PROTEIN LANGUAGE MODEL–ALIGNED SPECTRA EMBEDDINGS FOR DE NOVO PEPTIDE SEQUENCING

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the problem of *de novo* peptide sequencing in tandem mass spectrometry, where the goal is to predict the underlying peptide sequence given a spectrum's fragment peaks and precursor information. We present PLMNovo, a constrained learning framework that leverages pre-trained protein language models (PLMs) to guide the training process. In particular, we cast peptide-spectrum matching as a constrained optimization problem that enforces alignment between spectrum and peptide embeddings produced by a spectrum encoder and a PLM, respectively. We use a Lagrangian primal-dual algorithm to train the spectrum encoder and the peptide decoder by solving the proposed constrained learning problem, while optionally fine-tuning the pre-trained PLM. Through numerical experiments on established benchmarks, we demonstrate that PLMNovo outperforms several state-of-the-art deep learning-based *de novo* sequencing algorithms.

1 INTRODUCTION

Tandem mass spectrometry (MS/MS) is central to bottom-up proteomics, wherein proteins are digested into peptides, fragmented, and recorded as spectra that capture sequence information (Neagu et al., 2022). The central computational problem is to translate each spectrum into its underlying peptide sequence (and often its modifications), enabling downstream protein inference, quantification, and biological interpretation. Accurate and scalable peptide identification underpins biomedical applications ranging from pathway mapping to biomarker and therapeutic development (Schirle et al., 2012; Liu et al., 2013; Pejchinovski et al., 2024). Yet, it remains challenging due to, among other reasons, noisy and incomplete fragmentation, instrument variability, and the sheer volume of data generated in modern experiments (McDonnell et al., 2022; Mao et al., 2023; Du et al., 2025).

Most production pipelines rely on *database search* methods: candidate peptides are enumerated from an *in silico*–digested protein database (with specified enzyme rules and optional variable modifications), theoretical fragment spectra are generated, and candidate matches are scored with procedures that enable false discovery rate (FDR) control (Kapp & Schütz, 2007). This strategy is robust when the database is appropriate and the search space is limited. However, it inherits several structural limitations. The primary challenge with database search methods is that they cannot identify peptides absent from the database, such as single-amino-acid variants or sequences from unmodeled organisms (van Puyenbroeck et al., 2025). Moreover, the computational burden grows combinatorially, e.g., with variable post-translational modifications (PTMs). As the candidate space expands, runtimes increase, scores become less discriminative, and FDR control becomes extremely challenging (Neuhauser, 2013).

De novo peptide sequencing removes the reliance on a fixed database by inferring sequences directly from spectra, casting the task as a structured supervised prediction problem under strict mass constraints. Recent deep learning approaches, ranging from transformer-based encoder–decoder models (Yilmaz et al., 2022; 2024; Eloff et al., 2025) to graph-based pipelines (Mao et al., 2023), have learned to incorporate precursor mass and charge to map spectra to amino acid sequences. Trained on large-scale spectral libraries and adaptable across instruments and fragmentation modes, these models can recover peptides not found in any database and reveal biology otherwise missed by search-based pipelines.

Most state-of-the-art deep learning-based *de novo* peptide sequencing pipelines create intermediate representations, or embeddings, of the fragment peaks contained in the spectrum, which are then used by a peptide decoder to reconstruct the underlying amino acid sequence. In essence, during training, this is a multi-modal learning problem, where we have access to two different modalities of the same data (i.e., spectrum peaks and peptide sequence). While prior efforts have been made to regularize this embedding space using both modalities (Jin et al., 2024), they have relied on peptide encoders that are trained from scratch and lack prior biological knowledge of the corresponding amino acid sequences, as compared to large-scale pre-trained biological foundation models.

In this work, we leverage *protein language models (PLMs)* (Bepler & Berger, 2021; Rives et al., 2021; Elnaggar et al., 2021; Lin et al., 2023; Hayes et al., 2025) to enhance the training procedure of deep learning *de novo* peptide sequencing pipelines. We propose a *constrained learning* formulation of *de novo* peptide sequencing, referred to as PLMNovo, where the spectra embeddings generated by the spectrum decoder are forced to be aligned with their corresponding peptide embeddings generated by a pre-trained PLM. We use a primal-dual training algorithm that identifies the right balance between minimizing the primary amino acid classification objective, while respecting the above *alignment constraints* between the spectra and peptide embeddings. By training our model over a massive dataset of 2 million peptide-spectrum matches (PSMs), we demonstrate the superiority of PLMNovo over state-of-the-art baselines. Moreover, we provide additional results that highlight the interplay between our proposed alignment constraints and the characteristics of peptides and spectra.

Our contributions are as follows:

- We propose, for the first time, a constrained learning pipeline that integrates pre-trained protein language models (PLMs) into the *de novo* peptide sequencing procedure.
- We provide a Lagrangian duality-based method for solving the sequencing problem under peptide-spectrum embedding alignment constraints.
- We numerically demonstrate that our proposed method outperforms state-of-the-art deep learning *de novo* sequencing pipelines across standard benchmarks.

2 Related Work

2.1 *De Novo* Peptide Sequencing

Building on the limitations of database search, *de novo* sequencing has become essential in settings where relevant peptide sequences are missing or incomplete, such as metaproteomics, immunopeptidomics, antibody sequencing, and paleoproteomics. Recent progress has been primarily driven by deep learning since the introduction of DeepNovo (Tran et al., 2017). Contemporary models span convolutional, transformer, and graph-based architectures. Representative examples include PointNovo (Qiao et al., 2021), Casanovo and its follow-up versions (Yilmaz et al., 2022; 2024; Melendez et al., 2024), PepNet (Liu et al., 2023), GraphNovo (Mao et al., 2023), InstaNovo (Eloff et al., 2025), and MassNet (Jun et al., 2025). Beyond accuracy, interpretability has begun to receive attention, with π-xNovo (Wang et al., 2024b) utilizing multi-head attention to link predicted residues to specific spectral peaks, thereby offering post-hoc explanations of model decisions.

Alongside methodological advances, the field has grappled with evaluation and reliability. Key open issues include principled false discovery rate (FDR) control that jointly considers database search and *de novo* results, understanding the tradeoff between database size and detection power, and estimating the fraction of "foreign" spectra in a dataset to contextualize performance gains. Benchmarking also requires care: retraining and hyperparameter tuning are often necessary for fair comparisons, as default settings can yield suboptimal results and distort conclusions (Zhou et al., 2024). Despite these challenges, rapid progress, the expansion of public datasets, and improvements in instrumentation suggest that *de novo* deep learning pipelines will continue to mature and broaden their impact in proteomics.

