ANKA strain	Trypsin	49.4 / **57.1** (+15.6%)	65.0 / **66.6** (+2.5%)	70.3 / **73.7** (+4.8%)
PXD016315	*A. Thaliana*	Flower buds	Trypsin	55.3 / **61.7** (+11.6%)	69.6 / **71.6** (+2.9%)	76.2 / **78.2** (+2.6%)
PXD013233	*C. Elegans*	Whole organism	Trypsin	54.8 / **63.3** (+15.5%)	68.1 / **70.3** (+3.2%)	72.7 / **75.7** (+4.1%)
PXD016001	*E. Coli*	Whole cells	Trypsin	66.7 / **74.4** (+11.5%)	80.4 / **82.0** (+2.0%)	85.7 / **86.9** (+1.4%)
PXD008808	*T. truncatus*	Blubber	Trypsin	58.0 / **63.8** (+10.0%)	66.7 / **67.7** (+1.5%)	71.6 / **74.1** (+3.5%)
		Brain	Trypsin	59.0 / **65.7** (+11.4%)	68.4 / **70.2** (+2.6%)	73.7 / **76.7** (+4.1%)
		Kidney	Trypsin	61.3 / **67.9** (+10.8%)	70.9 / **73.4** (+3.5%)	76.8 / **79.8** (+3.9%)
		Liver	Trypsin	60.0 / **67.2** (+12.0%)	71.4 / **73.7** (+3.2%)	76.9 / **79.9** (+3.9%)
		Muscle	Trypsin	61.8 / **68.8** (+11.3%)	71.3 / **73.2** (+2.7%)	77.6 / **80.3** (+3.5%)
		Skin	Trypsin	61.8 / **68.0** (+10.0%)	71.2 / **73.2** (+2.8%)	76.7 / **79.5** (+3.7%)
PXD004452	*H. sapiens*	HeLa cell-line	Chymotrypsin	22.7 / **32.0** (+41.0%)	35.6 / **40.8** (+14.6%)	43.2 / **48.5** (+12.3%)
			Glu-C	20.9 / **30.5** (+45.9%)	31.6 / **36.4** (+15.2%)	36.0 / **41.6** (+15.6%)
			Lys-C	41.4 / **49.5** (+19.6%)	50.8 / **52.5** (+3.3%)	54.7 / **57.7** (+5.5%)
			Trypsin	52.4 / **61.7** (+17.7%)	62.1 / **65.1** (+4.8%)	67.6 / **71.6** (+5.9%)
MSV000094014	*H. sapiens*	Synthetic & C-terminal peptides	Non-enzymatic	27.0 / **42.3** (+56.7%)	42.5 / **49.8** (+17.2%)	51.1 / **59.4** (+16.2%)
MSV000079801	*H. sapiens*	IgG1	Asp-N	19.6 / **29.9** (+52.6%)	30.1 / **33.7** (+12.0%)	35.1 / **41.4** (+17.9%)
			Chymotrypsin	8.7 / **14.4** (+65.5%)	17.2 / **21.4** (+24.4%)	25.5 / **28.7** (+12.5%)
			Glu-C	15.0 / **26.8** (+78.7%)	26.8 / **35.1** (+31.0%)	32.6 / **41.0** (+25.8%)
			Lys-C	30.0 / **36.9** (+23.0%)	36.5 / **41.2** (+12.9%)	41.9 / **45.6** (+8.8%)
			Proteinase-K	10.6 / **13.1** (+23.6%)	16.1 / **17.6** (+9.3%)	18.5 / **20.5** (+10.8%)
			Trypsin	35.3 / **42.9** (+21.5%)	43.3 / **45.8** (+5.8%)	48.8 / **52.3** (+7.2%)
	M. musculus	IgG1	Asp-N	25.8 / **33.5** (+29.8%)	38.5 / **41.8** (+8.6%)	46.8 / **53.7** (+14.7%)
			Chymotrypsin	19.6 / **26.8** (+36.7%)	29.3 / **35.5** (+21.2%)	36.4 / **42.7** (+17.3%)
			Trypsin	43.5 / **49.5** (+13.8%)	51.7 / **55.9** (+8.1%)	60.0 / **64.9** (+8.2%)

provide partially overlapping benefits for improving model performance. In particular, π-HelixNovo incorporates a complementary spectrum strategy that enriches the ion information available from the corresponding experimental MS/MS spectrum. As a form of data augmentation, this strategy may already help the model learn spectral features that are useful for sequencing non-tryptic peptides, thereby reducing the additional benefit that NTS can provide.

Taken together, these per-dataset results further support the findings presented in the main text and confirm that NTS is a broadly applicable approach for addressing the training-data bottleneck in *de novo* peptide sequencing.