2.2 PROTEIN LANGUAGE MODELS

Large-scale language models for text have demonstrated that attention-based transformers (Vaswani et al., 2017) are powerful general-purpose sequence learners. The same recipe has been adopted for proteins: the availability of hundreds of millions of natural sequences (for example, from UniRef (Suzek et al., 2007; 2015) and BFD (Jumper et al., 2021)) enables pretraining at internet scale, turning protein sequences into the "language" on which to learn syntax (motifs), semantics (function), and long-range dependencies (contacts). This line of work has produced protein language models (PLMs) that learn rich contextual embeddings directly from amino-acid sequences and transfer surprisingly well to downstream biological tasks.

PLMs differ in architecture and objective but share the core idea of self-supervision. Masked language models, such as the ESM family (Rives et al., 2021; Lin et al., 2023; ESM Team, 2024; Hayes et al., 2025) and ProtT5 Elnaggar et al. (2021), learn to recover hidden residues from context and often scale to billions of parameters. Autoregressive generators, such as ProtGPT2 (Ferruz et al., 2022) and ProGen (Nijkamp et al., 2023), model next-token distributions and are used for controllable sequence generation. Moreover, context-based PLMs exploit multiple sequence alignment, or MSAs, that encode evolutionary variation explicitly (Rao et al., 2021; Truong Jr & Bepler, 2025; Akiyama et al., 2025). Recent efforts have extended the context length and improved efficiency (Chen et al., 2025a) and adopted diffusion-style generative processes for discrete sequences (Wang et al., 2024a).

The resulting representations from pre-trained PLMs have driven state-of-the-art or competitive performance across diverse applications, such as zero-shot and few-shot prediction of mutational effects (Meier et al., 2021; Brandes et al., 2023), functional annotation (Martínez-Redondo et al., 2025), and controllable protein and peptide design (Lee et al., 2024; Chen et al., 2025b). In proteomics specifically, PLM embeddings and priors have been utilized to enhance peptide property predictors, such as retention time, detectability, and MS/MS fragmentation pattern (Nakai-Kasai et al., 2025).

In this work, we present a novel application of PLMs in proteomics by integrating them into the *de novo* peptide sequencing pipeline, as described next.

3 METHOD

De novo peptide sequencing can be formulated as a supervised classification problem. Assume we have access to a set of N annotated training samples $\{(\mathbf{p}_i, \mathbf{s}_i, \mathsf{prec}_i)\}_{i=1}^{N}$. For every $i \in \{1, \ldots, N\}$, $\mathbf{p}_i \in \mathcal{Y}^{L_i}$ represents the peptide sequence of length L_i, with $\mathcal{Y}$ denoting the amino acid alphabet, and $\mathbf{s}_i \in (\mathbb{R}_+ \times \mathbb{R}_+)^{K_i}$ denotes the observed spectrum composed of K_i peaks, with each peak represented as a pair of m/z (mass to charge ratio) and intensity values. Moreover, $\mathsf{prec}_i \in \mathbb{R}_+^3$ denotes the precursor information of the i^{th} training sample, consisting of the precursor mass, precursor charge, and retention time. The goal of *de novo* peptide sequencing is to reconstruct the ground-truth peptide sequence given the observed spectrum and the precursor information. More precisely, we are interested in a parameterized function f_θ that solves the following empirical risk minimization problem:

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \ell_{\mathsf{CE}}\Big(f_\theta(\mathbf{s}_i, \mathsf{prec}_i), \mathbf{p}_i\Big), \tag{1}$$

where $\ell_{\mathsf{CE}}(\cdot, \cdot)$ denotes the cross-entropy reconstruction loss function between the predicted and ground-truth peptide sequences, and Θ denotes the set of all possible model parameters. The supervised learning formulation in (1) has been used in the majority of the recent work on deep learning-based *de novo* peptide sequencing (Yilmaz et al., 2022; 2024; Eloff et al., 2025). Most of these studies break down the end-to-end function f_θ into an encoder $g_{\theta_{\mathsf{enc}}}$ and a decoder $h_{\theta_{\mathsf{dec}}}$, where the encoder maps the observed spectrum into intermediate peak-level representations, which are then used by the decoder, alongside the precursor information, to reconstruct the peptide sequence at the output, i.e.,

$$f_\theta(\mathbf{s}_i, \mathsf{prec}_i) = h_{\theta_{\mathsf{dec}}}\Big(g_{\theta_{\mathsf{enc}}}(\mathbf{s}_i), \mathsf{prec}_i\Big), \tag{2}$$

with the encoder and decoder parameterized using transformer architectures (Vaswani et al., 2017).

In this paper, we take a different approach from the supervised learning formulation in (1). Given the success of pre-trained protein language models (PLMs) in deriving informative representations from amino acid sequences, we hypothesized that including a PLM in the training process could benefit the generalization power of the *de novo* sequencing pipeline. More specifically, for a given peptide-spectrum match (PSM) $(\mathbf{p}, \mathbf{s})$, we propose *aligning* the peptide embedding generated by a PLM with the spectrum embedding $g_{\theta_{\mathsf{enc}}}(\mathbf{s})$ created by the spectrum encoder.

For a peptide p of length L, let $m_{\theta_{\mathsf{PLM}}}(\mathbf{p}) \in \mathbb{R}^{L \times d'}$ denote the residue-level embeddings generated by a PLM $m_{\theta_{\mathsf{PLM}}}$. Moreover, assume the peak-level embeddings generated by the spectrum encoder $g_{\theta_{\mathsf{enc}}}$ lie in R^d, i.e., $g_{\theta_{\mathsf{enc}}}(\mathbf{s}) \in \mathbb{R}^{K \times d}$ for a spectrum $\mathbf{s}$ with K peaks. We then aggregate these embeddings using two pooling modules. To derive a unified spectrum-peptide co-embedding space, we further use a projection module to map the peptide embedding to $\mathbb{R}^d$. With a slight abuse of notation, we let $e_{\theta_{\mathsf{s}}}(\cdot) \in \mathbb{R}^d$ denote the pooling function for the spectrum embeddings and $e_{\theta_{\mathsf{p}}}(\cdot) \in \mathbb{R}^d$ represent the combined pooling and projection function for the peptide embeddings.

Our proposed method, PLMNovo, enforces the peptide-spectrum embedding alignment via a *constrained* learning approach (Chamon et al., 2022). Formally, we solve the following constrained optimization problem:

$$\min_{\theta_{\mathsf{enc}}, \theta_{\mathsf{dec}}, \theta_{\mathsf{s}}, \theta_{\mathsf{p}}, \theta_{\mathsf{PLM}}} \quad \frac{1}{N} \sum_{i=1}^{N} \ell_{\mathsf{CE}} \left[h_{\theta_{\mathsf{dec}}} \left(g_{\theta_{\mathsf{enc}}}(\mathbf{s}_i), \mathsf{prec}_i \right), \mathbf{p}_i \right], \tag{3a}$$

$$\text{s.t.} \quad \left\| e_{\theta_{\mathsf{s}}} \left(g_{\theta_{\mathsf{enc}}}(\mathbf{s}_i) \right) - e_{\theta_{\mathsf{p}}} \left(m_{\theta_{\mathsf{PLM}}}(\mathbf{p}_i) \right) \right\|_2^2 \leq \epsilon, \quad \forall i \in \{1, \ldots, N\}, \tag{3b}$$

where $\|\mathbf{x} - \mathbf{y}\|_2$ denotes the Euclidean distance between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. The learning problem in (3) attempts to find the model parameters that not only minimize the primary *de novo* sequencing objective in (3a), but also ensure that the spectra and peptide embeddings are closely aligned. The alignment is introduced as a *per-PSM constraint* in (3b), where the squared Euclidean distance between the aggregated spectrum and peptide embeddings, both in $\mathbb{R}^d$, is at most ϵ. The upper bound, ϵ, is treated as a hyperparameter: extremely low values of ϵ could make the problem infeasible or lead to degenerate solutions (where all embeddings collapse to a small subset of the embedding space), while $\epsilon \to \infty$ reverts the problem to the original unconstrained problem in (1). Therefore, its choice is critical in generating informative spectrum (and peptide) embeddings. Figure 1 illustrates an overview of PLMNovo.

Observe that the optimization variables in (3) include those of the spectrum encoder θ_{enc}, the peptide decoder θ_{dec}, the spectrum pooling θ_{s}, the peptide pooling and projection θ_{p}, and the PLM θ_{PLM}. While all these parameters are involved during the training phase, only the spectrum encoder and peptide decoder parameters (i.e., $\theta_{\mathsf{enc}}, \theta_{\mathsf{dec}}$) are used during inference. Furthermore, the PLM parameters can be kept frozen (at a pre-trained checkpoint), or trained end-to-end (e.g., via fine-tuning (Hu et al., 2022; Schmirler et al., 2024; Sledzieski et al., 2024)). We note that such an MSE-based alignment approach resembles the BYOL self-supervised learning method (Grill et al., 2020), but instead of augmentations, here we consider embeddings from two different data modalities and corresponding feature extractors.

3.1 Primal-Dual Training

To solve the constrained learning problem in (3), we move to the dual domain (Boyd & Vandenberghe, 2004) and write the Lagrangian function as

$$\begin{aligned} &\mathcal{L}(\theta_{\mathsf{enc}}, \theta_{\mathsf{dec}}, \theta_{\mathsf{s}}, \theta_{\mathsf{p}}, \theta_{\mathsf{PLM}}, \boldsymbol{\lambda}) \\ &= \frac{1}{N} \sum_{i=1}^{N} \ell_{\mathsf{CE}} \left[h_{\theta_{\mathsf{dec}}} \left(g_{\theta_{\mathsf{enc}}}(\mathbf{s}_i), \mathsf{prec}_i \right), \mathbf{p}_i \right] + \sum_{i=1}^{N} \lambda_i \left[\left\| e_{\theta_{\mathsf{s}}} \left(g_{\theta_{\mathsf{enc}}}(\mathbf{s}_i) \right) - e_{\theta_{\mathsf{p}}} \left(m_{\theta_{\mathsf{PLM}}}(\mathbf{p}_i) \right) \right\|_2^2 - \epsilon \right], \end{aligned} \tag{4}$$

where $\lambda_i \geq 0$ is the *Lagrangian dual multiplier* corresponding to the i^{th} training sample, and $\boldsymbol{\lambda} \in \mathbb{R}_+^N$ denotes the vector of all dual multipliers. The dual of problem (3) is then given by the saddle-point problem

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}_+^N} \min_{\theta_{\mathsf{enc}}, \theta_{\mathsf{dec}}, \theta_{\mathsf{s}}, \theta_{\mathsf{p}}, \theta_{\mathsf{PLM}}} \mathcal{L}(\theta_{\mathsf{enc}}, \theta_{\mathsf{dec}}, \theta_{\mathsf{s}}, \theta_{\mathsf{p}}, \theta_{\mathsf{PLM}}, \boldsymbol{\lambda}). \tag{5}$$

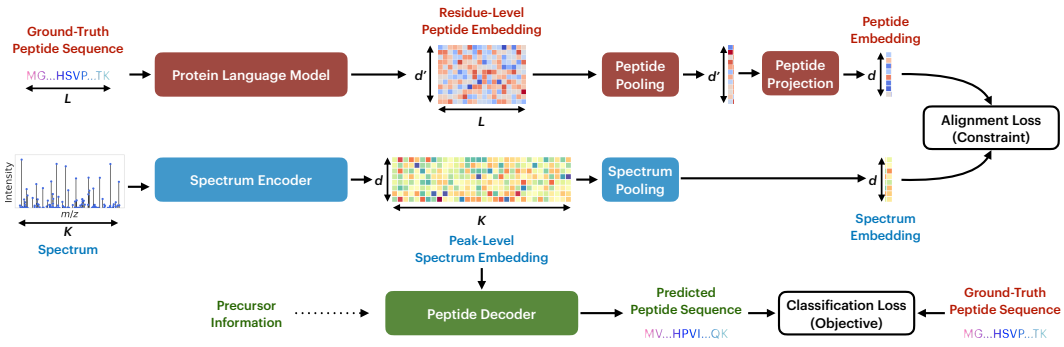


Figure 1: Our proposed architecture, PLMNovo, consists of an encoder-decoder pair, whose intermediate embeddings are constrained by a protein language model (PLM). In particular, the spectrum fragment peak information (comprising K (m/z, intensity) pairs) is mapped to a peak-level $K \times d$ spectrum embedding using a spectrum encoder. These intermediate embeddings are then fed into a peptide decoder, alongside the precursor mass and charge, to predict the corresponding peptide sequence. The predicted peptide sequence is compared to the ground-truth peptide sequence using the cross-entropy loss, which constitutes the primary objective function. Simultaneously, the ground-truth peptide sequence (comprising L amino acids) is mapped to a residue-level $L \times d'$ peptide embedding using a PLM. The spectrum and peptide embeddings are aggregated using two pooling modules, and the aggregated peptide embedding is projected to the same embedding space as the spectrum embedding (i.e., $\mathbb{R}^d$). The squared Euclidean distance between the resulting peptide and spectrum embeddings is then enforced to be bounded by a constant, which acts as a per peptide-spectrum match (PSM) constraint in the optimization problem. While the spectrum encoder-decoder pair, as well as the pooling and projection modules, are trained end-to-end, the PLM is fine-tuned from a pre-trained checkpoint via the alignment loss gradients.

To solve this problem, we can use a primal-dual approach (Boyd & Vandenberghe, 2004; Fioretto et al., 2021; Elenter et al., 2022), where we alternate between updating the model parameters and the dual multiplier. More specifically, in each primal iteration, the model parameters are updated using *gradient descent* on the Lagrangian, i.e.,

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta_{\boldsymbol{\theta}} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}, \tag{6}$$

where $\eta_{\boldsymbol{\theta}}$ denotes the primal learning rate, and to ease the notation, we aggregate all the model parameters into $\boldsymbol{\theta} = [\theta_{\mathsf{enc}}, \theta_{\mathsf{dec}}, \theta_{\mathsf{s}}, \theta_{\mathsf{p}}, \theta_{\mathsf{PLM}}]$. Then, in each dual iteration, the dual multipliers are updated using *projected gradient ascent* on the Lagrangian, i.e.,

$$\boldsymbol{\lambda} \leftarrow \left[\boldsymbol{\lambda} + \eta_{\boldsymbol{\lambda}} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}}\right]_+, \tag{7}$$

where $\eta_{\boldsymbol{\lambda}}$ denotes the dual learning rate, and $[\cdot]_+ := \max(\cdot, 0)$ represents elementwise mapping onto the $\mathbb{R}_+^N$. Combining (4) and (7), we can rewrite the update of each Lagrangian multiplier in closed form as

$$\lambda_i \leftarrow \left[\lambda_i + \eta_{\boldsymbol{\lambda}} \left(\left\| e_{\theta_{\mathsf{s}}}\left(g_{\theta_{\mathsf{enc}}}(\mathbf{s}_i)\right) - e_{\theta_{\mathsf{p}}}\left(m_{\theta_{\mathsf{PLM}}}(\mathbf{p}_i)\right)\right\|_2^2 - \epsilon\right)\right]_+. \tag{8}$$

The closed-form update in (8) implies that the dual multiplier corresponding to each PSM accumulates the alignment constraint violations over the course of the training process. In other words, if for a PSM, the spectra and peptide embeddings are perfectly aligned, its corresponding dual multiplier remains at zero, whereas in the case of a PSM for which there is a significant misalignment between the spectra and peptide embeddings, its dual multiplier keeps increasing. This shows how the proposed constrained formulation adapts the importance of different training sample alignments by dynamically adjusting the importance of each PSM in the Lagrangian in (4).

4 NUMERICAL RESULTS

4.1 EXPERIMENTAL SETTINGS

We base PLMNovo's implementation on Casanovo 4.2 (Melendez et al., 2024), a state-of-the-art autoregressive *de novo* sequencing pipeline trained on a dataset of 2 million PSMs, all of which are digested using trypsin, the standard enzyme used in tandem mass spectrometry (Melendez, 2024). We train all PLMNovo models on this dataset. This dataset originates from the MassIVE Knowledge Base (Wang et al., 2018) and is accompanied by a test split containing 200,000 tryptic samples. In what follows, we refer to this dataset as the MSKB dataset. Casanovo 4.2 is built on the previous versions of Casanovo (Yilmaz et al., 2022; 2024). We follow the exact hyperparameters used by Melendez et al. (2024), including using a 9-layer transformer-based encoder-decoder architecture with a $d = 512$-dimensional embedding space, and a beam search decoding mechanism with $k = 10$ beams. The rest of the hyperparameters and model details are mentioned in Appendix A. We retrain Casanovo v4.2 on the MSKB training set, and all results reported below corresponding to this architecture are based on our retrained version.

As for the PLMs, we use two PLM architectures from the ESM-2 family (Lin et al., 2023), namely the 8M (with $d' = 320$) and 650M (with $d' = 1280$) versions. We remove any post-translational modifications (PTMs) from the ground-truth peptide sequences before feeding them to the PLMs to respect their token vocabularies, which are based on the canonical amino acids. To manage computational complexity, we leverage low-rank adaptation, or LoRA (Hu et al., 2022), to fine-tune the PLM parameters, focusing on only the key and value parameter matrices in the self-attention layers, as recommended in prior work (Sledzieski et al., 2024). We utilize average pooling to aggregate both peptide and spectrum embeddings.

We perform grid search on two important hyperparameters: the alignment constraint bound ($\epsilon \in \{10^{-1}, 10^{-2}, 10^{-3}\}$) in (3b), and the LoRA fine-tuning rank ($r \in \{0, 2, 4\}$). Treating the MSKB test set as a validation split, for each PLM, we select the (ϵ, r) combination that leads to the lowest amino acid classification loss. Our implementation code can be found at `https://github.com/AnonMS2/PLMNovo`.

4.2 PERFORMANCE ON THE MSKB AND MULTI-ENZYME TEST SETS

The top portion of Table 1 compares the performance of PLMNovo and Casanovo v4.2 on the MSKB test set. Across all amino acid-level and peptide-level metrics, PLMNovo, especially with the 8M version of ESM-2, outperforms the base Casanovo v4.2 model, demonstrating the performance boost that the PLM-guided alignment constraints provide in PLMNovo.

While the MSKB test set provides an in-distribution evaluation setting, we also tested our models on a held-out non-tryptic multi-enzyme dataset (Melendez, 2024; Melendez et al., 2024). As the bottom portion of the Table 1 shows, PLMNovo also outperforms Casanovo v4.2 on this out-of-distribution dataset. Interestingly, ESM-2 650M significantly outperforms ESM-2 8M on this dataset, suggesting that the smaller-scale PLM may have led to slight overfitting of PLMNovo on tryptic data.

Dataset	Method	Classification Loss (↓)	AA Precision (↑)	AA Recall (↑)	Peptide Precision (↑)
MSKB (Tryptic)	Casanovo v4.2 (Melendez et al., 2024)	0.1983	0.8919	0.8884	0.7381
	PLMNovo (ESM-2 8M)	**0.1919**	0.8973	**0.8921**	**0.7411**
	PLMNovo (ESM-2 650M)	0.1941	**0.8976**	0.8915	0.7400
Multi-Enzyme (Non-Tryptic)	Casanovo v4.2 (Melendez et al., 2024)	0.9949	0.5200	0.5211	0.2467
	PLMNovo (ESM-2 8M)	0.9723	0.5249	0.5283	0.2410
	PLMNovo (ESM-2 650M)	**0.9714**	**0.5356**	**0.5323**	**0.2483**

Table 1: Amino acid-level and peptide-level performance comparison on the MSKB and multi-enzyme test sets (Melendez, 2024). Numbers in **bold** represent the best result under each (metric, dataset) combination.

4.3 Performance on the Nine-species Benchmark

We next evaluated PLMNovo, pre-trained on the MSKB training set, on the nine-species benchmark, which is a standard dataset used by the majority of prior work on *de novo* peptide sequencing. As Table 2 shows, while PLMNovo provides competitive performance in terms of peptide recall, it outperforms all other baselines in terms of the amino acid precision.

Metric	Method	Species									Average
		Bacillus	*C. bacteria*	*Honeybee*	*Human*	*M. mazei*	*Mouse*	*Ricebean*	*Tomato*	*Yeast*	
AA Precision (↑)	DeepNovo (Tran et al., 2017)	0.742	0.602	0.630	0.610	0.694	0.623	0.679	0.731	0.750	0.673
	PointNovo (Qiao et al., 2021)	0.768	0.589	0.644	0.606	0.712	0.626	0.730	0.733	0.779	0.687
	Casanovo (Yilmaz et al., 2022)	0.749	0.603	0.629	0.586	0.679	0.689	0.668	0.721	0.684	0.667
	AdaNovo (Xia et al., 2024)	0.739	0.642	0.650	0.618	0.728	0.646	0.719	0.740	**0.793**	0.697
	Casanovo v2 (Yilmaz et al., 2024)	0.790	0.681	0.706	**0.676**	**0.755**	**0.760**	0.748	0.785	0.752	0.739
	Casanovo v4.2 (Melendez et al., 2024)	0.793	0.678	0.705	0.668	0.687	0.756	0.753	0.791	0.768	0.733
	PLMNovo (ESM-2 8M)	0.794	0.684	0.707	0.670	**0.755**	0.759	0.764	0.790	0.755	0.742
	PLMNovo (ESM-2 650M)	**0.797**	**0.688**	**0.709**	**0.676**	**0.755**	**0.760**	**0.766**	**0.795**	0.771	**0.746**
Peptide Recall (↑)	DeepNovo (Tran et al., 2017)	0.449	0.253	0.330	0.293	0.422	0.286	0.436	0.454	0.462	0.376
	PointNovo (Qiao et al., 2021)	0.518	0.298	0.396	0.351	0.478	0.355	0.511	0.513	0.534	0.439
	CasaNovo (Yilmaz et al., 2022)	0.537	0.330	0.406	0.341	0.478	0.426	0.506	0.521	0.490	0.448
	AdaNovo (Xia et al., 2024)	0.528	0.372	0.431	0.373	0.496	0.467	0.546	0.530	0.593	0.481
	Casanovo v2 (Yilmaz et al., 2024)	**0.622**	**0.446**	**0.493**	**0.446**	**0.557**	**0.483**	**0.589**	**0.618**	**0.599**	**0.539**
	Casanovo v4.2 (Melendez et al., 2024)	0.603	0.421	0.478	0.437	0.498	0.468	0.558	0.608	0.584	0.517
	PLMNovo (ESM-2 8M)	0.602	0.423	0.484	0.434	0.544	0.470	0.565	0.606	0.571	0.522
	PLMNovo (ESM-2 650M)	0.601	0.425	0.483	0.432	0.544	0.464	0.576	0.606	0.579	0.523

Table 2: Performance comparison results in terms of amino acid precision and peptide recall on the nine-species dataset (Tran et al., 2017). The performance of the baseline methods, except for Casanovo v4.2, is reported from (Zhang et al., 2025b). **Bolded** and underlined results represent the first and second best performance per species (or averaged across species in the last column), respectively.

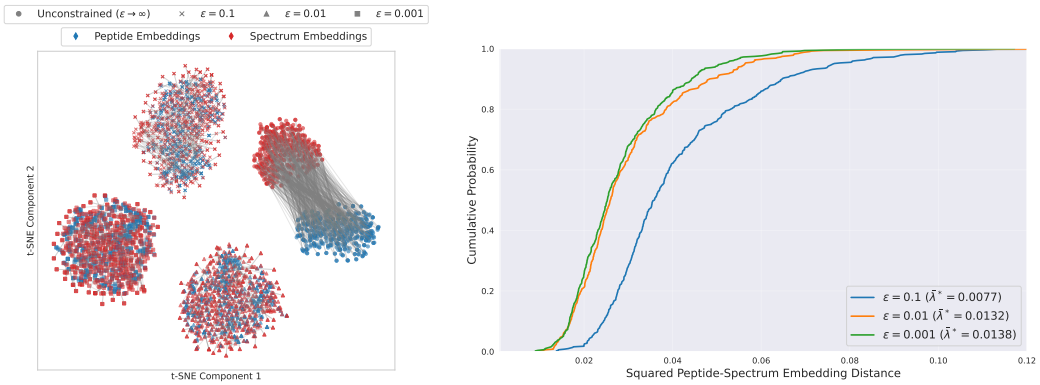
4.4 Impact of the Constraints on the Embedding Space

Figure 2-(a) provides a two-dimensional visualization of the embedding space occupied by the peptide and spectrum embeddings of the MSKB test set using t-SNE (Maaten & Hinton, 2008) under different alignment constraint bounds. As the figure shows, our proposed alignment constraints highly regularize the peptide-spectrum co-embedding space, where matching peptide and spectrum embeddings lie close to each other. This is in contrast to the unconstrained case ($\epsilon \rightarrow \infty$), where peptides and spectra embeddings are very far from each other.

Furthermore, as shown in Figure 2-(b), tighter constraint bounds generally lead to lower Euclidean distances for the PSMs in the embedding space, even though enforcement of such alignment gets more and more challenging as ϵ is reduced.

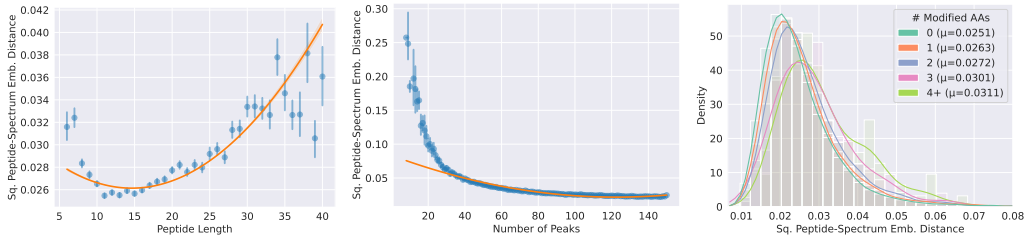
4.5 Interpretation of Embedding Alignment

Alignments for Different Peptide and Spectrum Scales. Figures 3-(a) and 3-(b) show the squared Euclidean distance between peptide and spectrum embeddings for different peptide lengths and number of spectrum peaks. Interestingly, the PSM co-embedding distance increases for extremely short or long peptides, and decreases with an increasing number of peaks. The latter phenomenon is intuitive, since having more peaks can lead to a more informative representation of the spectrum, hence improving the peptide decoding process. The former phenomenon has also been previously reported in the literature (Zhou et al., 2024), where longer peptide sequences are generally more complicated to decode, especially when using autoregressive methods, and shorter peptide sequences suffer from a lack of sufficient training data. The inclusion of more sophisticated pooling methods for spectrum and peptide embeddings (Zhang et al., 2020; Stärk et al., 2021; NaderiAlizadeh & Singh, 2025; Amir & Dym, 2025; Tartici et al., 2025; NaderiAlizadeh et al., 2025)), as well as non-autoregressive decoding algorithms (Zhang et al., 2025a;b)), could enhance the performance of PLMNovo, especially for longer peptide sequences.



(a) (b)

Figure 2: (a) t-SNE (Maaten & Hinton, 2008) visualization of the peptide-spectrum co-embedding space produced by PLMNovo at various alignment constraint bounds (ϵ) levels on a subset of 1000 PSMs from the MSKB test set. (b) Empirical cumulative distributions of the peptide-spectrum embedding distances under different constraint levels, alongside the corresponding optimal Lagrangian dual multipliers, averaged across the training samples.



(a) (b) (c)

Figure 3: Squared Euclidean distance of PSM embeddings vs. (a) peptide length and (b) number of peaks in the spectrum. (c) Distribution of the squared Euclidean PSM embedding distance for different numbers of modified amino acids in the peptide sequences.

Impact of Post-Translational Modifications. Post-Translational Modifications, or PTMs, are biological processes where proteins are chemically modified after translation. While prevalent in proteomics, modified amino acids resulting from PTMs, such as oxidation of methionine and deamidation of asparagine or glutamine, constitute a small fraction of the amino acids in tandem mass spectrometry training datasets. This has led to previous de novo sequencing models struggling to decode PTMs, thereby motivating methods especially designed to handle modified residues, such as AdaNovo (Xia et al., 2024) and PrimeNovo (Zhang et al., 2025a). We sought to understand the impact of PTMs on PLMNovo's learned co-embedding space. Figure 3-(c) shows the histograms of squared peptide-spectrum embedding distances for peptide sequences with different numbers of modified amino acids on the MSKB test set. As the figure demonstrates, the embedding gap between a spectrum and its corresponding peptide increases with the number of PTMs in the peptide sequence. We hypothesize that this limitation stems from the inability of common PLMs, such as the ESM-2 family, to handle PTMs. PLM architectures specifically designed to handle PTM tokens, such as PTM-Mamba (Peng et al., 2025), can potentially bridge this gap and enhance PLMNovo's ability to identify modified amino acids during the decoding process.

4.6 Ablation Study

Figure 4 presents an ablation study of PLMNovo in terms of different alignment constraint bounds and PLM fine-tuning ranks for 8M and 650M versions of ESM-2. As the figure shows, fine-tuning the PLM generally helps improve PLMNovo's performance compared to keeping the PLM frozen

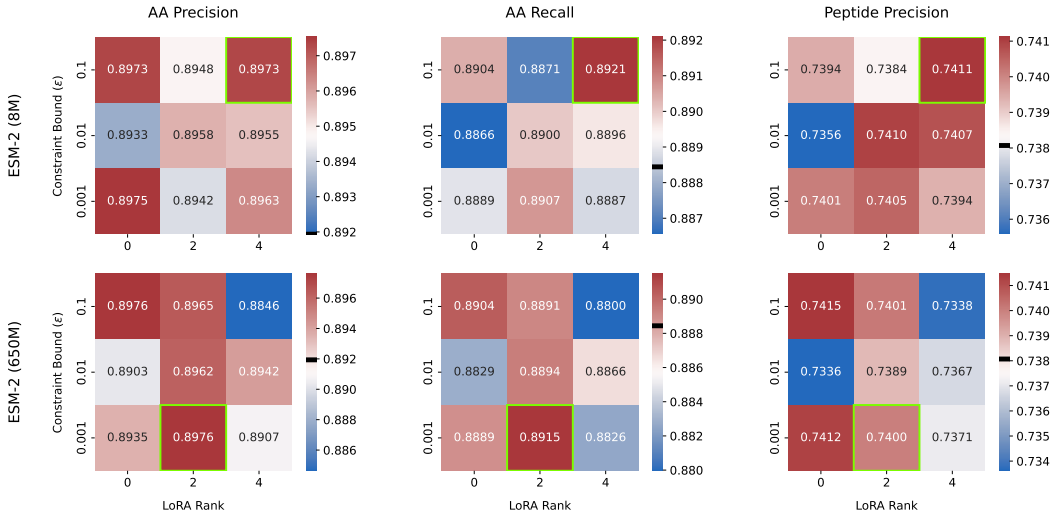


Figure 4: Ablation study of PLMNovo's performance on the MSKB test set in terms of the alignment constraint bound (ϵ) in (3b), as well as the fine-tuning rank of the PLM, where a rank of zero implies that the PLM was frozen. Boxes highlighted in green correspond to the selected hyperparameter combination. Furthermore, the black lines in the colorbars represent Casanovo 4.2's performance, which is equivalent to PLMNovo's unconstrained performance (i.e., $\epsilon \to \infty$).

(LoRA rank = 0). While the majority of the configurations in our grid search exceed the performance of Casanovo v4.2, these results show that tuning these hyperparameters is critical to maximize PLMNovo's predictive power.

5 Discussion and Concluding Remarks

We presented PLMNovo, a novel deep learning method for *de novo* peptide sequencing, which enforces the proximity of matching peptides and spectra in a co-embedding space. While the spectrum embeddings are generated using a spectrum encoder, we use pre-trained protein language models (PLMs) to create the peptide embeddings. We formulate the sequencing problem using a constrained learning framework and adopt a primal-dual algorithm to train the spectrum encoder and peptide decoder, while fine-tuning the PLM. Numerical experiments demonstrated that PLMNovo surpasses other baseline deep learning *de novo* peptide sequencing methods in various benchmarks.

In the future, we envision several immediate enhancements to PLMNovo, such as investigating other PLM architectures and families (e.g., (Elnaggar et al., 2021; Nijkamp et al., 2023; ESM Team, 2024; Truong Jr & Bepler, 2025; Peng et al., 2025)) and alternative embedding pooling strategies (e.g., (Tartici et al., 2025; NaderiAlizadeh & Singh, 2025)). While our pipeline is implemented based on Casanovo 4.2 (Melendez et al., 2024), we expect our gains to transfer to more recent *de novo* sequencing pipelines that employ techniques such as curriculum learning (Zhang et al., 2025b), sequence re-ranking (Qiu et al., 2025), missing fragmentation imputation (Du et al., 2025), and non-autoregressive decoding (Zhang et al., 2025a). Finally, extension of the proposed constrained learning approach to data-independent acquisition (DIA) (Sanders et al., 2025), a more challenging protocol than the data-dependent acquisition (DDA) setting studied in this work, is an essential avenue for further impact of PLMNovo in mass spectrometry research.

6 Use of Large Language Models

Large Language Models (LLMs) were utilized to aid in manuscript review and editorial refinement.

REFERENCES

Yo Akiyama, Zhidian Zhang, Milot Mirdita, Martin Steinegger, and Sergey Ovchinnikov. Scaling down protein language modeling with msa pairformer. *bioRxiv*, pp. 2025–08, 2025.

Tal Amir and Nadav Dym. Fourier sliced-Wasserstein embedding for multisets and measures. In *The Thirteenth International Conference on Learning Representations*, 2025.

Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell Systems*, 12(6):654–669.e3, 2021. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2021.05.017.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Nadav Brandes, Grant Goldman, Charlotte H Wang, Chun Jimmie Ye, and Vasilis Ntranos. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55(9):1512–1522, 2023.

Luiz FO Chamon, Santiago Paternain, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained learning with non-convex losses. *IEEE Transactions on Information Theory*, 69(3):1739–1760, 2022.

Bo Chen, Xingyi Cheng, Pan Li, Yangli-ao Geng, Jing Gong, Shen Li, Zhilei Bei, Xu Tan, Boyan Wang, Xin Zeng, et al. Xtrimopglm: unified 100-billion-parameter pretrained transformer for deciphering the language of proteins. *Nature Methods*, pp. 1–12, 2025a.

Leo Tianlai Chen, Zachary Quinn, Madeleine Dumas, Christina Peng, Lauren Hong, Moises Lopez-Gonzalez, Alexander Mestre, Rio Watson, Sophia Vincoff, Lin Zhao, et al. Target sequence-conditioned design of peptide binders using masked language modeling. *Nature Biotechnology*, pp. 1–9, 2025b.

Ye Du, Chen Yang, Nanxi Yu, Wanyu Lin, Qian Zhao, and Shujun Wang. Latent imputation before prediction: A new computational paradigm for de novo peptide sequencing. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qou3A0Bjzt.

Juan Elenter, Navid NaderiAlizadeh, and Alejandro Ribeiro. A Lagrangian duality approach to active learning. *Advances in Neural Information Processing Systems*, 35:37575–37589, 2022.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

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.

ESM Team. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. EvolutionaryScale Website, December 2024. URL https://evolutionaryscale.ai/blog/esm-cambrian. Accessed: September 25, 2025.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Ferdinando Fioretto, Pascal Van Hentenryck, Terrence WK Mak, Cuong Tran, Federico Baldo, and Michele Lombardi. Lagrangian duality for constrained deep learning. In *Machine learning and knowledge discovery in databases. applied data science and demo track: European conference, ECML pKDD 2020, Ghent, Belgium, September 14–18, 2020, proceedings, part v*, pp. 118–135. Springer, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zhi Jin, Sheng Xu, Xiang Zhang, Tianze Ling, Nanqing Dong, Wanli Ouyang, Zhiqiang Gao, Cheng Chang, and Siqi Sun. ContraNovo: a contrastive learning approach to enhance de novo peptide sequencing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 144–152, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

A Jun, Xiang Zhang, Xiaofan Zhang, Jiaqi Wei, Te Zhang, Yamin Deng, Pu Liu, Zongxiang Nie, Yi Chen, Nanqing Dong, et al. Massnet: billion-scale ai-friendly mass spectral corpus enables robust de novo peptide sequencing. *bioRxiv*, pp. 2025–06, 2025.

Eugene Kapp and Frédéric Schütz. Overview of tandem mass spectrometry (MS/MS) database search algorithms. *Current Protocols in Protein Science*, 49(1):25–2, 2007.

Minji Lee, Luiz Felipe Vecchietti, Hyunkyu Jung, Hyun Joo Ro, Meeyoung Cha, and Ho Min Kim. Robust optimization in protein fitness landscapes using reinforcement learning in latent space. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=0zbxwvJqwf.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

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.

Yansheng Liu, Ruth Hüttenhain, Ben Collins, and Ruedi Aebersold. Mass spectrometric protein maps for biomarker discovery and clinical research. *Expert review of molecular diagnostics*, 13 (8):811–825, 2013.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

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.

Gemma I Martínez-Redondo, Francisco M Perez-Canales, Belén Carbonetto, José M Fernández, Israel Barrios-Núñez, Marçal Vázquez-Valls, Ildefonso Cases, Ana M Rojas, and Rosa Fernández. FANTASIA leverages language models to decode the functional dark proteome across the animal tree of life. *Communications Biology*, 8(1):1227, 2025.

Kevin McDonnell, Enda Howley, and Florence Abram. The impact of noise and missing fragmentation cleavages on de novo peptide identification algorithms. *Computational and structural biotechnology journal*, 20:1402–1412, 2022.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

Carlo Melendez. Data for 'accounting for digestion enzyme bias in Casanovo' (melendez et al., 2024), June 2024. URL https://doi.org/10.5281/zenodo.12587317.

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

Navid NaderiAlizadeh and Rohit Singh. Aggregating residue-level protein language model embeddings with optimal transport. *Bioinformatics Advances*, 5(1):vbaf060, 2025.

Navid NaderiAlizadeh, Darian Salehi, Xinran Liu, and Soheil Kolouri. Constrained sliced wasserstein embedding. *arXiv preprint arXiv:2506.02203*, 2025.

Ayano Nakai-Kasai, Kosuke Ogata, Yasushi Ishihama, and Toshiyuki Tanaka. Leveraging pretrained deep protein language model to predict peptide collision cross section. *Communications Chemistry*, 8(1):137, 2025.

Anca-Narcisa Neagu, Madhuri Jayathirtha, Emma Baxter, Mary Donnelly, Brindusa Alina Petre, and Costel C Darie. Applications of tandem mass spectrometry (ms/ms) in protein analysis for biomedical research. *Molecules*, 27(8):2411, 2022.

Nadin Neuhauser. *Computational approaches to enhance mass spectrometry-based proteomics*. PhD thesis, Technische Universität München, 2013.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

Martin Pejchinovski, Pedro Magalhães, and Jochen Metzger. Mass spectrometry-based proteomics in drug discovery and development, 2024.

Fred Zhangzhi Peng, Chentong Wang, Tong Chen, Benjamin Schussheim, Sophia Vincoff, and Pranam Chatterjee. Ptm-mamba: a ptm-aware protein language model with bidirectional gated mamba blocks. *Nature Methods*, 22(5):945–949, 2025.

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.

Zijie Qiu, Jiaqi Wei, Xiang Zhang, Sheng Xu, Kai Zou, Zhi Jin, ZhiQiang Gao, Nanqing Dong, and Siqi Sun. Universal biological sequence reranking for improved de novo peptide sequencing. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=HtSdgubxsJ.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International conference on machine learning*, pp. 8844–8856. PMLR, 2021.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Justin Sanders, Bo Wen, Paul A Rudnick, Richard S Johnson, Christine C Wu, Michael Riffle, Sewoong Oh, Michael J MacCoss, and William Stafford Noble. A transformer model for de novo sequencing of data-independent acquisition mass spectrometry data. *Nature Methods*, pp. 1–7, 2025.

Markus Schirle, Marcus Bantscheff, and Bernhard Kuster. Mass spectrometry-based proteomics in preclinical drug discovery. *Chemistry & biology*, 19(1):72–84, 2012.

Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15(1):7407, 2024.

Samuel Sledzieski, Meghana Kshirsagar, Minkyung Baek, Rahul Dodhia, Juan Lavista Ferres, and Bonnie Berger. Democratizing protein language models with parameter-efficient fine-tuning. *Proceedings of the National Academy of Sciences*, 121(26):e2405840121, 2024.

Hannes Stärk, Christian Dallago, Michael Heinzinger, and Burkhard Rost. Light attention predicts protein location from the language of life. *Bioinformatics Advances*, 1(1):vbab035, 11 2021. ISSN 2635-0041. doi: 10.1093/bioadv/vbab035.

Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

Alp Tartici, Gowri Nayar, and Russ B Altman. Pool parti: a pagerank-based pooling method for identifying critical residues and enhancing protein sequence representations. *Bioinformatics*, 41 (6):btaf330, 2025.

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

Timothy Fei Truong Jr and Tristan Bepler. Understanding protein function with a multimodal retrieval-augmented foundation model. *arXiv preprint arXiv:2508.04724*, 2025.

Sam van Puyenbroeck, Denis Beslic, Tomi Suomi, Tanja Holstein, Thilo Muth, Laura L Elo, Lennart Martens, Robbin Bouwmeester, Tim Van Den Bossche, and Tine Claeys. Limitations of de novo sequencing in resolving sequence ambiguity. *bioRxiv*, pp. 2025–08, 2025.

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.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. In *International Conference on Machine Learning*, 2024a.

Yu Wang, Zhendong Liang, Tianze Ling, Cheng Chang, Tingpeng Yang, Linhai Xie, and Yonghong He. Transforming de novo peptide sequencing by explainable ai. 2024b.

Jun Xia, Shaorong Chen, Jingbo Zhou, Xiaojun Shan, Wenjie Du, Zhangyang Gao, Cheng Tan, Bozhen Hu, Jiangbin Zheng, and Stan Z. Li. Adanovo: Towards robust *De Novo* peptide sequencing in proteomics against data biases. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0zfUiSX5si.

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.

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, 2025a.

Xiang Zhang, Jiaqi Wei, Zijie Qiu, Sheng Xu, Nanqing Dong, ZhiQiang Gao, and Siqi Sun. Curriculum learning for biological sequence prediction: The case of de novo peptide sequencing. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=WbLXcMt2eG.

Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. FSPool: Learning set representations with featurewise sort pooling. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgBA2VYwH.

Jingbo Zhou, Shaorong Chen, Jun Xia, Sizhe Liu, Tianze Ling, Wenjie Du, Yue Liu, Jianwei Yin, and Stan Z. Li. Novobench: Benchmarking deep learning-based \emph{De Novo} sequencing methods in proteomics. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=RQlbMrA5XL.

A Additional PLMNovo Implementation Details

Architecture. Similar to Casanovo 4.2 (Melendez et al., 2024), PLMNovo utilizes transformers (Vaswani et al., 2017) to map spectral information onto amino acid sequences via an encoder-decoder architecture. Peak characteristics (m/z ratios and intensities) undergo distinct encoding procedures, with sinusoidal functions for mass values and learnable linear mappings for intensities, producing unified high-dimensional representations through summation. The encoder component processes these peak embeddings using multi-head attention to establish inter-peak relationships and contextual understanding across the entire spectrum. The peak embeddings then guide the decoder's sequential amino acid prediction task.

Sequence generation operates through step-wise autoregressive decoding initiated with precursor information. In particular, precursor mass and charge values undergo a sinusoidal transformation and a linear layer, respectively, before being integrated into unified embeddings. The decoder leverages both spectral context and precursor information to initiate the construction of amino acid sequences. At each decoding step, the decoder processes embeddings corresponding to the precursor characteristics and all previously predicted amino acids. Beam search maintains diversity by tracking the top k scoring hypotheses throughout decoding, expanding sequences until natural termination or mass constraint violation occurs, with the output sequence being the maximum-scoring complete sequence. Quality control applies mass accuracy filters, penalizing predictions whose theoretical precursor masses exceed specified deviation limits (50 ppm threshold) from observed values.

Hyperparameters. The length of predicted peptides is set to be between a minimum of 6 and a maximum of 100 amino acids. We select at most 150 peaks within each spectrum, with a minimum intensity of 0.01, and a m/z ratio between 50 and 2500. We use a batch size of 32 during training and train the model for 7 epochs (Melendez et al., 2024). We use the Adam optimizer with a primal learning rate of $\eta_{\boldsymbol{\theta}} = 5 \times 10^{-4}$ and a weight decay of 10^{-5} and a dual learning rate of $\eta_{\boldsymbol{\lambda}} = 10^{-2}$. The primal learning rate is warmed up linearly in the first 10^5 iterations, followed by a cosine-shaped decay with a half period of 6×10^5 iterations. A label smoothing factor of 10^{-2} is used when calculating the training classification loss. The encoder and decoder each have 9 self-attention layers, each containing 8 attention heads, with a latent representation dimension of 512 and a fully-connected layer dimension of 1024. For LoRA, we set $\alpha = 4 \times r$ for any selected rank r (Sledzieski et al., 2024).